Do Evaluation Ratings Affect Teachers' Professional Development Activities?

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

September 2015

We document substantial differences across teachers who receive different performance ratings from the educator evaluation system in Tennessee in terms of both the intensity of their self-reported professional development activities, and the extent to which these activities are guided by feedback from their performance evaluations. Using a regression discontinuity design, we test whether ratings per se causally influence these gaps and find that they do not. We then perform an exploratory analysis of an alternative mechanism to explain the gaps. Namely, we examine the link between the feedback that teachers receive during classroom observations and their professional development activities, and find a strong association. We conclude that future research in this area should aim to more rigorously test hypotheses that relate teachers' experiences during classroom observations to their professional development activities.

Keywords: Teacher professional development; teacher evaluation; teacher rating effects; multinomial regression discontinuity analysis

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 Consortium on Research, Evaluation and Development (the Consortium) at Vanderbilt University's Peabody College, which is funded by the State of Tennessee's Race to the Top grant from the United States Department of Education (grant #S395A100032). 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 Consortium and Tennessee Department of Education for providing data and expert insight to conduct our analysis, in particular, Susan Burns, Sara Heyburn, Trish Kelly, Erin O'Hara, and Matthew Pepper. The usual disclaimers apply.

1. Introduction

Research consistently identifies teacher quality as one of the most important determinants of student success, both academically and beyond (Chetty, Friedman and Rockoff 2014a, 2014b; Hanushek and Rivkin, 2010; Kane et al., 2013). However, Weisberg et al. (2009) show that the significant variation in teacher quality that has been established in empirical research is not reflected in traditional teacher evaluation systems, which overwhelmingly award high ratings to teachers that exhibit little variance. Consistent with the prevalence of high ratings, Weisberg et al. (2009) find that most teachers are not provided with critical feedback that can be used to guide improvement. Indeed, across the 12 districts they study, Weisberg et al. (2009) report that no areas for development were identified during the performance evaluations of nearly three out of every four teachers.

One potential mechanism by which emerging, more-rigorous teacher evaluations systems can improve instructional quality is by providing actionable feedback for improvement to a larger fraction of the teaching workforce. Intuitively, moving to a rating system where a smaller fraction of teachers receive the highest possible score should lead to better information being provided to teachers about areas for development. Moreover, as a matter of practice, emerging systems have built-in feedback mechanisms designed for precisely this purpose.

The present study aims to improve our understanding of the relationships between teachers' professional development activities and the performance feedback they receive from new, more rigorous teacher evaluation systems. To learn about these relationships we examine the evaluation

system implemented in the state of Tennessee beginning with the 2011-2012 school year. Tennessee has been at the forefront of policy reform in the area of teacher evaluation nationally, and was a winner of the initial federal Race to the Top competition.[1] Unlike the distributions of ratings from traditional evaluation systems as documented by Weisberg et al. (2009), the rating distribution in Tennessee better reflects the considerable variation in teacher effectiveness identified by research. The evaluation system also includes formal mechanisms to provide feedback to teachers about their strengths and weaknesses, including recommendations for how to improve, which is different from traditional teacher development activities recently critiqued by The New Teacher Project (2015). For example, teachers are guided to participate in workshops held by schools and districts, engage in one-on-one work with mentor teachers and instructional coaches, and conduct self-directed learning. Following Harris and Sass (2011), we generally define these types of behaviors as professional development.

To perform our analysis we merge teacher ratings from the Tennessee evaluation system with their responses to a statewide survey that asks a variety of questions about their professional development choices (among other things). We begin by documenting substantial differences in teachers' self-reported professional development activities across rating levels, showing that teachers with lower ratings report spending more time on professional development activities in general, and are more likely to make professional development choices based on feedback from their evaluations,

[1] According to the GAO (2013), Tennessee was one of only six Race to the Top (RTTT) states to have fully implemented both the teacher and principal evaluation systems by the 2012-13 school year.

than their higher-rated counterparts. Next we examine the extent to which ratings per se causally affect teachers' professional development choices using a regression discontinuity (RD) design. Identification is achieved by comparing teachers who are observationally similar but receive different ratings due to nonlinearities in how teachers' underlying scores in the Tennessee system translate into final ratings.

There are several reasons one might expect ratings alone to affect professional development choices, including (1) direct and indirect rewards and sanctions associated with the ratings, (2) psychological effects of different rating levels on teachers' attitudes and perceived efficacy, and (3) ratings may provide new information to teachers about their performance (Boyd et al., 2011; Dee and Wyckoff, 2015; Jacob, 2005; Taylor and Tyler, 2012; Vigdor, 2009; Drake et al, 2015). Previous studies by Dee and Wyckoff (2015) and Koedel, Li, Springer and Tan (2015) use a similar identification strategy and find that teacher behaviors and perceptions of work are meaningfully affected by rating assignments. However, unlike these previous studies, we find no evidence that rating assignments causally affect teachers' self-reported professional development activities.

Having ruled out an effect of ratings per se, we perform an exploratory analysis of an alternative mechanism that may help to explain the large gaps in professional development activities that we document across differentially rated teachers in Tennessee. Specifically, we consider the possibility that teachers are particularly responsive to the feedback they receive from the classroom observation component of the evaluation process. Responsiveness to this dimension of the evaluation would not show up in our RD analysis because the feedback that teachers receive from

classroom observations is not discontinuous at the system-wide rating thresholds that the RD model leverages for identification (see below). Our exploration of the connection between classroom-observation feedback and teachers' professional development choices reveals a strong association. Although our data are not suited for a rigorous causal investigation, our findings suggest that this is an area worthy of further exploration as researchers aim to determine how newly emerging evaluation systems contribute to self-directed teacher improvement.

2. Background: Teacher Evaluations in Tennessee

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. All five models require a post-observation conference to discuss teachers' strengths and weaknesses based on their classroom performance, which may contribute to teachers' professional development choices. 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.

The final evaluation ratings for teachers during the 2012-2013 evaluation year, which is the focus of this study, are comprised of three components. For teachers with available growth

measures based on student performance on standardized tests, 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. As a practical matter, the score on the latter component is primarily driven by classroom observation scores.[2] 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.[3]

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 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). 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 findings from the RD models because it means that teachers with very similar underlying scores but

[2] This will become clear in Section 6, when we look more closely at how classroom observation scores correlate with the qualitative feedback that teachers received from their evaluations. The biggest item beyond classroom observations that is incorporated into this final component for some teachers is a student survey, which accounted for part of the rating for roughly 15 percent of teachers during the 2012-2013 school year.

[3] Approximately 45 percent of teachers in Tennessee have an individual growth measure (see Table 1).

different discrete ratings were not provided with 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 (spring 2014 – see below) and they did not apply universally. As an example of a policy that was in place, in 2011 the Tennessee General Assembly voted to explicitly tie evaluation ratings to new tenure decisions (*Public Chapter 70,* 2011) by requiring teachers to receive ratings of four or five during the last two years of the pre-tenure probationary period to earn tenure. Teachers who do not receive tenured status at the end of their five-year probationary period may either be rehired under a year-to-year contract or dismissed.

Another example is that teachers working in disadvantaged schools under the Tennessee Achievement School District (ASD) program who earn a rating of three or higher can 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. This suggests that the primary drivers of any behavioral effects of teachers ratings that we find will be driven by other factors such as (1) the transmission of new information about relative performance to individual teachers; (2) the psychological effects of different rating assignments; and (3) informal policies that reward and sanction teachers based on their ratings (e.g., promotions, access to desirable teaching assignments, use of improvement plans or targeting PD to low performers).

3. Data

3.1 Ratings, Administrative and Survey Data

We use 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 for our primary analysis. The rating data are based on teacher performance during the 2012-2013 school year, the second year that the evaluation system had been in place. The administrative data include information on each teacher's gender, race, education level and years of experience, also from the 2012-2013 school year. The survey data are from the annual Tennessee Consortium Survey administered during the spring of the 2013-2014 school year, after teachers received their ratings from 2012-2013 school year. The survey is designed to improve the state's understanding of how the performance evaluation is implemented and how feedback is provided to and processed by teachers.

We use data from the second year of the evaluation system rather than data from the first year for two reasons. First, Dee and Wyckoff (2015) show that teachers became more responsive to the IMPACT evaluation system in Washington DC as it matured, and thus we anticipate that relationships between teachers' evaluation ratings and professional development activities in Tennessee will be more likely to exist during the second year of implementation. Second, the statewide survey was modified between the first and second years of its use, and the 2013-2014 survey asks more questions about professional development activities. Notably, we also replicate

much of the analysis presented below using the limited data on professional development activities from the 2012-2013 survey and obtain qualitatively similar results.

Table 1 provides descriptive statistics for all teachers in the Tennessee ratings database (which includes teachers regardless of whether they received a rating) side-by-side with the analytic sample. In total, our final sample of 24,009 teachers represents about 35 percent of the teachers in Tennessee. Appendix Table A.1 documents the reasons that teachers are omitted from the analytic sample. Two reasons are noteworthy. First, we exclude all teachers evaluated by the COACH model because it produces a lumpy distribution of teacher performance measures that is not compatible with the RD research design, which we use to estimate the causal effects of ratings on professional development activities. Approximately 6 percent of Tennessee teachers are evaluated using the COACH model (see Table 1). Second, we exclude nearly 60 percent of Tennessee teachers because they did not submit a survey and thus did not provide outcomes for our analysis.[4,5]

Our inability to include COACH-evaluated teachers is only a minor limitation given that these teachers represent a small share of the Tennessee teaching workforce. The issue of survey non-response is a larger issue. Although the response rate to the survey is not particularly low for a

[4] Note that Koedel, Li, Springer and Tan (2015) are able to include teachers evaluated by the COACH model but exclude teachers evaluated by the TEM model in their study based on teachers' first-year ratings from 2011-2012. The reason for the switch is that the evaluation models are constantly under development and thus their underlying reporting structures are subject to change from year-to-year. However, it is important to recognize that for both studies, the TEAM model covers the overwhelming majority of teachers, and thus our findings are mostly driven by teachers who are evaluated using this model.

[5] One reason that teachers did not submit a survey is that they left the Tennessee teaching workforce between the rating and survey years, although the vast majority of teachers who did not submit a survey remained in the workforce.

non-mandatory instrument (e.g., see Rockoff et al., 2012; Stutz, 2014; Watts and Becker, 2008), the large fraction of teachers who are dropped from the sample because they did not submit a survey raises two potential concerns. First, the omission of these teachers could affect our ability to identify rating effects on teacher behaviors if the rating treatments also affect survey submissions (a related problem in the context of RD analysis arises in McCrary and Royer, 2011). Second, even if ratings do not affect survey submissions, the generalizability of the findings will be limited to the extent that teachers who submitted a survey are systematically different from other teachers. We delay the investigation of whether the rating treatments cause survey submission until Section 4.2. To examine whether teachers who submitted a survey are different in other ways, Table 1 compares the observable characteristics of all Tennessee teachers in the ratings data file, which to a rough approximation represents the universe of Tennessee teachers, to the teachers in the analytic sample.[6] Overall, the analytic sample is observationally similar to the full sample of Tennessee teachers, which suggests that our findings will generalize, at least to some extent, to the broader teaching population.[7]

[6] Per Table A.1, information for some teachers in the universe sample for some of the characteristics shown in Table 1 is unavailable. Teachers with missing data for particular data elements are omitted from the calculations of the descriptive statistics for those elements in the table. In omitted results we have compared teachers who do and do not submit a survey conditional on having otherwise complete data files (again, see Table A.1) and obtain results that are substantively similar to what we show in Table 1 (in fact, the differences are even smaller for several characteristics because the largest differences between the columns in Table 1 are owing to the exclusion of COACH teachers).

[7] While most of the differences reported in Table 1 are statistically significant, this is driven in large part by the fact that we have very large samples of teachers. For example, even differences between the universe and analytic samples which are clearly not different substantively, like across teacher education levels, are different statistically.

Turning to the substance of Table 1, in addition to generally describing the teaching workforce, it also documents the distribution of teacher ratings in the system. 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 using available data and the RD research design. This is a limitation as this is possibly one of the most important margins for finding effects, particularly if a low rating may trigger the teacher being put on a formal improvement plan, which will include professional development expectations.

However, moving beyond level-1 there are significant fractions of teachers who receive scores at levels 2 through 5, which facilitates our investigation of rating effects on professional development activities at the thresholds 2/3, 3/4 and 4/5. Compared to the districts studied by Weisberg et al. (2009) in which 94 percent of teachers receive one of the top two ratings, just 66 percent of teachers in Tennessee receive one of the top two ratings during the 2012-2013 evaluation. This indicates that ratings from the new Tennessee system are considerably more dispersed than in traditional evaluation systems in public education.[8]

[8] The distribution of ratings is more differentiated in 2012-2013 than it was in 2011-2012. In 2011-2012, only 0.4 percent of all teachers received a level-1 rating and 73 percent of teachers received one of the two top ratings (Koedel, Li, Springer and Tan, 2015).

3.2 Measuring Professional Development with the Survey

We have identified four questions from the Tennessee survey that elicit feedback from teachers regarding their professional development activities. Table 2 summarizes the content of each question and Appendix B shows how each question was presented to teachers on the survey.

Following Koedel, Li, Springer and Tan (2014), for analytic tractability we collapse teacher responses to each question into four categories: (1) positive, (2) negative, (3) non-response by choice and (4) non-response due to position change (i.e., respondent was directed to skip question). To code each answer as either a positive or negative response, we grouped teachers' more-detailed responses into the two categories as described in Appendix B. Positive responses are identified as indicating more intense professional development, or professional development that is more responsive to evaluation feedback; and the converse for negative responses. For non-responses, on average across the four questions 8.3 percent of survey respondents did not answer by choice and 14.3 percent were directed to skip the questions (among other questions) due to a position change.[9]

Table 2 presents initial descriptive evidence on the patterns of positive responses across rating levels. It reports the proportions of positive responses to each question by final rating in the evaluation system, where the denominator in each cell is the total number of submitted surveys

[9] Position changes that resulted in teachers being directed to skip the professional development questions include changes to observer teachers, mentor teachers, counselors, librarians, assistant principals, principals and non-teaching positions. Individuals who changed positions were directed to a different set of questions based on their new positions during the survey. In addition, although instructional coaches and specialists were presented the same questions as the teachers, we categorize these educators' responses as non-response due to position change in order to focus on behaviors specific to teachers.

(including those for which a non-response was recorded for the relevant question). The table shows that teachers who are assigned higher ratings report less intense and less responsive professional development activities. The differences across rating levels are quite large. In Appendix Table A.2, we show that a qualitatively similar pattern is present if we report the positive-response share conditional on teachers who submit either a positive or negative response.

Although the descriptive statistics in Table 2 (and Appendix Table A.2) illustrate a clear association between teachers' ratings and their professional development activities, attributing causality is not straightforward. We overcome the causal inference challenge, at least with respect to testing the hypothesis that the ratings themselves contribute to the observed gaps, using a regression-discontinuity design that we detail in the next section. As noted above, given that we find no evidence to suggest that ratings per se causally influence the patterns in Table 2, we also perform an exploratory analysis focusing on the classroom-observation component of the evaluation. For the exploratory analysis, we supplement the data we have discussed thus far with additional qualitative data capturing the feedback that teachers received from their classroom observations. We elaborate on the qualitative data below in Section 6.

4. Regression Discontinuity Methodology

4.1 Specification

We use RD models to identify the causal effects of teacher ratings on their professional development choices. Our RD design compares differences in self-reported professional development outcomes for teachers whose underlying performance scores are similar, but who

receive different ratings because of the discrete function that translates the underlying scores into final ratings in the Tennessee system. The key assumption for causal inference within the RD framework is that teachers with similar underlying scores are similar in other respects, and thus conditional on the underlying score the discontinuous rating assignments can be viewed as effectively random (Hahn, Todd and Van der Klaauw, 2001; Imbens and Lemieux, 2008; Lee and Lemieux, 2010).

We first examine whether the discontinuities in the data are sharp or fuzzy. Figure 1 shows the probability of a teacher receiving treatment (i.e., the higher rating) as a function of her underlying performance measure for each ratings-pair that we study (2/3, 3/4, 4/5). The graphs in the figure aggregate the scores for individual teachers into 5-point bins and are centered on the threshold value for the higher rating. Figure 1 shows that for the ratings pair 2/3, the discontinuity in converting the underlying performance measures into final ratings is sharp, but the underlying scores do not translate perfectly into final ratings at the other two thresholds.

A standard approach to dealing with fuzzy discontinuities like the ones shown in the figure for the 3/4 and 4/5 thresholds is to use the discontinuity as an instrument for treatment and estimate a local average treatment effect (LATE). However, in our application the outcomes of interest are multinomial and are thus best evaluated using a nonlinear model. To recover consistent estimates of the treatment effects using the standard instrumental variables approach in our nonlinear application requires strong assumptions that are difficult to verify. Therefore, we instead estimate reduced-form RD models that essentially ignore the fuzziness in converting the running

variable into final ratings. The practical implication of this modeling decision is in the interpretation of our estimates – rather than estimating treatment effects of higher ratings, we estimate "intention-to-treat" (ITT) effects. We can still perform standard tests for statistical significance with our ITT estimates, and they will convey similar information to treatment effects in sign subject to scaling, which depends on the degree of fuzziness in the discontinuities. Because the discontinuities at the 3/4 and 4/5 thresholds are close to being sharp, our ITT estimates will not be far off from what we would expect to estimate as treatment effects directly (Gupta, 2011; Hernan and Hernandez-Diaz, 2012).

We construct the multinomial response outcome for each question as follows:

$$Y_i = \begin{cases} 0 \ \textit{negative response} \\ 1 \ \textit{positive response} \\ 2 \ \textit{skip voluntarily} \\ 3 \ \textit{skip due to position change} \end{cases}$$

We then estimate the corresponding RD models for each question at each performance threshold using the specification (for levels 2/3, 3/4 and 4/5 – recall from above that we do not evaluate the 1/2 threshold because very few teachers in Tennessee receive a level-1 rating):

$$\log[\frac{P(Y_i = j \mid Z_i = z_i)}{P(Y_i = 0 \mid Z_i = z_i)}] = \beta_{j0} + X_i\beta_{j1} + [f(S_i)]\beta_{j2} + [f(S_i) * I(S_i \geq T)]\beta_{j3} + [I(S_i \geq T)]\beta_{j4} + \varepsilon_{ji} \quad (1)$$

where $P(Y_i = j \mid Z_i = z_i) = \exp(z_i\beta_j) / [1 + \sum_{h=1}^{3} \exp(z_i\beta_h)]$ for j=1, 2 and 3; and

$P(Y_i = 0 \mid Z_i = z_i) = 1 / [1 + \sum_{h=1}^{3} \exp(z_i\beta_h)]$.

In equation (1), X_i is a vector of observable teacher 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 regardless whether the teacher actually receives a higher rating, and ε_i is the error term, which we cluster at the school level. The *X*-vector in our primary specification includes teacher gender, race, degree level and experience. We also use an expanded *X*-vector that includes school characteristics in supplementary models. Although we considered several functions for $f(S_i)$, in our primary models we specify $f(S_i)$ as a simple linear function of the running variable on both sides of the discontinuity.[10] The parameters of interest are in the vector β_4, which are the ITT effects of receiving a higher rating on responses to the professional development questions relative to the negative baseline (and in particular, β_{14}, which compares positive and negative responses).

4.2 Validation of the RD Design

The RD design offers a credible approach for identifying the causal effects of teacher ratings on professional development choices subject to several assumptions. In this section we review and test these assumptions to provide evidence on the extent to which the RD design can be useful for informing our research question. The validation analysis in this section closely follows Koedel, Li, Springer and Tan (2015).

---

[10] We use both the Akaike information criterion (*AIC*) test and Bayesian information criterion (*BIC*) test to determine the polynomial order for the primary specification. The linear functional form always performs best under the BIC and performs best more than any other functional form under the *AIC*. Adding higher order polynomial terms (up to a quartic) of the running variable to the models does not influence our findings qualitatively.

As noted above, the most important context-specific issue we face in inferring causality from our RD models is that a large fraction of Tennessee teachers did not submit a survey. The key threat to identification is that if teachers' decisions to submit a survey are determined in part by their ratings, the RD estimates conditional on submitting a survey will be biased by attrition from the dataset that is itself caused by treatment. 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) above. The supplementary models are estimated for all Tennessee teachers and the dependent variable is a binary indicator for whether a survey was submitted. The models include the same variables shown in equation (1) and are estimated as linear probability models given that the outcome variable is binary. The purpose of the supplementary models is to determine whether treatment, in this case a higher rating, impacts the likelihood of survey submission. If so, this would suggest a source of sample-selection bias in the estimates obtained from the restricted sample of teachers who submitted a survey.[11]

Table 3 displays the estimated effects of treatment on survey submission at each of the three performance thresholds we evaluate. At each cutoff, the effect is small and statistically insignificant, indicating that the survey participation rate is not caused by ratings, at least subject to the local interpretation of the RD estimates (which is the most relevant interpretation for informing the

[11] 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 whose records are missing key information from these regressions, leaving a total sample size of 57,631 across all three discontinuity thresholds. Appendix Table A.1 shows the reasons that teachers are excluded from these models.

credibility of our main findings below). This in turn suggests that the RD estimates based on the sample of Tennessee teachers who submitted a survey will not be biased by sample selection.[12]

In addition to the survey-submission issue, which is specific to our context, we also perform two general tests that are commonly used to detect potential violations of the RD assumptions. The first test examines whether there are other discontinuities in the data that align with the primary discontinuity at each threshold. If other variables are discontinuous at the main discontinuity threshold, 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 main discontinuities in teachers' evaluation scores, we estimate a series of reduced-form models as follows:

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

[12] As noted above, one reason that teachers did not submit a survey is that they left the Tennessee teaching workforce (see notes to Appendix Table A.1). These teachers are incorporated into the results in Table 3 in the sense that they are treated no differently than teachers who remained in the workforce but did not submit a survey. For the purpose of gaining inference about identification it is not necessary to differentiate between the various reasons that teachers did not submit a survey.

[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.

The dependent variable in equation (2) is now X_i, which is a teacher characteristic. For simplicity, each teacher characteristic is converted into a binary indicator variable and the equation is specified as a linear probability model.[14]

Table 4 presents results from the series of RD regressions linking the rating discontinuities to teacher characteristics. Not all covariates are balanced. For example, at the 3/4 threshold white teachers are more likely to receive a higher rating, and at the 4/5 threshold group-1 teachers (with individual growth scores) are more likely to receive a higher rating. However, it is not obvious that the observed lack of balance in Table 4 is any worse than what would be expected by chance.

We construct a simulation-based test as in Koedel, Li, Springer and Tan (2015) to determine the likelihood of observing the number of unbalanced covariates reported in each column of the table by chance (also see Cullen, Jacob and Levitt, 2005; Fitzpatrick, Grissmer and Hastedt, 2011). To construct the simulations we first split the analytic dataset vertically, separately blocking off teachers' covariates (dependent variables) and underlying scores (independent variables). The critical 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 with the real data. At each iteration of the simulations, we randomly sort the block of teacher scores, then re-connect it to the covariate block to assign each teacher a random rating. Then we estimate the model in equation (2)

[14] For example, we do not code the teacher experience bins shown in Table 1 into a single, multinomial variable – instead, we estimate a separate regression for each bin where the dependent variable is a binary indicator for whether the teacher belongs in that bin.

for each threshold, storing the number of covariates that are unbalanced at the 5-percent level under random assignment. We repeat this procedure 3,000 times to construct empirical distributions of covariate imbalance.

Based on the random-assignment simulations, the bottom of the table reports the probabilities that we would observe at least the number of unbalanced covariates by chance that we actually observe with the real data for each threshold. For the 2/3 threshold, since there are no significant results, the p-value is necessarily 1.0. The p-value is 0.43 at both the 3/4 and 4/5 thresholds. Thus, we conclude that the degree of covariate imbalance in Table 4 is not out of 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, forthcoming; Van Der Klaauw, 2002).

We perform density tests around the three thresholds and report the results in Appendix C. The tests are clean for the 2/3 and 3/4 cutoffs, but indicate bunching at the 4/5 cutoff such that there are significantly more teachers who receive the high rating close to the threshold (at the

5-percent level). This does not necessarily imply that the bunching of the running variable at the 4/5 cutoff is the result of non-random manipulation by individual teacher evaluators, and indeed, McCrary (2008) notes that failure of the density test alone does not provide sufficient grounds to reject the validity of an RD research design. Nonetheless, the bunching around the 4/5 cutoff suggests some caution in interpreting our findings at that threshold.[15]

5. Results

5.1 Primary Results

Table 5 presents results from our primary RD models as specified in equation (1). The table reports the ITT effect of the higher rating at each discontinuity threshold for each survey question, presented as a relative risk ratio coefficient for positive responses relative to negative responses. The relative risk ratios represent changes in the relative probability of a positive response against a negative response after the receipt of treatment. A coefficient above 1.0 indicates that the performance rating caused an increase in the likelihood of a positive response relative to a negative response, and a coefficient of less than 1.0 indicates the opposite. T-statistics are in parentheses for whether the coefficients are significantly different from 1.0, where a value of 1.0 would indicate a null effect.

[15] Koedel, Li, Springer and Tan (2015) show that in the previous evaluation year in Tennessee, there was an *ex post* rounding of scores for teachers near the cutoff, which was implemented without discretion and thus should not cause any bias. It may very well be that teacher scores in our data were also rounded, but if this is the case then the rounding was done prior to our gaining access to the data. We do not have any direct evidence of score rounding during the 2012-2013 evaluation year.

The coefficients reported in the table are estimated within the context of the full multinomial model and thus in conjunction with coefficients for the two types of non-response outcomes as well, where a negative response is always the baseline. We focus most of our discussion on the positive-to-negative-response coefficients shown in Table 5 and briefly discuss the coefficients for the non-response outcomes in Section 5.2 (with results provided in Appendix D).

Table 5 shows that the ITT treatment effects are statistically insignificant at the 5 percent level for all four questions at each of the three thresholds. Moreover, the directions of the coefficients are mixed and their magnitudes are small. For example, taking the estimate for question 4 at the 2/3 threshold at face value would imply that teachers with the higher rating are 1.126 times more likely to respond positively rather than negatively to the question compared to teachers with the lower rating. In terms of a marginal effect, if this estimate were statistically significant it would imply only a 1.3 percentage point increase in the likelihood of a positive response caused by the higher rating, or just a 3-percent increase off of the baseline positive response rate to that question (per Table 2). On the whole, we interpret the results in Table 5 as showing that the assignment of a higher rating itself does not impact teachers' professional development choices.

One potential explanation for our null results is that our models are underpowered. While it is true that our point estimates must be attenuated by the estimation of ITT parameters, at least at the 3/4 and 4/5 thresholds, it is unlikely that the attenuating effect of the fuzziness in our data is substantively important given that all of the discontinuities are very close to sharp (per Figure 1).

Another potential source of power loss is that we evaluate the survey questions in a disaggregated fashion.

It may be that we can improve power in our study by consolidating the information in the professional development questions into a single "professional development index." In results that we ultimately omit for brevity, we pursued this possibility with limited success. The problem is that we were unable to identify a latent construct of professional development that linked the four questions in a meaningful way. More directly, we performed a factor analysis using the four survey questions and found no single factor with an eigenvalue above 0.80. Substantively, this result suggests that the communality shared by the questions is small, and thus the questions do not lend themselves to being summarized by an index or indices of reduced dimension. All of that said, we did perform a full factor analysis and estimated index-based models following the general methodological approach of Koedel, Li, Springer and Tan (2015), and obtained substantively similar findings to what we report in Table 5 – i.e., null results.

5.2 Missing Response Outcomes

In this subsection we briefly discuss our findings for the missing-response outcomes from the multinomial model. Appendix Tables D.1 and D.2 show estimates for each question and each threshold analogous to what we report for the positive-response outcomes in Table 5, but for voluntary non-response and position change non-response outcomes respectively.[16] The coefficients

[16] There are very few voluntary non-responses for question 1 (roughly 2 percent), which led to our removing voluntary non-responders from the analytic sample to address a convergence problem with the multinomial logit. Correspondingly, Table D.1 only presents results for questions 2, 3 and 4 for voluntary non-responses.

shown in the tables are estimated in the same multinomial models as the coefficients in Table 5. The relative risk ratio coefficients are again relative to negative responses. Consistent with the null results in Table 5, Table D.1 and D.2 provide no evidence to suggest that teachers' professional development activities affect the likelihood of either type of non-response relative to a negative response.

5.3 Robustness and Sensitivity

We examine the robustness of the findings from our primary multinomial model in several ways. First, in Appendix Table E.1 we replicate the results in Table 5 after incorporating information about the schooling environment. Although it seems unlikely, it is possible that teachers working in different schooling environments could be rated in such a way that whether a teacher ends up on one side of the cutoff or another near the threshold is associated with school characteristics. In turn, this could bias the estimates in Table 5. When we add a vector of school characteristics to equation (1) that includes the shares of students at each school by gender, race and free/reduced-price lunch status, Table E.3 shows that the findings are essentially unchanged from what we report in Table 5.

We also examine the sensitivity of the findings to alternate bandwidth specifications around the discontinuity thresholds. Our main results in Table 5 use data from all teachers with ratings on either side of the threshold, which includes teachers with scores within 75 points of the cutoff in either direction. In Appendix Tables E.2 through E.4 we consider alternative bandwidths of 60, 50 and 40 points. The tables show that while there are mixed changes in either the magnitude or directions of the coefficients, the null results are robust to shrinking the bandwidth. None of the

estimated coefficients at any of the three thresholds are statistically significant in Tables E.2 through E.4.

A third robustness issue relates to whether all of the teachers in the analytic sample were actually treated. Although all teachers were provided access to their ratings online, approximately 7.5 percent of survey respondents indicate on the survey that they did not see their ratings. We do not do anything differently for these teachers in the main analysis. However, if these teachers truly did not receive their ratings, then for the purpose of our study they never received treatment and their inclusion in the analytic sample will attenuate the results.

In Appendix Table E.5 we again replicate the results in Table 5, but this time we exclude teachers who indicate that they had not seen their ratings. All the estimates remain statistically insignificant, further supporting our finding of null effects of rating assignments on teachers' professional development activities.

6. An Alternative Explanation and Exploratory Analysis

The descriptive statistics reported in Table 2 indicate non-negligible gaps in the intensity and responsiveness of professional development activities across teachers with different overall ratings. Our regression discontinuity analysis uncovers no evidence to suggest that ratings per se are a causal contributor to these gaps, but this leaves open many possibilities. In this section we discuss one possibility in particular – that teachers are responsive in their professional development choices to the feedback they receive during classroom observations. We offer an exploratory analysis consistent with this explanation.

Before going into the details of our exploratory analysis, we first note that the possibility that feedback from classroom observations at least partly driving the gaps across rating levels shown in Table 2 is not ruled out by our regression discontinuity findings (or lack thereof). The reason is that while there are large gaps in the qualitative feedback that teachers receive across rating levels, as would be expected, there are not discontinuities in qualitative feedback that align with the discontinuities in teachers' overall ratings. We make these points in Tables 6 and 7 drawing on supplementary, qualitative feedback data from classroom observations provided to teachers who were evaluated using the TEAM model (who comprise 82 percent of our full analytic sample, per Table 1).[17]

As part of the evaluation process in TEAM, teacher observers report on one area for "reinforcement" (ENFORCE) and one area for "refinement" (REFINE) for each teacher. An area of reinforcement is something that is working well in the classroom, while an area for refinement is where there is room for improvement. Pre-specified lists are provided to observers from which to choose the ENFORCE and REFINE areas, and observers are additionally given space to provide ENFORCE and REFINE comments. It is worth noting that while our measure of feedback is based on the actual written feedback regarding an area of strength and an area for improvement, a teacher also receives feedback through formal and informal conversations with the observer.

[17] Qualitative feedback data are unavailable from TDOE for teachers evaluated using the other, smaller evaluation models used by some Tennessee school districts. We also exclude about 15 percent of TEAM teachers from this supplementary analysis because qualitative feedback data are missing.

Table 6 is analogous to Table 2 but restricted to TEAM teachers. It shows differences across overall rating levels in one (admittedly rough) measure of the intensity of the qualitative feedback that teachers receive - the average number of words in the comment section for each ENFORCE and REFINE area. Our word count strategy is modeled after similar strategies used by political scientists to extract policy positions from political texts (Laver, Benoit, and Garry, 2003). A clear pattern in the table is that lower-rated teachers have shorter ENFORCE entries and longer REFINE entries. The average word-count difference (ENFORCE – REFINE) is statistically different across all rating levels. Thus, Tables 2 and 6 combine to show that teachers with different overall ratings in the Tennessee system make different professional development choices (Table 2) and receive different qualitative feedback from the Tennessee system (Table 6).

Table 7 illustrates why our RD design is not informative about the causal relationship between qualitative feedback and teachers' professional development activities. The table reports estimates from linear regression discontinuity models where the dependent variables are (a) the average number of ENFORCE words, (b) the average number of REFINE words, and (c) the average of the difference. Unsurprisingly, the discontinuity thresholds that determine teachers' overall ratings as shown in Figure 1 do not align with discontinuities in the feedback that teachers receive (note: this result is consistent with the RD assumptions described in Section 4.2; put differently, non-zero results in Table 7 would undermine the credibility of the RD analysis).

Although our data do not lend themselves to the design of a credible causal analysis of how differences in feedback affect teachers' professional development choices, in Tables 8 and 9 we

provide evidence of what is, at the least, a strong association. Table 8 mimics Table 6, but this time we show differences in our ENFORCE and REFINE word-count measures when teachers are grouped by their *classroom observation scores only* – which we aggregate in the same way that the final rating is calculated – rather than their overall scores as in Table 6.[18] Note that the differences across rating levels widen in Table 8 relative to Table 6, which intuitively shows that the feedback teachers receive from their classroom observations is more closely aligned with their classroom observation scores than their overall scores.

Next, Table 9 is an analog to Table 2, but where teachers are grouped by their classroom observation scores rather than overall ratings. If feedback from classroom observations is more highly correlated with the classroom observation score than the overall score, as is indicated by Table 8, and if this feedback is indeed contributing to teachers' professional development choices, then a widening of the gaps in Table 9 relative to Table 2 should be observed. This is precisely what we see. For example, whereas the gap in positive responses to question 1 in Table 2 between teachers with overall ratings of (2) and (5) is 10 percentage points (0.63 – 0.53), the same gap between teachers with observation ratings of (2) and (5) in Table 9 is 14 percentage points (0.66-0.52). A similar comparison for question 2 yields gaps of 15 and 19 percentage points,

[18] Recall from above that the "classroom observation" score can include other information for some teachers. However, per our naming convention, the score is primarily driven by ratings for teachers based on classroom observations. Consistent with this assertion, the tables that follow document a strong correlation between the classroom observation score and the qualitative feedback that teachers receive from their observers.

respectively, and the general pattern of greater differentiation across rating levels within questions is upheld throughout Table 9.

Although inference from the exploratory analysis we show here is limited by our lack of a research design that can support causal inference, we document evidence of an important association. These results suggest an area of focus for future research aimed at understanding how emerging teacher evaluation systems can be used to promote teacher self-improvement.

7. Conclusion

There is substantial variation in teacher quality throughout the workforce and access to high-quality teachers has been shown to be a critical determinant of student success in school and beyond (Chetty, Friedman and Rockoff 2014a, 2014b; Hanushek and Rivkin, 2010; Kane et al., 2013). The educational and economic importance of teacher quality motivates the development of policies aimed at measuring and improving the effectiveness of teachers. Indeed, the process of implementing new, more rigorous teacher evaluations is moving forward or being considered in a number of states and school districts across the United States, with Tennessee being at the forefront of education policy in this regard.

The more detailed performance information provided to teachers in emerging systems like the one in Tennessee may improve workforce quality in a number of ways, including by changing teachers' professional development choices. A number of mechanisms could drive teachers to change, ranging from the provision of better information to teachers about how to improve (Taylor

and Tyler, 2012) to the presence of implicit or explicit rewards and sanctions in new systems that are tied to performance (Boyd et al., 2011; Dee and Wyckoff, 2015; Vigdor, 2009).

The objective of the present study is to use the Tennessee experience to learn about how rigorous evaluations influence teachers' professional development activities broadly defined. Descriptively, we show that there are large differences in both the intensity of teachers' professional development activities and the extent to which these activities are associated with evaluation feedback across teachers with different overall ratings in the Tennessee system. Lower-rated teachers are significantly more likely to report more intensive and responsive professional development. Our primary analytic models leverage discontinuities in how the Tennessee evaluation system converts underlying scores into final ratings to determine whether the assignment of ratings itself influences the gaps in professional development that we document descriptively. Previous studies in other locales, and in Tennessee, have identified substantive effects of rating assignments on attrition and measured performance (Dee and Wyckoff, 2015), and perceptions of work (Koedel, Li, Tan and Springer, 2015). However, our results provide no indication that teachers' professional development choices are influenced by rating assignments per se. That is, all else equal, teachers who receive higher ratings are no more or less likely to be engaged in intense and feedback-responsive professional development compared with their counterparts who receive lower ratings.

We extend our analysis to provide suggestive evidence on another hypothesis: that teachers are responding to the feedback that they receive during classroom observations when selecting professional development activities. Although our data and methods are not suited to construct a

causal test of this hypothesis, our exploratory findings suggest that this is an area worthy of further exploration. Specifically, we show that classroom observation scores are more closely aligned with the qualitative feedback that teachers receive from their observers than overall ratings, and that the gaps in teachers' professional development choices across rating categories widen when we define the categories based on the classroom observation component only, rather than the overall rating.

The mechanism through which teachers receive feedback is an important area for future study. While one can easily imagine that a school principal as observer can play an important role in driving professional development decisions, and ultimately instructional practice, it is not obvious how additional feedback from school leaders influences the quantity or types of professional development their teachers take. Are principals who are more engaged in the evaluation process better equipped to connect teachers to professional development relevant to their needs? Does additional feedback from school leaders help teachers get more out of existing professional development options? Our work begins to provide an important foundation for understanding the relationship between rigorous teacher evaluation systems and professional development activities.

References

Angrist, Joshua D. 2001. Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors Simple Strategies for Empirical Practice. *Journal of Business & Economic Statistics* 19(1), 2-16.

Boyd, Don, Hamilton Lankford, Susanna Loeb, Matthew Ronfeldt, and James Wyckoff. 2011. The Role of Teacher Quality in Retention and Hiring: Using Applications to Transfer to Uncover Preferences of Teachers and Schools. *Journal of Policy Analysis and Management* 30(1), 88-100.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff. 2014a. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review* 104(9), 2593-2632.

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

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.

Dee, Thomas and James Wyckoff. 2013. Incentives, Selection, and Teacher Performance: Evidence from IMPACT. NBER working paper.

Drake, T.A., Goldring, R., Grissom, J.A., Cannata, M.A., Neumerski, C.M., Rubin, M., and Schuermann, P. (2015). In *Improving Teacher Evaluation Systems: Making the Most of Multiple Measures*, eds. Jason A. Grissom and Peter Youngs. New York: Teachers College Press.

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.

Gupta, Sandeep K. 2011. Intention-to-treat concept: A review. *Perspectives in clinical research* 2(3), 109-112.

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. and Steven G. Rivkin. 2010. Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review* 100(2), 267-271.

Harris, Douglas N. and Tim R. Sass. 2011. Teacher Training, Teacher Quality and Student Achievement. *Journal of Public Economics* 95(7/8), 798-812.

Hernan, Miguel A. and Sonia Hernandez-Diaz. 2012. Beyond the intention-to-treat in comparative effectiveness research. *Clinical Trials* 9(1), 48-55.

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

Jacob, Brian. 2005. Accountability Incentives and Behavior: The Impact of High Stakes Testing in the Chicago Public Schools. *Journal of Public Economics* 89(5), 761-796.

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

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller and Douglas O. Staiger. 2013. Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Seattle, WA: Bill and Melinda Gates Foundation.

Koedel, Cory, Jiaxi Li, Matthew Springer and Li Tan. 2015. The Impact of Performance Ratings on Job Satisfaction for Public School Teachers. Working Paper.

Laver, M., Benoit, K., and Garry, J. (2003). Extracting Political Positions from Political Texts Using Words as Data. *American Political Science Review*, 97(2), 311-331.

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

Lee, David S. and Thomas Lemieux. 2009. Regression Discontinuity Designs in Economics. *Journal of Economic Literature* 48(2), 281-355.

McCrary, Justin. 2008. Manipulation of the Running Variable in the Regression Discontinuity Design: a Density Test. *Journal of Econometrics* 142(2), 698-714.

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.

Rockoff, Jonah E, Douglas O. Staiger, Thomas J. Kane and Eric S. Taylor. 2012. Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools. *American Economic Review* 102 (7), 3184-3213.

Stutz, Terrence. 2014. Texas Teachers, Principals Ignore Survey Ordered by Legislature. Dallas News (08.03.2014). http://www.dallasnews.com/news/education/headlines/20140803-texas-teachers-principals-ignore-survey-ordered-by-legislature.ece

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

The New Teacher Project. 2012. The Irreplaceables: Understanding the Real Retention Crisis in America's Urban Schools. Brooklyn, NY.

The New Teacher Project. 2015. The Mirage: Confronting the Hard Truth About Our Quest for Teacher Development. Brooklyn, NY.

United States Government Accountability Office. 2013. Race to the Top: States Implementing Teacher and Principal Evaluation Systems Despite Challenges. Washington, DC.

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

Vigdor, Jacob L. 2009. Teacher Salary Bonuses in North Carolina. *Performance Incentives: Their Growing Impact on American K-12 Education*, edited by Matthew Springer, Brookings, 2009.

Watts, Michael and William E. Becker. 2008. A Little More than Chalk and Talk: Results from a Third National Survey of Teaching Methods in Undergraduate Economics Courses. *The Journal of Economic Education* 39(3), 273-286.

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.

Table 1. Descriptive Statistics for the Universe and Analytic Samples.

Variable	All Tennessee Teachers	Analytic Sample
Female teacher	0.80	0.83
Black teacher	0.12	0.11
White teacher	0.87	0.88
Other race	0.01	0.00
Bachelor degree	0.41	0.38
Education specialist	0.07	0.09
Master degree	0.42	0.43
Other education	0.10	0.10
Years of experience: 0-1	0.12	0.10
Years of experience: 2-4	0.13	0.12
Years of experience: 5-9	0.21	0.20
Years of experience: 10-14	0.18	0.19
Years of experience: 15-19	0.12	0.15
Years of experience: 20+	0.23	0.24
Group 1 (with individual growth score)	0.45	0.47
AFET evaluation model	0.00	0.00
COACH evaluation model	0.06	0.00
TEAM evaluation model	0.77	0.82
TEM evaluation model	0.14	0.15
TIGER evaluation model	0.02	0.03
Level 1 (Sig. Below Expectations)	0.02	0.01
Level 2 (Below Expectations)	0.09	0.08
Level 3 (At Expectations)	0.22	0.22
Level 4 (Above Expectations)	0.33	0.34
Level 5 (Sig. Above Expectations)	0.33	0.36
N	68171	24009

Notes: Among all Tennessee teachers, just 2.2 percent of teachers received an evaluation rating of "1," which indicates significantly below expectation (1.2 percent in the analytic survey sample). Group-1 teachers are those with an individual growth score as a component of the final performance measure. An education specialist degree is an advanced terminal degree for individuals who wish to go beyond the MA level but do not wish to pursue a doctorate. As noted in the text, some teachers in the "full universe sample" are missing some of the information reported in this table (see Appendix Table A.1 for details). Teachers with missing information are omitted from the calculations on an item-by-item basis.

Table 2. Proportion of Positive Responses to each Question.

	Below Expectation (2)	At Expectation (3)	Above Expectation (4)	Significantly Above Expectation (5)
Question 1: Spent time improving instructional skills	0.63	0.59	0.54	0.53
Question 2: Changed teaching based on evaluation results	0.66	0.61	0.56	0.51
Question 3: Changed non-teaching duties based on evaluation results	0.26	0.23	0.21	0.19
Question 4: Evaluation feedback influenced prof. development activities	0.43	0.41	0.39	0.37
N	1852	5302	8103	8556

Notes: See Appendix B for more information about how positive and negative responses are coded for each question. The table reports the share of positive responses where the denominator is the total number of submitted surveys within each rating level. Appendix Table A.2 reports analogous positive-response shares where the denominator excludes teachers who skipped the question on a question-by-question basis.

Table 3. Regression Discontinuity Estimates of the Effects of Ratings on Survey Submission.

Dependent Variable	Level 2-3	Level 3-4	Level 4-5
Response Rate	-0.004 (0.015)	-0.005 (0.012)	0.004 (0.010)
N	17662	32175	39055

***/**/* 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. None of the estimates is statistically significant at the 10-percent level or higher.

Table 4. Regression Discontinuity Estimates of the "Effects" of Ratings on Teacher Characteristics that Should Not be Affected.

Variable	Level 2-3	Level 3-4	Level 4-5
Female	0.012	0.002	0.017
	(0.020)	(0.013)	(0.012)
African American	-0.001	-0.020	-0.003
	(0.017)	(0.011)*	(0.011)
White	0.006	0.024	0.002
	(0.018)	(0.012)**	(0.011)
BA degree	-0.018	-0.002	0.021
	(0.024)	(0.017)	(0.014)
MA degree	-0.006	0.018	0.006
	(0.024)	(0.016)	(0.015)
Educational specialist	0.004	-0.005	-0.000
	(0.013)	(0.010)	(0.009)
Years of experience: 0-1	-0.025	-0.019	0.000
	(0.018)	(0.012)*	(0.008)
Years of experience: 2-4	0.007	0.005	0.003
	(0.017)	(0.012)	(0.010)
Years of experience: 5-9	0.011	0.004	0.009
	(0.019)	(0.013)	(0.012)
Years of experience: 10-14	0.015	0.024	-0.003
	(0.019)	(0.013)*	(0.012)
Years of experience: 15-19	0.021	0.003	0.004
	(0.018)	(0.012)	(0.011)
Group-1	-0.027	0.017	0.071
	(0.026)	(0.019)	(0.016)***
Overall p-value	1.00	0.43	0.43
N	7154	13405	16644

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

Note: Models are specified as linear probability models. Standard errors are clustered at the school level and reported in parentheses. P-values indicate the probability of obtaining the observed number of statistically significant coefficients by chance in each column (at the 5-percent level based on 3,000 bootstrap repetitions). See notes from Table 1 for details about the variables listed here.

Table 5. Effects of Ratings on Self-Reported Professional Development Choices.

Dependent Variable	Level 2-3	Level 3-4	Level 4-5
Question 1	1.095	0.968	1.005
	(0.76)	(-0.41)	(0.06)
Question 2	0.953	0.990	1.028
	(-0.35)	(-0.11)	(0.34)
Question 3	0.870	1.036	1.084
	(-1.13)	(0.39)	(0.99)
Question 4	1.126	1.096	0.970
	(1.04)	(1.15)	(-0.45)
N	7154	13405	16644

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

Notes: Question numbers correspond to those in Table 2. Models are specified as multinomial logistic regressions. Each estimate in each cell comes from a separate regression. The values in each cell are relative risk ratios where the baseline comparison outcome is a negative response. T-statistics for the multinomial logit coefficients are reported in parentheses and as in the preceding analysis, clustering is at the school level. The number of observations used for the question-1 regressions is smaller than what is reported in the final row of the table because we exclude voluntary non-responses from the analytic sample for that question, per the discussion in the text.

Table 6. Average Word Counts in the ENFORCE and REFINE Comment Sections from Teachers' Classroom Observations, by Teachers' Overall Rating Levels.

	Below Expectation	At Expectation	Above Expectation	Significantly Above Expectation
	(2)	(3)	(4)	(5)
Average: Reinforcement	26.07	27.23	27.40	28.16
Average: Refinement	32.13	29.68	28.93	27.14
Average: Difference	**-6.06**	**-2.45**	**-1.52**	**1.02**
N	1408	3900	5566	5772

Notes: As discussed in the text, our sample for this supplementary analysis is limited to teachers who were evaluated using the TEAM model, which is the predominant evaluation model in Tennessee. We also exclude 15 percent of TEAM teachers for whom qualitative feedback data are unavailable. The average differences in the word counts reported in the final row of the table are statistically different across all rating levels.

Table 7. Tests for whether Discontinuities in the Qualitative Feedback Received by Teachers (Measured by Word Counts) Coincide with the Discontinuities that Convert Underlying Scores into Final Overall Ratings.

Dependent Variable	Level 2-3	Level 3-4	Level 4-5
Average: Re-enforcement	0.360	-1.517	0.201
	(0.35)	(-1.69)*	(0.23)
Average: Refinement	0.363	-0.684	0.207
	(0.26)	(-0.65)	(0.23)
Average: Difference	**-0.003**	**-0.833**	**-0.006**
	(0.00)	**(-1.51)**	**(-0.01)**
N	5308	9466	11323

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

Notes: T-statistics are reported in parentheses and as in the preceding analysis, clustering is at the school level. As discussed in the text, our sample for this supplementary analysis is limited to teachers who were evaluated using the TEAM model, which is the predominant evaluation model in Tennessee. We also exclude 15 percent of TEAM teachers for whom qualitative feedback data are unavailable.

Table 8. Average Word Counts in the ENFORCE and REFINE Comment Sections from Teachers' Classroom Observations, Grouping Teachers based on their Scores on the Classroom Observation Component of the Rating Only.

	Obs. Rating 2.00-2.74	Obs. Rating 2.75-3.49	Obs. Rating 3.50-4.24	Obs. Rating 4.25-5.00
	(2)	(3)	(4)	(5)
Average: Reinforcement	28.36	27.42	27.22	28.02
Average: Refinement	43.68	33.53	28.26	25.62
Average: Difference	**-15.32**	**-6.11**	**-1.03**	**2.40**
N	346	3405	7694	5255

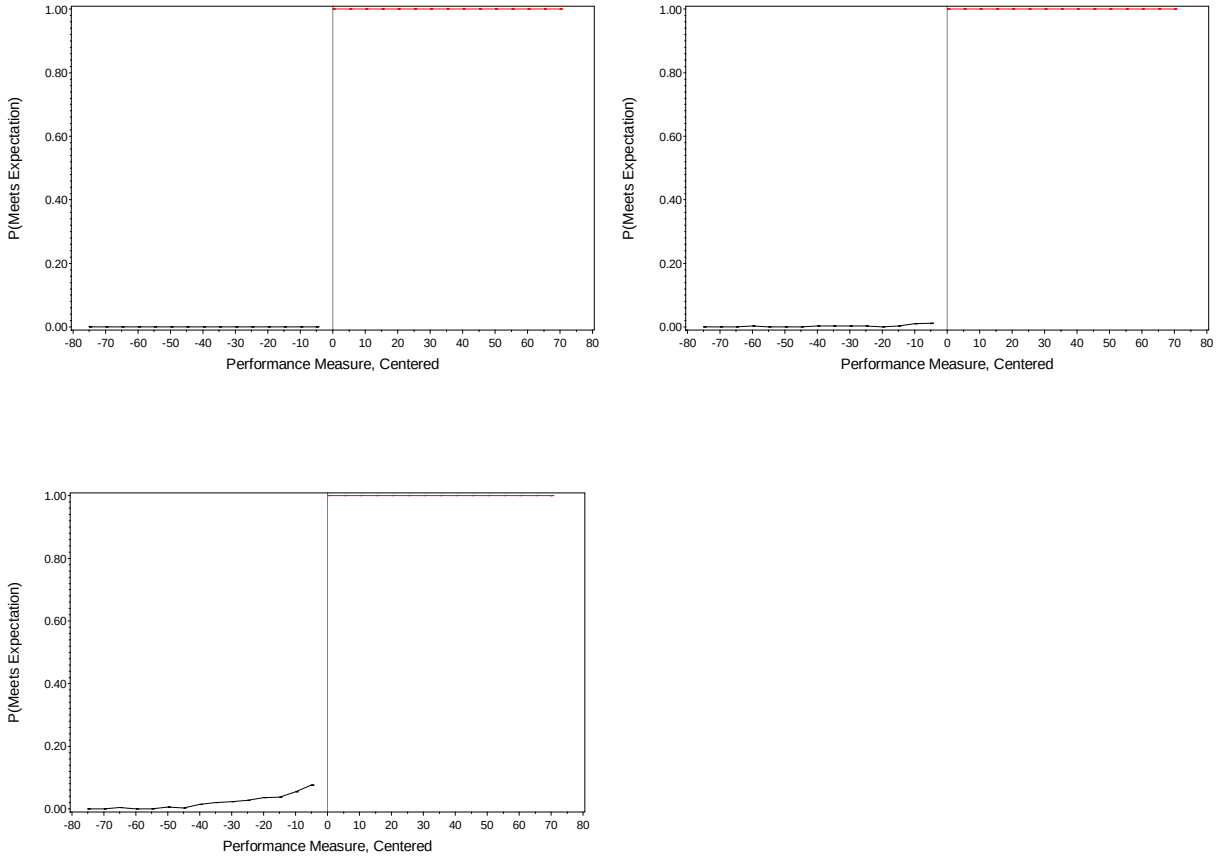
Notes: As discussed in the text, our sample for this supplementary analysis is limited to teachers who were evaluated using the TEAM model, which is the predominant evaluation model in Tennessee. We also exclude 15 percent of TEAM teachers for whom qualitative feedback data are unavailable. The average differences in the word counts reported in the final row of the table are statistically different across all classroom-observation rating levels.

Table 9. Proportion of Positive Responses to each Professional Development Question, Grouping Teachers based on their Scores on the Classroom Observation Component of the Rating Only.

	Obs. Rating 2.00-2.74	Obs. Rating 2.75-3.49	Obs. Rating 3.50-4.24	Obs. Rating 4.25-5.00
	(2)	(3)	(4)	(5)
Question 1	0.66	0.59	0.56	0.52
Question 2	0.68	0.64	0.58	0.49
Question 3	0.34	0.26	0.21	0.17
Question 4	0.48	0.42	0.39	0.36
N	471	4556	10879	7956

Notes: This table is analogous to Table 2 and is based on data from the full analytic sample.

Figure 1. Illustration of Sharp Discontinuities at the Cut Scores Between Levels 2/3, 3/4 and 4/5.



Notes: We aggregate teachers 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 five points above it.

Appendix A
Data Appendix

Table A.1. Construction of the Analytic Sample.

Total Number of Records in the TN Evaluation Dataset		68171
	Records Deleted	Remaining sample
Evaluated using COACH model	-3992	64179
Missing final rating	-6009	58170
Missing indicator for group-1 status	-536	57634
Assigned rating and calculated rating (from the running variable) are not consecutive	-3	57631
Did Not Take the First to the Top Survey	-33622	24009

Table A.2. Proportion of Positive Responses to Each Question Conditional on Non-Missing Response.

	Below Expectations (2)	Meets Expectations (3)	Above Expectations (4)	Sig. Above Expectations (5)
Question 1: Spent time improving instructional skills	0.71	0.69	0.64	0.66
Question 2: Changed teaching based on evaluation results	0.82	0.78	0.74	0.69
Question 3: Changed non-teaching duties based on evaluation results	0.32	0.29	0.28	0.27
Question 4: Evaluation feedback influenced prof. development activities	0.53	0.52	0.52	0.50

Note: This table is analogous to Table 2 in the text, but conditions on non-missing responses (that is, all missing responses are removed from the denominator in each cell).

Appendix B
Coding and Other Details for the Professional Development Questions

Below we show each question listed in Table 2 in full form as it was presented to teachers.[1] The options in italics are the ones that we coded to indicate a positive response. Non-italicized options were coded to indicate a negative response. The coding choices are based in part on the underlying distribution of teacher answers to each question. 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.*

Q3. I changed how I perform non-teaching duties based on my evaluation results.[2]

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

[1] With the exception of Question 1, which is actually 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.

[2] 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 from other online sources also include parent and teacher meetings, curriculum development, etc.

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

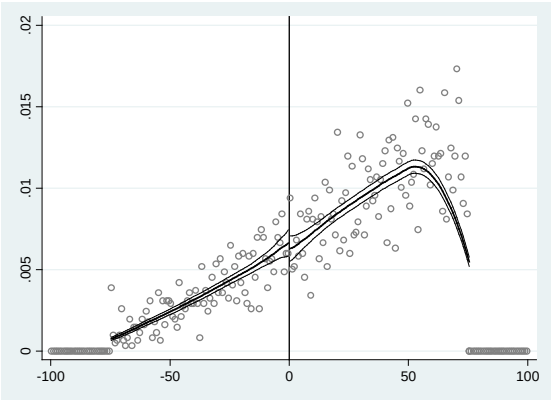
1. Strongly Disagree.
2. Disagree.
3. *Agree.*
4. *Strongly Agree.*

Appendix C
Density Tests

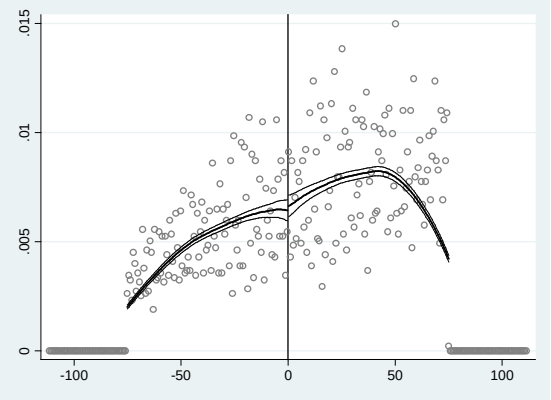
Figure C.1 shows the densities of the running variable. The densities are smooth through the 2/3 and 3/4 thresholds but significant bunching at the 5-percent level is present at the 4/5 cutoff.

Table C.1 reports regression results corresponding to the figures.

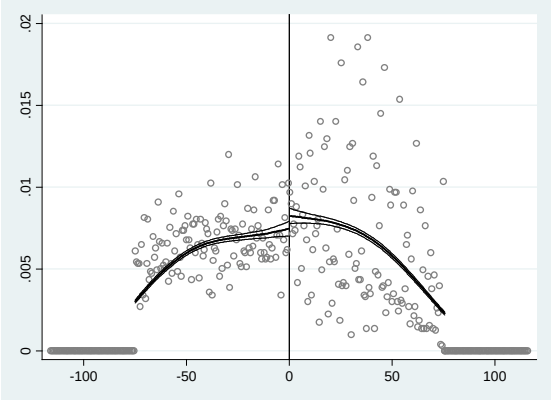
Figure C.1. Densities of Teachers' Underlying Score Variables Centered on the Actual Scores that Determined Teachers' Ratings.



(a) Level 2/3



(b) Level 3/4



(c) Level 4/5

Table C.1. Estimated Discontinuities Using McCrary's Method at Different Cutoff Levels.

	Level 2-3	Level 3-4	Level 4-5
Using actual assignment variable	-0.064	0.027	0.100
	(0.092)	(0.055)	(0.042)**

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

Appendix D
Results for Two Types of Item Non-Response Relative to Negative Response

Table D.1. Effects of Ratings on Teacher Voluntary Non-Response to Survey Questions.

Dependent Variable	Level 2-3	Level 3-4	Level 4-5
Question 2	0.943	1.029	1.185
	(-0.30)	(0.21)	(1.47)
Question 3	0.942	1.108	1.158
	(-0.36)	(0.88)	(1.48)
Question 4	1.125	1.056	1.164
	(0.67)	(0.45)	(1.48)
N	7154	13405	16644

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

Notes: Question numbers correspond to those in Table 2. Models are specified as multinomial logistic regressions. Each estimate in each cell comes from a separate regression. The values in each cell are relative risk ratios where the baseline comparison outcome is a negative response. T-statistics for the multinomial logit coefficients are reported in parentheses. As described in the text, there are very few voluntary non responses to Question 1 (approximately 2 percent of the sample). This creates convergence issues in the model and we deal with this issue by dropping all voluntary non-responders to question 1 from the analytic sample.

Table D.2. Effects of Ratings on Teacher Non-Response to Survey Questions Due to Position Change.

Dependent Variable	Level 2-3	Level 3-4	Level 4-5
Question 1	1.240	1.082	0.887
	(1.07)	(0.67)	(-1.17)
Question 2	1.123	1.098	0.915
	(0.55)	(0.73)	(-0.81)
Question 3	1.118	1.125	0.918
	(0.61)	(1.08)	(-0.91)
Question 4	1.255	1.155	0.888
	(1.22)	(1.26)	(-1.21)
N	7154	13405	16644

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

Notes: Question numbers correspond to those in Table 2. Models are specified as multinomial logistic regressions. Each estimate in each cell comes from a separate regression. The values in each cell are relative risk ratios where the baseline comparison outcome is a negative response. T-statistics for the multinomial logit coefficients are reported in parentheses.

Appendix E
Robustness Analyses

Robustness to Controlling for Schooling Context

Table E.1 shows results analogous to the results in Table 5 but where we also include school-level control variables for the shares of students by race, gender, and free/reduced-price lunch eligibility in the models. The findings are very similar to what we report in Table 5, providing no indication that the RD estimates are biased by systematic ratings-assignment differences around the thresholds that align with factors related to schooling environments.

Table E.1. Replication of Table 5 with Additional Controls for Schooling Context.

Dependent Variable	Level 2-3	Level 3-4	Level 4-5
Question 1	1.096	0.968	1.009
	(0.77)	(-0.40)	(0.13)
Question 2	0.948	0.986	1.032
	(-0.38)	(-0.15)	(0.39)
Question 3	0.870	1.034	1.089
	(-1.13)	(0.37)	(1.05)
Question 4	1.126	1.099	0.973
	(1.04)	(1.18)	(-0.40)
N	7154	13405	16644

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

Notes: Question numbers correspond to those in Table 2. Models are specified as multinomial logistic regressions. Each estimate in each cell comes from a separate regression. The values in each cell are relative risk ratios where the baseline comparison outcome is a negative response. T-statistics for the multinomial logit coefficients are reported in parentheses.

Robustness to Alternative Bandwidths

Tables E.2 through E.4 show results from the main specification for each question at each threshold with narrower bandwidths around the discontinuity. The primary results in Table 5 use all teachers with scores on either side of each rating cutoff, which corresponds to a bandwidth of 75 in each direction. Here we consider bandwidths of 60, 50 and 40. The estimates based on the full bandwidths, as reported in Table 5, are displayed in the first row of each table for comparison.

Table E.2 shows results for all three alternative bandwidths for the 2/3 threshold. Table E.3 presents analogous information at the 3/4 threshold, and Tables E.4 at the 4/5 threshold. In Tables E.2 and E.3, several coefficients change in sign as we shrink the bandwidth from 75 to 40. However, none of the differences imply statistically significant changes to our main findings.

Table E.2. Robustness of Findings to Alternative Bandwidths. Cutoff for Ratings 2/3.

Bandwidth	N	Q1	Q2	Q3	Q4
Full	7154	1.095	0.953	0.870	1.125
		(0.76)	(-0.35)	(-1.13)	(0.87)
$\|S_i\| \leq 60$	5758	1.045	0.907	0.872	1.045
		(0.34)	(-0.65)	(-1.01)	(0.36)
$\|S_i\| \leq 50$	4749	1.033	0.894	0.856	0.979
		(0.23)	(-0.66)	(-1.06)	(-0.16)
$\|S_i\| \leq 40$	3788	1.085	1.023	---	0.992
		(0.50)	(0.12)		(-0.06)

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

Notes: Question numbers correspond to those in Table 2. The estimates in the first row of the table are taken directly from Table 5 for comparison. Models are specified as multinomial logistic regressions. Each estimate in each cell comes from a separate regression. The values in each cell are relative risk ratios for positive responses against negative responses. T-statistics for the multinomial logit coefficients are reported in parenthesis. As in the preceding analysis, standard errors are clustered at the school level. Note that for question 3 the model does not converge when the bandwidth shrinks to 40.

Table E.3. Robustness of Findings to Alternative Bandwidths. Cutoff for Ratings 3/4.

Bandwidth	N	Q1	Q2	Q3	Q4
Full	13405	0.968	0.990	1.036	1.096
		(-0.41)	(-0.11)	(0.39)	(1.15)
$\|S_i\| \leq 60$	11018	0.950	0.936	1.039	1.134
		(-0.58)	(-0.64)	(0.38)	(1.20)
$\|S_i\| \leq 50$	9270	0.899	0.881	1.025	1.110
		(-1.10)	(-1.14)	(0.23)	(1.07)
$\|S_i\| \leq 40$	7429	0.981	0.880	0.981	1.094
		(-0.18)	(-1.01)	(-0.16)	(0.81)

Notes: See notes from Table E.2.

Table E.4. Robustness of Findings to Alternative Bandwidths. Cutoff for Ratings 4/5.

Bandwidth	N	Q1	Q2	Q3	Q4
Full	16644	1.005	1.028	1.084	0.970
		(0.06)	(0.34)	(0.99)	(-0.45)
$\|S_i\| \leq 60$	14319	1.029	1.052	1.133	0.970
		(0.36)	(0.58)	(1.42)	(-0.41)
$\|S_i\| \leq 50$	12270	1.112	1.038	1.096	0.974
		(1.26)	(0.39)	(0.98)	(-0.33)
$\|S_i\| \leq 40$	9970	1.148	1.085	1.066	0.924
		(1.44)	(0.76)	(0.62)	(-0.86)

Notes: See notes from Table E.2.

Robustness to Excluding Teachers Who Report Not Receiving a Rating

Table E.5 reports results analogous to Table 5 after restricting the sample to exclude teachers who indicated that they did not receive their evaluation ratings from the system. As reported in the text, approximately 7.5 percent of teachers in the analytic dataset (n=1,783) indicated that they did not receive their rating. Also note that some teachers did not answer the question about whether they received a rating – we only exclude teachers in Table E.5 who explicitly indicated that they did not receive their ratings. Although in principle our inclusion of teachers who may not have seen their ratings in the main regressions should cause some attenuation bias in our findings, Table E.5 shows that even if we exclude these teachers our null results are comfortably retained.

Table E.5. Effects of Ratings on Professional Development Choices for Teachers Who Report Receiving their Ratings.

Dependent variable	Level 2-3	Level 3-4	Level 4-5
Question 1	1.050	0.959	1.008
	(0.38)	(-0.49)	(0.11)
Question 2	0.951	0.991	1.010
	(-0.33)	(-0.09)	(0.12)
Question 3	0.922	1.020	1.067
	(-0.61)	(0.21)	(0.78)
Question 4	1.151	1.098	0.981
	(1.13)	(1.10)	(-0.27)
N	6389	12211	15657

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

Notes: Question numbers correspond to those in Table 2. Models are specified as multinomial logistic regressions. Each estimate in each cell comes from a separate regression. The values in each cell are relative risk ratios where the baseline comparison outcome is a negative response. T-statistics for the multinomial logit coefficients are reported in parenthesis. As in the preceding analysis, standard errors are clustered at the school level.