

WORKING PAPER

The Impact of Performance Ratings on Job Satisfaction for Public School Teachers

OCTOBER 2014

CORY KOEDEL
JIA XI LI
MATTHEW G. SPRINGER

The Tennessee Consortium on Research, Evaluation, and Development is a research and policy center at Peabody College of Vanderbilt University, which is funded through the State of Tennessee's Race to the Top grant.

Please visit www.tnconsortium.org to learn more about our program of research and recent publications.

JEL CODES: I20, J45, J48

Keywords: personnel evaluation; teacher evaluation; job satisfaction; teacher quality; regression discontinuity in multinomial model

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, 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. Koedel is in the Department of Economics and Truman School of Public Affairs, and Li is in the Department of Economics, at the University of Missouri. Springer is in the Peabody College of Education and Human Development at Vanderbilt University. The usual disclaimers apply.

Tennessee Consortium on Research, Evaluation, and Development

Peabody #44 • 230 Appleton Place • Nashville, Tennessee 37203

PHONE: 615-322-5538 FAX: 615-322-6018

www.tnconsortium.org

The Impact of Performance Ratings on Job Satisfaction for Public School Teachers

Cory Koedel
Jiaxi Li
Matthew G. Springer

October 2014

Abstract

Spurred by the federal Race to the Top competition, the state of Tennessee implemented a comprehensive statewide educator evaluation system in 2011. The new system is designed to increase the rigor of evaluations and better differentiate teachers based on performance. We merge teacher performance evaluations from the new system with data from post-evaluation teacher surveys to examine the effects of teacher ratings on job satisfaction. Using a regression-discontinuity design, we show that ratings causally affect teachers' perceptions of work – all else equal, teachers who receive higher ratings are made to be more satisfied with teaching. Our findings offer the first causal evidence of which we are aware on the relationship between performance ratings and job satisfaction for teachers, and have important policy implications in light of the move toward more rigorous teacher evaluation systems nationally.

1. Introduction

State and local education agencies across the United States are working to improve teacher quality through the adoption of rigorous teacher evaluation systems.¹ The development of these systems is motivated in part by a large body of research showing that teachers differ dramatically in their effectiveness (for a recent overview see Hanushek and Rivkin, 2010). Beyond directly affecting students' short-term academic success, more effective teachers have also been shown to positively affect later-life outcomes for students including wages, college attendance, and teenage childbearing (Chetty, Friedman and Rockoff, 2014). The variation in teacher quality that has been consistently documented in empirical research is not reflected in the traditional teacher evaluation systems that are used in most school districts in the United States, which tend to assign similar and favorable performance ratings to the overwhelming majority of teachers (Weisberg et al., 2009). A rationale for developing new, more-rigorous evaluation systems is to construct ratings that better reflect empirically-observed performance differences across teachers, which can subsequently be used to improve workforce quality in a number of ways (Boyd et al., 2011; Condie, Lefgren and Sims, 2014; Dee and Wyckoff, 2013; Ehlert et al., 2014; Goldhaber and Theobald, 2013; Hanushek, 2009; Springer, Rodriguez and Swain, 2014; Taylor and Tyler, 2012; Winters and Cowen, 2013).

In many respects, Tennessee has become a national leader for the design and implementation of teacher evaluation systems.² In January 2010, the Tennessee General Assembly passed Senate Bill 5, also known as the First to the Top (FTTT) Act, reforming dozens of areas of

¹ A number of states have enacted legislation mandating performance-based evaluations for teachers and high stakes have been attached in some cases – examples include Senate Bill 736 in Florida (2011) and House Bill 1001 in Colorado (2012). Similar legislation is being considered or in the process of being implemented in other states, including Michigan and Pennsylvania. Some large school districts are also independently developing performance-based teacher evaluations. The Houston Independent School District (Shifrer, Turley and Heard, 2013), L.A. Unified School District (Strunk, Weinstein and Makkonnen, 2013), Pittsburgh Public School District (Chute, 2013), and Washington DC Public School District (Arcaira et al., 2013) are examples.

² According to the GAO (2013), Tennessee was one of only six Race to the Top (RTTT) states to fully implement both their teacher and principal evaluation systems by the 2012-13 school year.

state education policy. Most notably, the new law requires that all school personnel be evaluated annually, and personnel decisions must be based at least in part on the evaluations. The four models approved by the Tennessee State Board of Education to evaluate teachers each contain several components that are combined to produce measures of overall teacher effectiveness. Unlike the traditional teacher evaluation systems examined by Weisberg et al. (2009), the new ratings in Tennessee are sufficiently dispersed to indicate meaningful variation in effectiveness across the teaching workforce (see below).

The purpose of present study is to examine one aspect of how the new, more-rigorous teacher evaluation system in Tennessee has influenced teachers. Specifically, we examine the effects of teacher ratings under the new system on job satisfaction. Job satisfaction among teachers has been the subject of active research for decades, particularly with respect to its relationship with teacher retention and teacher shortages (Borg and Riding, 1991; Chapman and Lowther, 1982; Ingersoll, 2001; Johnson, 2006; Liu and Ramsey, 2008; Stockard and Lehman, 2004; The New Teacher Project, 2012).³ Given the rapidly changing policy landscape surrounding teacher evaluations in K-12 public schools, and the clear shift toward more differentiated ratings based on performance nationally, a question we aim to answer is whether the new ratings lead to differences in job satisfaction between more- and less-effective teachers.⁴

To perform our analysis we merge ratings data for teachers in Tennessee with data from a statewide survey conducted after teachers received their first-year ratings from the system. The survey includes a number of questions about professional wellbeing. We identify the causal effects of teacher ratings on job satisfaction by leveraging discontinuities in the way that the evaluation

³ The relationship between job satisfaction and turnover has also been studied in other fields, like nursing – e.g., see Coomber and Barriball (2007) and Mueller and Price (1990).

⁴ The New Teacher Project (2012) argues that overall job satisfaction and retention among teachers are less important policy objectives than differentiated job satisfaction and retention (Goldhaber, Gross and Player, 2011, also explicitly focus on the retention of effective teachers). Put differently, students in K-12 schools will benefit if job satisfaction and retention improve among effective teachers and decline among ineffective teachers.

system converts underlying performance measures into teachers' overall ratings, as in Dee and Wyckoff (2013) and Springer et al. (2014). In particular, we estimate regression-discontinuity (RD) models that compare job satisfaction between teachers who receive similar underlying performance scores, but whose scores place them on different sides of the threshold values that were used to convert underlying scores into overall ratings.

Our findings indicate that teachers' ratings causally affect their satisfaction with work. All else equal, teachers who receive higher ratings are more likely to indicate that they are satisfied in teaching. These results are encouraging because they suggest that differentiated teacher ratings can be used to disproportionately promote job satisfaction for more effective teachers. In turn, this may lead to improvements in workforce quality over time to the extent that job satisfaction influences teacher retention.

2. Background: Teacher Evaluations in Tennessee

As part of the federal Race to the Top (RTTT) grant competition, the United States Department of Education (USDOE) called for states and their participating school districts to improve teacher and organizational effectiveness by developing comprehensive educator evaluation systems.⁵ The USDOE (2009) particularly called for systems to (1) measure student growth for each individual student, (2) design and implement evaluation systems that include multiple rating categories that take into account data on student growth as a significant factor, (3) evaluate teachers and principals annually and provide feedback, including student growth data, and (4) use these evaluations to inform decisions regarding professional development, compensation, promotion, retention, tenure, and certification. State-awarded RTTT grants have intensified efforts around the

⁵ Race-to-the-Top (RTTT) is a competitive grant program created under the American Recovery and Reinvestment Act of 2009. RTTT provides incentives for states to reform K-12 education in such areas as turning around low performing schools and improving teacher and principal effectiveness. To date, the United States Department of Education (USDOE) has awarded 19 states over \$4.35 billion to implement RTTT reforms (Tennessee's RTTT award was for \$501 million). These states serve approximately 22 million students and employ 1.9 million teachers in 42,000 schools, representing roughly 45 percent of all K-12 students and 42 percent of all low-income students (USDOE, 2014).

construction and implementation of performance-based measures for teacher evaluation systems, though the degree of implementation has varied considerably (GAO, 2013).

In July 2011, the Tennessee State Board of Education approved four teacher evaluation models. The models are the Tennessee Educator Acceleration Model (TEAM), Project Coach (COACH), Teacher Effectiveness Measure (TEM), and Teacher Instructional Growth for Effectiveness and Results (TIGER). All of the models follow the requirements adopted by the State Board of Education and have the same goals – to monitor teacher performance and encourage teacher development.

We omit data from Tennessee districts that implemented the TEM model from our study for technical reasons that we discuss below. Among the remaining three models – COACH, TEAM and TIGER – the basic components of teacher evaluations are the same, although there are some differences in how the evaluations are implemented. For example, both the TEAM and TIGER models rely on announced and unannounced classroom observations while the COACH model uses unannounced observations only. There is also variability in the number and frequency of observations – TEAM requires six observations totaling more than 90 minutes for apprentice teachers and four observations totaling more than 60 minutes for tenured teachers, TIGER requires a minimum of six observations totaling more than 60 minutes, and COACH requires that all teachers be observed at least ten times a year for a minimum of five minutes per observation. All three models have slightly different approaches to feedback, but all require a post-observation conference and written feedback to the teacher provided by the observer.

State law specifies that 50 percent of the evaluation must be based on student achievement data in all evaluation models. Among teachers for whom individual growth measures based on how their students perform on standardized tests can be constructed, 35 percent of the final rating is based on student performance on these tests. For teachers outside of grades and subjects for which

standardized tests are available, school- and/or district-wide growth is substituted for the individual growth measure.⁶ The remaining 15 percent is based on additional measures of student achievement chosen through mutual agreement by the educator and evaluator. The non-achievement-based share of each rating is based on qualitative measures including teacher observations, student perception surveys, personal conferences, and a review of prior evaluations and work.

The effectiveness scores range from 0 to 500 in all four 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).⁷ Ratings 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 because it means that teachers with very similar underlying scores but 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 the survey and they did not apply universally (in fact, some policies remain to be implemented as of our writing this draft). 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

⁶ Approximately one third of teachers in Tennessee have an individual growth measure (see Table 1).

⁷ In practice, scores just below the consequential threshold values were rounded up so that teachers who ended up just below each threshold were awarded the higher rating without exception. For example, teachers with scores close to but just below 425 received a rating of “5” (i.e., “significantly above expectation”). As we illustrate below, the discontinuities that convert the underlying scores into final ratings are sharp after accounting for the rounding (see Figure 1). We use the actual threshold values (i.e., allowing for the rounding) in the empirical work that follows. More discussion of the rounding of scores is provided in Section 4.2.

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 District program who earn a rating of three or higher can earn salary increases and/or promotions not available to teachers with lower ratings.⁸

3. Data

3.1 Ratings, Administrative and Survey Data

We merge teacher ratings data with administrative data from the Tennessee Department of Education and data from an annual survey given to all Tennessee teachers as part of the FTTT initiative. The ratings data are based on teacher performance during the 2011-2012 school year. The administrative data include information on each teacher's gender, race, education level and years of experience, also from the 2011-2012 school year. The survey data are from the annual Tennessee Consortium Survey administered during the 2012-2013 school year, after teachers received their first set of ratings from the evaluation system. 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. Critical for our study, the survey includes a number of questions about job satisfaction.

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 descriptive statistics for our analytic sample. In total, our final sample of 20,492 teachers, while large, represents less than a third of the teachers in Tennessee. Appendix Table A.1 documents the reasons that

⁸ See here: <http://achievementschooldistrict.org/wp-content/uploads/2012/12/Achievement-Teacher-Compensation-Summary-Feb-2013.pdf>.

teachers are omitted from our analytic sample. We briefly discuss the two most important reasons here. First, as noted above, we exclude teachers evaluated using the TEM model due to its unique approach to calculating teachers' overall scores, which produces a very lumpy distribution of teacher performance measures that is not compatible with our regression-discontinuity research design. The TEM model was adopted only in Memphis during the 2011-2012 school year, which accounted for approximately ten percent of the teachers in Tennessee (Memphis accounts for roughly ten percent of the population in Tennessee as well). Second, the most significant reduction in sample size in our study occurs because a large fraction of teachers – almost two-thirds – did not submit a survey.⁹

Although the response rate to the survey is not particularly low for a 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 the ratings effects on job satisfaction if the ratings treatments also affect survey submissions (a related problem arises in McCrary and Royer, 2011). Second, even if ratings do not affect survey submissions, the generalizability of our findings will be limited to the extent that teachers who submitted a survey are systematically different from other teachers. We delay our investigation of whether the ratings treatments cause survey submission until Section 4.2. To examine whether teachers who submitted a survey are different in other ways, in Table 1 we compare the observable characteristics of all Tennessee teachers in the ratings data file, which to a rough approximation represents the universe

⁹ 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. The notes in Appendix Table A.1 provide more information on individuals who exited the profession. Appendix Table A.1 also shows that we also lose some data due to missing demographic information (less than 5 percent of teachers). We drop teachers with missing demographic information from the analytic sample because our primary specifications are non-linear and we encountered convergence problems in some models when we attempted to include these teachers (with proper controls for the missing data in the models). In results omitted for brevity we verified the robustness of our findings to including these teachers by either (1) removing all demographic controls from the models or (2) estimating linear variants of our models for which convergence is not an issue.

of Tennessee teachers, to the teachers in our analytic sample.¹⁰ Outside of the difference in the share of TEM teachers, which also influences the racial share in our sample because the teaching workforce in Memphis is disproportionately black, our analytic sample looks similar to the full sample of Tennessee teachers. This gives us confidence that our findings will generalize, at least to some extent, to the broader teaching population.¹¹

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 we study. Very few teachers receive a score that puts them at level-1 (specifically, 0.4 percent of all Tennessee teachers receive a level-1 rating, which is rounded to zero in the table). Because of the small sample size, we cannot evaluate the effects of ratings around the 1/2 threshold with reasonable precision. However, moving beyond level-1 there are significant fractions of teachers who receive ratings at levels 2 through 5, which facilitates our investigation of ratings effects at the thresholds 2/3, 3/4 and 4/5. The distribution of ratings may seem somewhat generous in the absence of context, but Weisberg et al. (2009) find that in the districts they study that use a rating system with more than two options, 94 percent of teachers receive one of the top two ratings. Just 73 percent of teachers in Tennessee receive one of the top two ratings, indicating that the rigor of the Tennessee system is substantially higher than the “industry standard” in public education.

¹⁰ 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 A.1 (in fact, the differences are even smaller because the largest differences between the columns in Table 1 are owing to the exclusion of TEM teachers).

¹¹ 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.

3.2 *Measuring Job Satisfaction with the Survey*

There are ten questions on the survey instrument that elicit feedback from teachers regarding their satisfaction with work. Table 2 splits the ten questions into two groups – one group consists of questions about general job satisfaction and the other about job satisfaction specific to the school. The table summarizes the content of each question and Appendix B shows each question as it was presented to teachers on the survey.

We divide teacher responses to each question into four categories: (1) positive, (2) negative, (3) non-response by choice and (4) non-response due to a position change (i.e., respondent was directed to skip question). To code each answer as either a positive or negative response, we collapse teachers' more-detailed responses into the two categories as indicated in Appendix B. On average across the ten questions, 13.6 percent of survey respondents did not answer by choice. Just over 12 percent of survey respondents were directed to skip the ten job satisfaction questions (among other questions) due to a position change.¹²

Table 2 shows the proportions of positive responses to each question across teachers by final rating in the evaluation system. The denominator in each cell is the total number of submitted surveys, including those for which a non-response was recorded for the relevant question. With the exception of question 3, 4 and 9, teachers who are assigned a higher rating convey higher satisfaction with work. Even for questions 3, 4 and 9, teachers with higher ratings do not indicate that they are meaningfully less-satisfied with work. Some of the differences across teacher ratings are modest (e.g., question 2) and others are large (e.g., question 7). In Appendix Table A.2, we show that a

¹² Position changes that resulted in teachers being directed to skip the job-satisfaction questions include changes to observer teachers, mentor teachers, math/reading specialists, 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.

qualitatively similar but stronger pattern emerges in the data 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) point to an association between teacher ratings and job satisfaction, attributing causality is not straightforward. The reason is that we cannot tell if effective teachers are more satisfied because they are identified as effective, or they are more satisfied simply because effective teachers like teaching better (e.g., see Lawler and Porter, 1967; Schwab and Cummings, 1970). We overcome this identification problem with the RD approach that we describe in the next section

Finally, Table 3 presents a correlation matrix that documents dependence in the job-satisfaction data. Unsurprisingly, teacher responses are positively correlated across the job-satisfaction questions and the correlations are moderate to high. Correlations are larger within the “general” and “school-specific” sets of questions than across, although the differences in many cases are modest. The key piece of information to take away from the correlation matrix is that our outcome measures are strongly related – this is an issue that we will account for directly in the analysis below when we test for the effects of evaluation ratings on teachers’ overall satisfaction with work.

4. Methodology

4.1 Regression Discontinuity Specification

We use an RD approach to identify the causal impacts of differential evaluation ratings on teacher job satisfaction. The RD design compares 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. 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 (Imbens and Lemieux, 2008; Lee and Lemieux, 2010).

In order to present a more comprehensive picture of how teachers responded to the survey questions and utilize all of the available information conveyed in their responses, we construct a multinomial outcome variable for each question as follows:

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

We then estimate multinomial logistic RD models using a “sharp” discontinuity design (see Figure 1 below). We estimate separate models for each question at each performance threshold that we consider (these are the thresholds between levels 2/3, 3/4 and 4/5; recall from above that we do not consider the 1/2 threshold because very few teachers in Tennessee receive a level-1 rating). The models take the following form:

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

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

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

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 (i.e., the regression discontinuity indicator), and ε_i is the error term,

which we cluster at the school level.¹³ The X -vector in our primary specification includes teacher gender, race, degree level and experience. We also consider an expanded X -vector that includes school characteristics. 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.¹⁴ The parameters of interest are in the vector β_4 , which under the RD assumptions are the causal effects of receiving a higher rating on responses to the job satisfaction 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 job satisfaction 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.

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 (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 within each ratings-pair, the discontinuity in converting the

¹³ 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 does not seem egregious in our application, in results omitted for brevity we investigated its implications empirically by estimating variants of our models where we clustered the data on the 0.5-unit intervals, as suggested by Lee and Card (2008). Our findings from the alternatively-clustered models are substantively similar to what we report below.

¹⁴ We use the Akaike information criterion (AIC) test to determine the polynomial order for our primary specification. In results omitted for brevity we show that adding higher order polynomial terms (up to quartic) of the running variable to the models does not influence our results qualitatively.

underlying performance measures into final ratings is sharp, supporting the above-specified RD design.

As noted above, the most important context-specific issue we face in inferring causality from our RD estimates is that a large fraction of Tennessee teachers did not submit a survey. Note that survey submission is not synonymous with question non-response. Teachers who submitted a survey but skipped questions are included in our analytic sample and analyzed via the multinomial-response model described in the previous section. It is teachers who failed to submit a survey entirely who are potentially problematic. The primary 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 our dataset that is itself caused by treatment (i.e., the ratings).

To test whether teachers' decisions to submit a survey were influenced by the discontinuities that convert underlying scores to ratings, we estimate supplementary RD models analogous to what we show in equation (1) above. The supplementary models are estimated for all Tennessee teachers and the dependent variable is a binary indicator for whether each teacher submitted a survey. The models include the exact same variables shown in equation (1), but 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 our restricted sample of teachers who submitted a survey.¹⁵

¹⁵ Our 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 57,694 across all three discontinuity thresholds. Appendix Table A.1 shows the reasons that teachers are excluded from these models.

Table 4 displays the estimated effects of treatment on survey submission at each of the three performance thresholds we evaluate. At each cutoff, the estimated treatment effect is substantively small and statistically insignificant, which suggests that the survey participation rate is not caused by the ratings treatment, at least subject to the local interpretation of the RD estimates (which is the most relevant interpretation for informing the credibility of our main analysis). The evidence in Table 4 suggests that our RD estimates of the effects of ratings on job satisfaction, based on the sample of Tennessee teachers who submitted a survey, will not be biased by sample selection.¹⁶

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 running-variable values near the cutoff are not otherwise similar.¹⁷

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 models of the following form:

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

In equation (2), X_i is a teacher characteristic variable 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.

¹⁶ 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 4 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.

¹⁷ 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.

For simplicity, each teacher characteristic that we evaluate using equation (2) is converted into a binary indicator variable and the equation is specified as a linear probability model.¹⁸

Table 5 presents results from the series of RD regressions linking the ratings discontinuities to teacher characteristics. Not all covariates are balanced. For example, in our analysis of the 2/3 threshold, group-1 teachers (with individual growth scores) are less likely to receive the higher rating, while at the 3/4 threshold they are more likely to receive the higher rating. However, in addition to the fact that the estimates in the table do not imply a clear pattern of potential selection bias, it is also not obvious that the observed lack of balance in Table 5 is any worse than what would be expected by chance. In order to determine the likelihood of observing the number of unbalanced covariates reported in each column of the table by chance, we construct a simulation-based test following Cullen, Jacob and Levitt (2005) and Fitzpatrick, Grissmer and Hastedt (2011).

To construct the simulations we start by splitting 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) 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.

¹⁸ 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.

Based on our 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, 3/4 and 4/5 thresholds, the p-values are 0.44, 0.17 and 0.15, respectively. Thus, we conclude that the degree of covariate imbalance in Table 5 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, 2010; Van Der Klaauw, 2002).

We perform density tests around the three thresholds in our study and report the results in Appendix C. The initial density 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 1-percent level). In our application, however, it does not appear that the bunching of the running variable at the 4/5 cutoff is the result of non-random manipulation by individual teacher evaluators. Instead, the disproportionate weight in the density on the high side of the threshold appears to be driven largely by a system-wide, *ex post* rounding of teacher scores. Specifically, although 425 is the official threshold for earning a final rating of “5” in the system, teachers with scores close to but just below this threshold were, without exception, assigned ratings of “5” as well. It is not surprising that this type of rounding would cause bunching just above the threshold.

If we perform the density tests based on hypothetical teacher ratings without allowing for the rounding that occurred in practice, the bunching of the density at the 4/5 cutoff improves considerably and is only significant at the 10 percent level (see Appendix C for details). The implication is that the failure of the density test at the 4/5 cutoff in our data is driven primarily by the system-wide rounding of teacher ratings, and not the non-random manipulation of scores for individual teachers by evaluators. It is the latter scenario, which is conceptually similar to the canonical test-score example from above, that would be problematic for identification in our study. 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. Inference depends on the source of the failure of the test, which in our case seems unlikely to be cause for concern given that the rounding occurred throughout the system without case-by-case discretion. As noted in Section 2, in the empirical analysis we use teachers' ratings based on their actual, rounded scores (that is, the ratings teachers actually received), which is why all discontinuities shown in Figure 1 are sharp.¹⁹

5. Results

5.1 Primary Results

Table 6 presents the results from our main regressions as shown in equation (1). The table reports the effect of the higher rating at each discontinuity for each job satisfaction 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 in Table 6 indicates that the ratings treatment caused an increase in the likelihood of a positive response relative to a negative response, and a coefficient of less than one indicates the opposite. T-statistics are in parentheses for

¹⁹ Even with the rounding issue it is correct to interpret the discontinuities in our data as “sharp” – in contrast to what we observe with the Tennessee data, a fuzzy discontinuity occurs when individuals with the same running variable values can be assigned to different discrete performance categories with a probability above zero.

whether the coefficients are significantly different from 1.0, where a value of 1.0 would indicate a null effect.

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 6. We briefly comment on the coefficients for the non-response outcomes in Section 5.2, and show results for these outcomes in Appendix D.

First consider our findings for the individual questions in Table 6. To illustrate how the estimates in the table can be interpreted, consider our estimate for question 3 at the 4/5 threshold. This estimate indicates that teachers with the higher rating are 1.146 times more likely to respond positively rather than negatively to the question compared to teachers with the lower rating. The table shows significant positive effects of the higher rating at the 2/3 threshold for questions 5 and 6 (and question 8 if we include results significant at the 10 percent level), and at the 4/5 threshold for questions 3, 6 and 7 (and question 10 if we include results significant at the 10 percent level). None of our estimates imply statistically significant effects of the higher rating on job satisfaction at the 3/4 threshold, and none of our estimates at any threshold indicate a statistically significant negative effect of the higher rating on job satisfaction.

Looking at the findings in the table holistically, the estimates for the 2/3 and 4/5 thresholds are overwhelmingly positive: 9/10 and 10/10 of the point estimates are greater than one and 2/10 and 3/10 estimates are statistically significant at the 5-percent level or better at these thresholds, respectively. However, recall from Table 3 that teacher responses to the job-satisfaction questions are positively correlated, which makes it difficult to infer how likely it is that these results would arise by chance. For example, because of the moderate to strong correlations between outcome variables

in Table 6, observing either a particularly high or low number of positive coefficients by chance may not be that rare.

In the face of the strong dependence across job-satisfaction outcomes, we perform another simulation exercise in order to determine the likelihood of observing what we see in Table 6 by chance. The new simulations are analogous to the simulations from Table 5. We again split the data panel vertically for the simulations, this time separating out the blocks of teacher survey responses (dependent variables) and teacher characteristics/scores (independent variables). We re-sort the survey response data using a random number and re-merge the dataset to assign random job satisfaction outcomes to teachers. At each iteration with the randomly-assigned job satisfaction outcomes, we estimate our models at all three thresholds and store the number of coefficients above 1.0. We do this 3,000 times, and at the bottom of Table 6 we report the probability of observing at least as many coefficients greater than 1.0 as we estimate with the real data by chance at each threshold.

Again, at the 2/3 threshold we estimate coefficients above 1.0 for nine of the ten questions with the real data. Based on our simulations, the probability of observing at least nine out of ten coefficients greater than 1.0 at that threshold is nine percent. For the 3/4 and 4/5 thresholds, the probabilities of estimating at least as many coefficients above 1.0 as we estimate with the real data are 80 and 3 percent, respectively. Thus, the simulation-based statistical tests reported in the bottom row of Table 6 confirm that the visual patterns in the estimates reflect statistically significant differences in job satisfaction for teachers with higher ratings at the 4/5 threshold, and to a lesser extent the 2/3 threshold. Similarly, the simulations confirm the visually-apparent null result at the 3/4 threshold.²⁰

²⁰ A reasonable alternative metric by which to measure the prevalence of positive responses would be the number of statistically significant coefficients above 1.0, although our approach makes better use of the totality of information

The heterogeneity in the strength of our findings across thresholds in Table 6 does not appear to be strongly tied to obvious, documented statewide policy differences in the rewards and sanctions associated with the different ratings at the time of the survey. For example, the 2011 law that ties teacher ratings to tenure decisions for untenured teachers (*Public Chapter 70*, 2011) indicates that the 3/4 threshold is of first-order importance, as teachers with ratings of 4 and 5 are treated equally, and we find no effect of the higher rating on job satisfaction at the 3/4 threshold. Our marginally significant results at the 2/3 threshold align with the key threshold used to determine eligibility for salary increases and promotions in the Tennessee Achievement District program, but this applies only to teachers at particularly low-performing Tennessee schools and thus seems unlikely to drive our broader findings for teachers throughout the state. Two plausible explanations for the heterogeneity in the strength of our findings across thresholds that we cannot evaluate directly with our data are (1) there may be psychological effects associated with ratings based on how they are described to teachers, with level-2 “Below Expectation” and level-5 “Significantly Above Expectation” ratings leading to larger effects on job satisfaction due to their descriptions, and (2) teachers may experience local rewards and sanctions based on their ratings around the thresholds where we find larger effects, and/or they may perceive that future local and statewide rewards and sanctions will be based on these ratings.

5.2 *Missing Response Outcomes*

In this section we briefly discuss our findings for the missing-response outcomes. 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 6, but for voluntary non-response and position change non-response outcomes, respectively. The coefficients shown in the tables are estimated in

contained by the estimates in Table 6. For completeness, we have also performed simulations based on the number of statistically significant positive coefficients and we obtain results similar to what we report in the table.

the same multinomial models as the coefficients reported in Table 6. The relative risk ratio coefficients are again relative to negative responses.

The prevalence of coefficients above 1.0 in Appendix Table D.1 suggests that on the whole, teachers who receive higher ratings are, if anything, more likely to voluntarily skip a job-satisfaction question rather than answer it negatively. Although relative risk ratios above 1.0 are not more common than we would expect by chance in Appendix Table D.1, per the p-values reported at the bottom of the table, we view these results as broadly consistent with our main findings in Table 6. The results for non-response due to a position change are more mixed, with estimated effects that are nominally small and statistically insignificant throughout Table D.2. These results are consistent with position changes not being affected by discontinuities in the ratings.

5.3 *Robustness and Extensions*

We examine the robustness of our findings in several ways. First, in Appendix Table E.1 we replicate the results in Table 6 after incorporating information about the schooling environment. Although it seems unlikely, 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 cause bias in our estimates in Table 6. 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.1 shows that our findings are essentially unchanged from what we report in Table 6.

Next we examine the sensitivity of our findings to alternate bandwidth specifications around the discontinuity thresholds. Our main results in Table 6 use data from all teachers with ratings on either side of each threshold, which includes teachers within a range of scores that is 75 points in either direction. In Appendix Tables E.2 through E.7 we consider alternative bandwidths of 60, 50 and 40 points for each estimate reported in Table 6. The tables show that although our standard

errors increase as the bandwidths narrow, our findings are generally robust to shrinking the bandwidth around the discontinuity thresholds.

A third robustness issue that we examine relates to whether all of the teachers in our analytic sample were actually treated. Although all teachers were provided access to their ratings online, approximately ten percent of survey respondents indicate on the survey that they did not see their ratings.²¹ We do not do anything differently for these teachers in our 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 our analytic sample will attenuate our results.

In Appendix Table E.8 we again replicate the results in Table 6, but this time we restrict the analytic sample to exclude teachers who indicate that they had not seen their ratings. The results in the appendix table show that the ratings treatments had positive effects on job satisfaction for this restricted sample that are as large, or larger, than what we find in Table 6. These findings are consistent with some attenuation bias being reduced when we restrict the sample to individuals who report receiving treatment.

6. 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, 2014; Hanushek, 2011; Hanushek and Rivkin, 2010). The educational and economic importance of teacher quality motivates the development of policies aimed at improving the effectiveness of teachers. A critical first step in the improvement process is to develop evaluation systems that better identify the large differences in effectiveness across teachers that have been consistently shown in research but have gone unrecognized in traditional

²¹ Instructions to teachers for how to obtain their ratings can be found here:
http://team-tn.org/wp-content/uploads/2013/08/Teacher-Access-Guide-for-CODE_July_2014.pdf

teacher evaluation systems (Weisberg et al., 2009). More informative ratings for teachers can be used in a number of ways to improve student outcomes. For example, they can be used to improve feedback to teachers (Taylor and Tyler, 2012), better direct educator-to-educator learning (Ehlert et al., 2014), re-assign teachers to tasks that are better-suited to their strengths (Condie, Lefgren and Sims, 2014), and to re-shape the teaching workforce (Boyd et al., 2011; Dee and Wyckoff, 2013; Goldhaber and Theobald, 2013; Hanushek, 2009; Springer, Rodriguez and Swain, 2014; Winters and Cowen, 2013), among other possibilities.

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. Tennessee has been at the forefront of education policy in this regard, and the present study aims to use the Tennessee experience to learn about how more-rigorous evaluations influence job satisfaction among teachers. More specifically, we ask whether the assignment of ratings itself affects job satisfaction. Using a regression discontinuity design that locally identifies the effects of higher ratings near thresholds that convert teachers' underlying scores into discrete final ratings, our results show that job satisfaction is affected by ratings. Teachers who receive higher ratings indicate that they are happier in teaching than otherwise similar teachers who receive lower ratings. This result is particularly apparent for teachers near the threshold for being rated "Significantly Above Expectation" (the top rating in the system in Tennessee), and to a lesser extent for teachers near the threshold between ratings of "Below Expectation" (level-2) and "At Expectation" (level-3).

Our findings are intuitive and validate a simple but important aspect of policies that more-rigorously evaluate teachers. The Tennessee experience shows that if an education agency identifies quality metrics that it deems to be important for the teaching workforce, and then rigorously rates teachers on these metrics, it can differentially affect job satisfaction among teachers. All else equal, teachers who exhibit the quality indicators deemed to be important by the education

agency (and who earn higher ratings as a result) will be more satisfied with their jobs than their counterparts who do not fare as well on the evaluation. Thus, ratings systems themselves can be a useful human-resource tool for differentially promoting employee wellbeing in a way that favors more effective teachers.

To the extent that job satisfaction positively affects persistence in the teaching workforce, a topic that has been studied extensively in previous research (Borg and Riding, 1991; Chapman and Lowther, 1982; Grissom, Nicholson-Crotty and Harrington, forthcoming; Ingersoll, 2001; Johnson, 2006; Liu and Ramsey, 2008; Stockard and Lehman, 2004; The New Teacher Project, 2012), our findings raise the possibility that the implementation of more-rigorous teacher evaluation systems may lead to long-term improvements in workforce quality by influencing self-selection. However, this interpretation of our results is subject to several caveats. First, while our RD design lends credence to the internal-validity of our findings, a general limitation of the approach is that our findings are specific to teachers near the rating thresholds. It is not obvious in our application why our findings for teachers near the thresholds would be meaningfully different than for teachers further out, particularly given that teachers were not made aware of their underlying scores, but we cannot directly test this with our data. Perhaps a more important outstanding question that cannot be resolved by our study is how the implementation of the Tennessee evaluation system has affected teacher wellbeing overall. Put differently, while we are able to show that teachers who earned higher ratings were made to be more satisfied with work relative to their lower-rated peers, we do not have pre-implementation survey data to determine the overall impact of the new evaluation system on job satisfaction. A concern is that all teachers may be less satisfied with teaching post-implementation, in which case the differentiating effects of ratings that we identify could be swamped by the “tide” of an overall negative effect. Of course, we also cannot rule out an overall positive effect of the more-rigorous evaluation system, and following on this point it is of some comfort that available

evidence does not seem to suggest that job satisfaction among teachers on the whole is adversely affected by increased accountability (Grissom, Nicholson-Crotty and Harrington, forthcoming). However, we hesitate to infer too much based on prior evidence because previous versions of accountability have looked different than what is emerging in new teacher evaluation systems like the one that we study in Tennessee. The question of the total workforce effect is one that merits attention in future research as additional rigorous evaluation systems come online and more data become available.

References

- Arcaira, Erikson, Beatrice Birman, Stephen Coleman, Erin Dunlop, Michael Feuer, Maxine Freund, Steve Glazerman, Jane Hannaway, Heather Harding, Jaclyn MacFarlane, Taunya Nesin, Umut Ozek, Andrea Palmiter, Brenda Turnbull and Elias Walsh. 2013. Evaluation of the DC Public Education Reform Amendment Act (PERAA): Report No. 2, School Year 2011-2012. Report Published by the Education Consortium for Research and Evaluation at The George Washington University.
- Borg, Mark G. and Richard J. Riding. 1991. Occupational Stress and Satisfaction in Teaching. *British Educational Research Journal* 17(3), 263–281.
- Boyd, Donald, Hamilton Lankford, Susanna Loeb and James Wyckoff. 2011. Teacher Layoffs: An Empirical Illustration of Seniority v. Measures of Effectiveness. *Education Finance and Policy* 6(3), 439-454.
- Chapman, David W. and Malcolm A. Lowther. 1982. Teachers' Satisfaction with Teaching. *Journal of Educational Research* 75(4), 241-247.
- 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.
- Chute, Eleanor. 2013. New Teacher Evaluation Process Set to Begin in Pittsburgh Public Schools. *Pittsburgh Post-Gazette* (08.13.2013).
- Condie, Scott, Lars Lefgren and David Sims. 2014. Teacher Heterogeneity, Value-Added and Education Policy. *Economics of Education Review* 40(1), 76-92.
- Coomber, Billie and K. Louise Barriball. 2007. Impact of Job Satisfaction Components on Intent to Leave and Turnover for Hospital-Based Nurses: A Review of the Research Literature. *International Journal of Nursing Studies* 44(2), 297-314.
- 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.
- Dobbie, Will and Roland G. Fryer, Jr. 2011. The Impact of Youth Service on Future Outcomes: Evidence from Teach For America. NBER Working Paper No. 17402.
- Ehlert, Mark, Cory Koedel, Eric Parsons and Michael Podgursky. 2014. Choosing the Right Growth Measure: Methods Should Compare Similar Schools and Teachers. *Education Next* 14(2), 66-71.
- 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, Bethany Gross and Daniel Player. 2011. Teacher Career Paths, Teacher Quality, and Persistence in the Classroom: Are Public Schools Keeping Their Best? *Journal of Public Policy and Management* 30(1), 57-87.

Goldhaber, Dan and Roddy Theobald. 2013. Managing the Teacher Workforce in Austere Times: The Determinants and Implications of Teacher Layoffs. *Education Finance and Policy* 8(4), 494–527.

Grissom, Jason A., Sean Nicholson-Crotty and James R. Harrington (forthcoming). Estimating the Effects of No Child Left Behind on Teachers' Work Environments and Job Attitudes. *Educational Evaluation and Policy Analysis*.

Hanushek, Eric A. 2009. Teacher Deselection, in *Creating a New Teaching Profession* eds. Dan Goldhaber and Jane Hannaway. Urban Institute, Washington, DC.

Hanushek, Eric A. 2011. The Economic Value of Higher Teacher Quality. *Economics of Education Review* 30 (3), 466-479.

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

Herman, Jeanne B. 1973. Are Situational Contingencies Limiting Job Attitude—Job Performance Relationships? *Organizational Behavior and Human Performance* 10(2), 208-224.

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

Ingersoll, Richard. 2001. Teacher Turnover and Teacher Shortages: An Organizational Analysis. *American Educational Research Journal* 38(3), 499–534.

Jepsen, Christopher, Peter Mueser and Kenneth Troske. 2010. Labor-Market Returns to the GED Using Regression Discontinuity Analysis. University of Missouri Department of Economics Working Paper No. 10-14.

Johnson, Susan M. 2006. The Workplace Matters: Teacher Quality, Retention and Effectiveness. Washington DC: National Education Association. <http://www.eric.ed.gov/PDFS/ED495822.pdf>

Johnson, Susan M., Matthew A. Kraft and John P. Papay. (2012). How context matters in high-need schools: The effects of teacher working conditions on their professional satisfaction and their students achievement. *Teachers College Record* 114(10), 1-39.

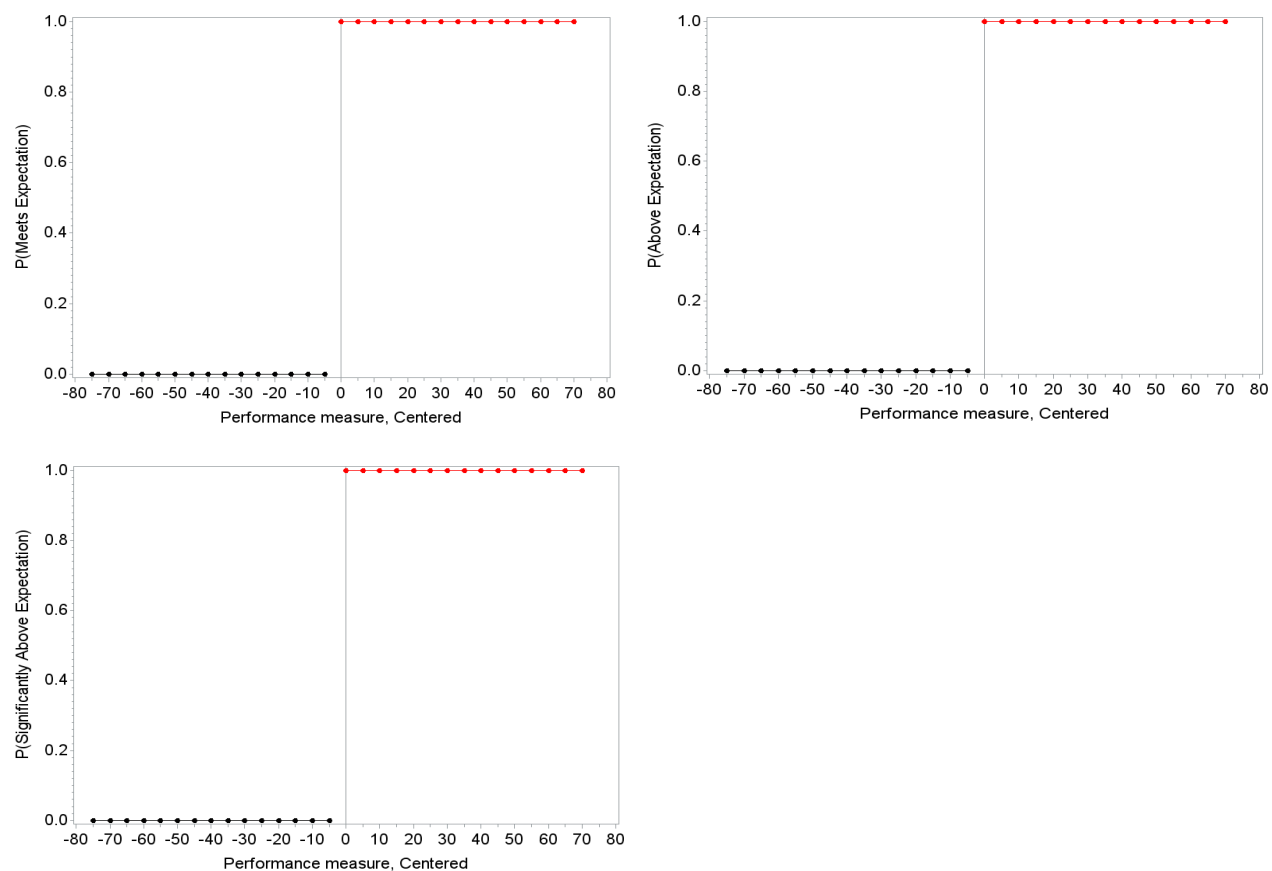
Judge, Timothy A. and Joyce E. Bono. 2001. Relationship of Core Self-Evaluations Traits—Self-Esteem, Generalized Self-Efficacy, Locus of Control, and Emotional Stability—With Job Satisfaction and Job Performance: A Meta-Analysis. *Journal of Applied Psychology* 86(1), 80-92.

Koch, Steven F. and Jeffrey S. Racine. 2013. Health Care Facility Choice and User Fee Abolition: Regression Discontinuity in a Multinomial Choice Setting. Working Paper.

- Landy, Frank J. 1989. *Psychology of Work Behavior*. Published by Pacific Grove, Calif. : Brooks/Cole Pub. Co.
- Lawler, Edward E. III, and Lyman W. Porter. 1967. The Effect of Performance on Job Satisfaction. *Industrial Relations* 7(1), 20-28.
- 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.
- Liu, Xiaofeng Steven and Jase Ramsey. 2008. Teachers' Job Satisfaction: Analyses of the Teacher Follow-up Survey in the United States for 2000-2001. *Teaching and Teacher Education* 24(5), 1173-1184.
- 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.
- Mueller, Charles W. and James L. Price. 1990. Economic, Psychological, and Sociological Determinants of Voluntary Turnover. *Journal of Behavioral Economics* 19(3), 321-335.
- 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.
- Schwab, Donald P. and Larry L. Cummings. 1970. Theories of Performance and Satisfaction: A Review. *Industrial Relations* 9(4), 408-430.
- Shifrer, Dara, Ruth Lopez Turley and Holly Heard. 2013. Houston Independent School District's Aspire Program: Estimated Effects of Receiving Financial Awards. Houston Educational Research Consortium Policy Report.
- Springer, Matthew G., Luis A. Rodriguez, and Walker A. Swain. (2014). Effective Teacher Retention Bonuses: Evidence from Tennessee. Nashville, TN: Vanderbilt University.
- Stockard, Jean and Michael B. Lehman. 2004. Influences on the Satisfaction and Retention of 1st-Year Teachers: The importance of Effective School Management. *Educational Administration Quarterly* 40(5), 742-71.
- Strunk, Katharine O., Tracey L. Weinstein and Reino Makkonnen. 2013. Sorting out the Signal: Do Multiple Measures of Teachers' Effectiveness Provide Consistent Information to Teachers and Principals? Working Paper, University of Southern California.

- 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.
- Triandis, Harry C. 1959. A critique and Experimental Design for the Study of the Relationship Between Productivity and Job Satisfaction. *Psychological Bulletin* 56(4), 309-312.
- United States Government Accountability Office. 2013. Race to the Top: States Implementing Teacher and Principal Evaluation Systems Despite Challenges. Washington, DC.
- United States Department of Education. 2009. Race to the Top. 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.
- 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.
- Winters, Marcus A. and Joshua M. Cowen. 2012. Grading New York: Accountability and Student Proficiency in America's Largest School District. *Educational Evaluation and Policy Analysis* 34(3), 313-327.
- 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 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. As a practical matter, final ratings are determined based on underlying scores after rounding, and thus the graphs in this figure portray final ratings as a function of rounded underlying scores (see text and Appendix C for further discussion).

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

Variable	All Tennessee Teachers	Analytic Sample
Female teacher	0.80	0.82
Black teacher	0.12	0.06
White teacher	0.87	0.93
Other race	0.01	0.00
Bachelor degree	0.41	0.38
Education specialist	0.07	0.08
Master degree	0.42	0.44
Other education	0.10	0.10
Years of experience: 0-1	0.10	0.08
Years of experience: 2-4	0.14	0.13
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	0.14
Years of experience: 20+	0.24	0.25
Group 1 (with individual growth score)	0.32	0.36
Professional license	0.85	0.89
Apprentice license	0.14	0.11
Partial-year license exemption	0.01	0.00
COACH evaluation model	0.06	0.05
TEAM evaluation model	0.82	0.91
TEM evaluation model	0.10	0.00
TIGER evaluation model	0.02	0.03
Level 1 (Sig. Below Expectations)	0.00	0.00
Level 2 (Below Expectations)	0.07	0.06
Level 3 (At Expectations)	0.20	0.21
Level 4 (Above Expectations)	0.32	0.31
Level 5 (Sig. Above Expectations)	0.41	0.42
N	69796	20492

Notes: Among all Tennessee teachers, just 0.4 percent of teachers received an evaluation rating of “1,” which indicates significantly below expectation (0.2 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. Job Satisfaction in Brief and the Proportion of Positive Responses to each Question.

	Below Expectation (2)	At Expectation (3)	Above Expectation (4)	Significantly Above Expectation (5)
<i>General Satisfaction with Teaching</i>				
Question 1: Teacher prefers a higher paying job to teaching	0.42	0.44	0.44	0.45
Question 2: Teacher is less enthusiastic about teaching	0.32	0.33	0.33	0.34
Question 3: Teacher regrets teaching career choice	0.36	0.36	0.36	0.36
Question 4: Teacher plans to quit teaching	0.31	0.31	0.30	0.29
<i>Satisfaction at School</i>				
Question 5: Teaching in this school is not worth the stress	0.50	0.54	0.54	0.56
Question 6: Teacher believes that the staff is satisfied at this school	0.46	0.51	0.52	0.53
Question 7: Teacher likes how the things are run at this school	0.43	0.49	0.49	0.51
Question 8: Teacher would like to transfer to another school	0.46	0.52	0.51	0.52
Question 9: Teacher is tired of teaching at this school	0.50	0.61	0.50	0.50
Question 10: Teacher plans to return to this school next year	0.64	0.66	0.66	0.64
N	1228	4215	6345	8655

Notes: Positive responses are coded as the ones that convey satisfaction with teaching. For example, for question 6, the most positive response is "Strongly Agree"; for question 1, it is "Strongly Disagree." 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. Correlation Matrix for Positive Responses across Questions.

Question	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Q1	1.00									
Q2	0.54	1.00								
Q3	0.58	0.49	1.00							
Q4	0.45	0.44	0.49	1.00						
Q5	0.57	0.50	0.44	0.37	1.00					
Q6	0.47	0.44	0.39	0.33	0.68	1.00				
Q7	0.45	0.44	0.37	0.32	0.65	0.74	1.00			
Q8	0.50	0.45	0.38	0.32	0.65	0.62	0.64	1.00		
Q9	0.54	0.54	0.45	0.40	0.57	0.50	0.47	0.51	1.00	
Q10	0.48	0.40	0.39	0.36	0.63	0.58	0.57	0.63	0.50	1.00

Notes: Question numbers correspond to those in Table 3. The upper box blocks off general questions and the lower box blocks off school-specific questions.

Table 4. Regression Discontinuity Estimates of the Effects of Ratings on Survey Submission.

Dependent Variable	Level 2-3	Level 3-4	Level 4-5
Response Rate	0.008 (0.017)	-0.011 (0.012)	0.010 (0.009)
N	15156	29753	42350

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 5. Regression Discontinuity Estimates of the “Effects” of Ratings on Teacher Characteristics that Should Not be Affected, for the Purpose of Validating the Research Design.

Variable	Level 2-3	Level 3-4	Level 4-5
Female	0.009 (0.025)	-0.025 (0.017)	-0.013 (0.012)
African American	0.023 (0.019)	0.013 (0.010)	-0.008 (0.007)
White	-0.032 (0.019)*	-0.013 (0.011)	0.012 (0.007)*
BA degree	0.020 (0.028)	0.036 (0.020)*	-0.002 (0.015)
MA degree	-0.017 (0.029)	-0.033 (0.020)	0.016 (0.015)
Educational specialist	-0.015 (0.016)	-0.007 (0.012)	0.007 (0.009)
Years of experience: 0-1	0.027 (0.016)*	0.020 (0.012)*	-0.009 (0.009)
Years of experience: 2-4	-0.006 (0.020)	0.035 (0.014)**	0.032 (0.010)***
Years of experience: 5-9	0.004 (0.023)	0.006 (0.016)	0.019 (0.012)
Years of experience: 10-14	-0.029 (0.022)	-0.019 (0.016)	0.004 (0.012)
Years of experience: 15-19	-0.018 (0.019)	-0.025 (0.014)*	-0.020 (0.010)**
Group-1	-0.073 (0.031)**	0.066 (0.024)***	0.004 (0.015)
Professional license	-0.019 (0.018)	-0.027 (0.014)*	0.004 (0.010)
Overall p-value	0.44	0.17	0.15
N	5443	10560	15000

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

Note: 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. 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 Table 1 for details about the variables listed here.

Table 6. Effects of Ratings on Job Satisfaction for Teachers.

Dependent Variable	Level 2-3	Level 3-4	Level 4-5
<i>General Satisfaction with Teaching</i>			
Question 1	1.139 (0.90)	0.966 (-0.33)	1.095 (1.22)
Question 2	1.007 (0.05)	0.944 (-0.60)	1.104 (1.40)
Question 3	1.095 (0.68)	1.015 (0.15)	1.146 (1.98)**
Question 4	1.125 (0.87)	0.956 (-0.45)	1.030 (0.43)
<i>Satisfaction at School</i>			
Question 5	1.360 (2.13)**	0.896 (-1.00)	1.116 (1.32)
Question 6	1.303 (2.00)**	1.070 (0.64)	1.176 (2.01)**
Question 7	1.242 (1.62)	0.981 (-0.18)	1.325 (3.36)***
Question 8	1.338 (1.91)*	0.853 (-1.48)	1.138 (1.59)
Question 9	0.984 (-0.11)	1.009 (0.09)	1.059 (0.76)
Question 10	1.172 (0.86)	0.967 (-0.26)	1.207 (1.90)*
Overall p-value	0.09*	0.80	0.03**
N	5443	10560	15000

***/**/* 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 in the preceding analysis, standard errors are clustered at the school level. P-values indicate the probability in each column of the observed number of relative risk ratios greater than one occurring by chance (based on 3,000 bootstrap repetitions).

Appendix A Data Appendix

Table A.1. Construction of the Analytic Sample.

Total Number of Records in the TN Evaluation Dataset		69796
	Records Deleted	Remaining sample
Observer teachers	-805	68991
Evaluated using TEM model	-6822	62169
Missing demographic data	-3086	59083
Missing final rating due to teaching status (e.g., part-time, non-apprentice, not licensed)	-556	58527
Missing final rating for other reason	-833	57694
Did Not Take the First to the Top Survey	-37202	20492

Notes: Observer teachers perform classroom observations for other teachers and answer a different set of questions on the survey, hence they are excluded. There were also 3,584 teachers who left the Tennessee teaching workforce after the first year of ratings before they could have completed a survey. These individuals are not explicitly identified in the table and are spread throughout the rows. They represent 5.1 percent of the total universe sample, and 7.3 percent of deleted records. Among the 37,202 records that were deleted for not submitting a survey in the final row of the table, teachers who left teaching in Tennessee are essentially proportionally represented (2,517 individuals, or 6.8 percent of records deleted for that reason).

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)	
		N	Ratio	N	Ratio	N	Ratio	N	Ratio
<i>General Satisfaction with Teaching</i>									
	Question 1	935	0.55	3165	0.59	4719	0.60	6149	0.63
	Question 2	945	0.41	3208	0.43	4760	0.44	6212	0.47
	Question 3	953	0.46	3247	0.47	4811	0.47	6299	0.50
	Question 4	954	0.40	3249	0.40	4838	0.39	6332	0.39
<i>Satisfaction at School</i>									
	Question 5	936	0.65	3173	0.72	4745	0.73	6171	0.78
	Question 6	941	0.60	3196	0.67	4780	0.69	6211	0.74
	Question 7	928	0.57	3172	0.65	4726	0.65	6153	0.72
	Question 8	933	0.61	3173	0.69	4735	0.69	6169	0.74
	Question 9	932	0.66	3186	0.67	4742	0.67	6174	0.70
	Question 10	958	0.82	3266	0.85	4866	0.86	6335	0.88

Note: This table is analogous to Table 2 in the text.

Appendix B

Coding and Other Details for the Job Satisfaction Questions

Below we show each job-satisfaction question listed in Table 2 in full form as it was posed to teachers. 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, which are not reported here for brevity. More information is available from the authors upon request.

Q1. If I could get a higher paying job I'd leave education as soon as possible.

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

Q2. I don't seem to have as much enthusiasm now as I did when I began in education.

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

Q3. If you could go back to your college days and start over again, would you become an educator or not?

- a. *Certainly would become an educator.*
- b. *Probably would become an educator.*
- c. Chance about even to become an educator.
- d. Probably would not become an educator.
- e. Certainly would not become an educator.

Q4. How long did you plan to remain in education?

- a. *As long as I am able.*
- b. Until I am eligible for retirement benefits from this job.
- c. Until I am eligible for retirement benefits from a previous job.
- d. Until I am eligible for Social Security benefits.
- e. Until a specific life event occurs (e.g., parenthood, marriage).
- f. Definitely plan to leave as soon as I can.
- g. Undecided at this time.

Q5. The stress and disappointments involved in being at this school aren't really with it.

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

Q6. The staff at this school like being here; I would describe us as a satisfied group.

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

Q7. I like the way things are run at this school.

5. Strongly Disagree.
6. Disagree.
7. *Agree.*
8. *Strongly Agree.*

Q8. I think about transferring to another school.

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

Q9. I think about staying home from school because I'm just too tired to go.

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

Q10. Do you plan to return to this school next year?

- a. *Yes.*
- b. No.
- c. I don't know.

Appendix C

Density Tests

Figure C.1 shows the densities of the running variable with the rounding of scores that occurred in practice, centered on the discontinuity thresholds. These figures represent the data we use for the empirical analysis. The densities are smooth through the 2/3 and 3/4 thresholds but significant bunching at the 1-percent level is present at the 4/5 cutoff.

Figure C.2 replicates Figure C.1 but without allowing for the rounding of teacher scores. The densities remain smooth through the 2/3 and 3/4 thresholds and the bunching in the density at the 4/5 cutoff attenuates considerably. It is only significant at the 10-percent level.

Table C.1 reports regression results corresponding to the figures. See the text at the end of Section 4.2 for a discussion and interpretation of the density-test findings.

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

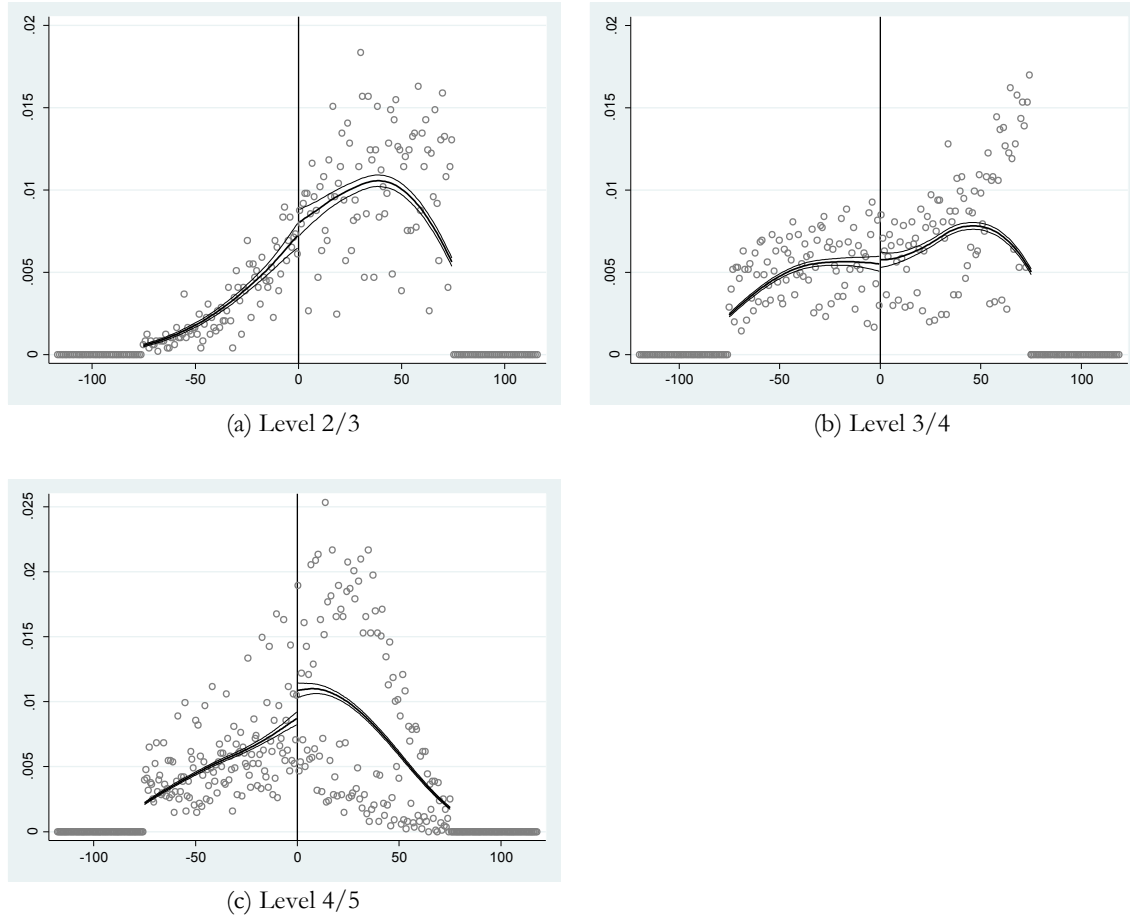


Figure C.2. Densities of Teachers' Underlying Score Variables Centered on the Scores that Determined Teachers' Ratings, without Allowing for Rounding.

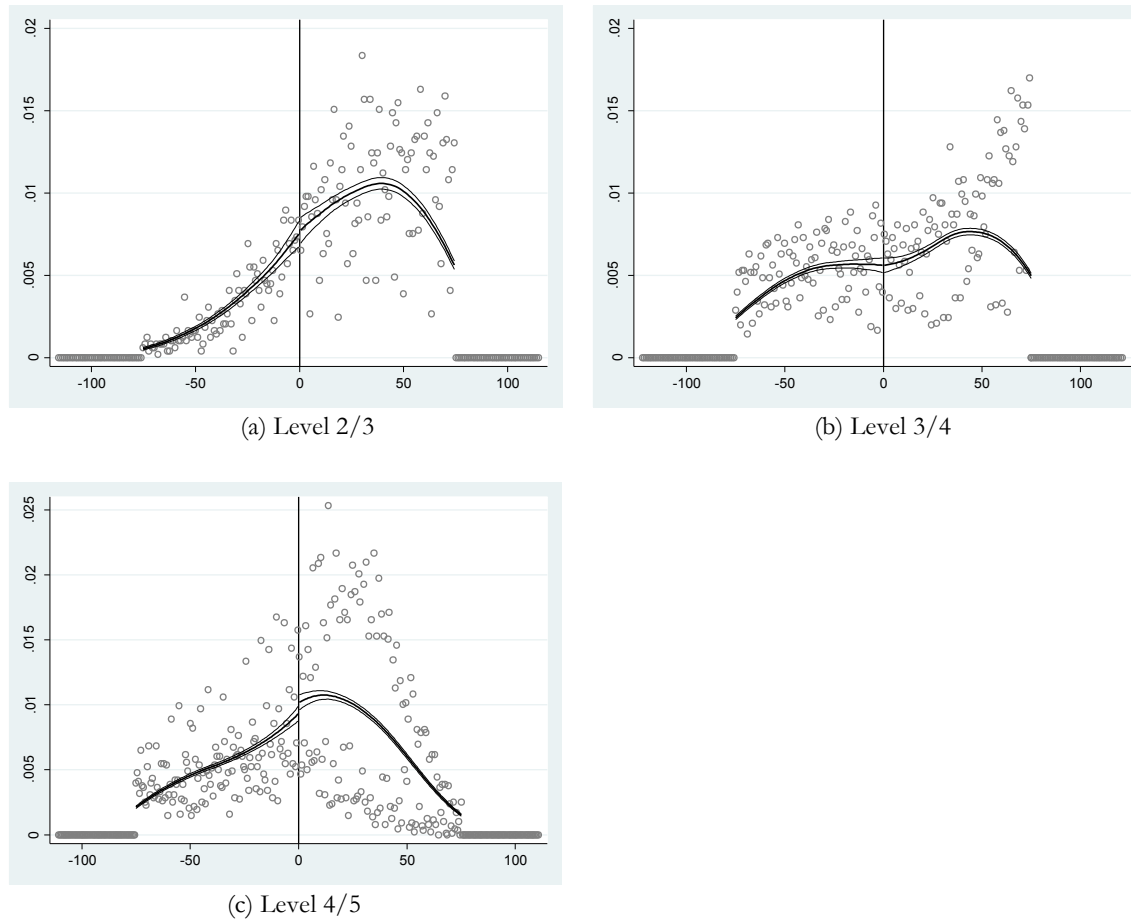


Table C.1. Estimated Discontinuities Using McCrary's Method at Different Cutoff Levels.

	Level 2-3	Level 3-4	Level 4-5
Using actual assignment variable	0.093 (0.075)	0.045 (0.061)	0.217 (0.039)***
Using non-rounded assignment variable	0.016 (0.076)	-0.003 (0.059)	0.084 (0.043)*

***/**/* 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
<i>General Satisfaction with Teaching</i>			
Question 1	1.029 (0.15)	1.137 (0.94)	1.145 (1.42)
Question 2	0.951 (-0.25)	1.150 (1.07)	1.082 (0.85)
Question 3	1.008 (0.04)	1.273 (1.74)*	1.156 (1.53)
Question 4	0.949 (-0.26)	1.185 (1.25)	1.083 (0.85)
<i>Satisfaction at School</i>			
Question 5	1.262 (1.13)	1.096 (0.64)	1.187 (1.59)
Question 6	1.133 (0.60)	1.140 (0.90)	1.216 (1.86)*
Question 7	1.009 (0.04)	1.066 (0.46)	1.258 (2.27)**
Question 8	1.107 (0.49)	1.020 (0.14)	1.145 (1.30)
Question 9	0.960 (-0.20)	1.189 (1.23)	1.076 (0.74)
Question 10	1.097 (0.37)	1.187 (1.00)	1.310 (2.17)**
Overall p-value	0.42	0.21	0.20
N	5443	10560	15000

***/**/* 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. P-values indicate the probability in each column of the observed number of relative risk ratios greater than one occurring by chance (based on 3,000 bootstrap repetitions).

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
<i>General Satisfaction with Teaching</i>			
Question 1	0.956 (-0.22)	1.101 (0.64)	0.952 (-0.45)
Question 2	0.888 (-0.60)	1.098 (0.65)	0.933 (-0.65)
Question 3	0.928 (-0.37)	1.143 (0.92)	0.962 (-0.37)
Question 4	0.923 (-0.40)	1.107 (0.71)	0.908 (-0.93)
<i>Satisfaction at School</i>			
Question 5	1.107 (0.47)	1.042 (0.25)	0.978 (-0.19)
Question 6	1.052 (0.25)	1.163 (0.97)	1.009 (0.08)
Question 7	1.000 (-0.00)	1.097 (0.60)	1.083 (0.71)
Question 8	1.061 (0.28)	1.004 (0.02)	0.981 (-0.17)
Question 9	0.880 (-0.58)	1.133 (0.82)	0.928 (-0.68)
Question 10	1.015 (0.06)	1.095 (0.53)	1.063 (0.46)
Overall p-value	0.50	0.23	0.61
N	5443	10560	15000

***/**/* 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. P-values indicate the probability in each column of the observed number of relative risk ratios greater than one occurring by chance (based on 3,000 bootstrap repetitions).

Appendix E

Robustness Analyses

Robustness to Controlling for Schooling Context

Table E.1 shows results analogous to the results in Table 6 but where we also include school-level 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 show in Table 6, 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.

We also considered constructing models that include school fixed effects. However, a challenge we faced is that the number of teachers who submitted a survey is small in many Tennessee schools. Specifically, the median number of teachers per school who submitted a survey in our data is 12, and at the 25th percentile it is just 7. The small school-level sample sizes are driven by (1) the fact that many schools in Tennessee are small, and (2) we restrict the analytic sample to teachers who submitted a survey. As a practical matter, attempting to include school fixed effects in our multinomial regressions caused convergence problems with the models. It is also conceptually unappealing for the RD to focus on within-school comparisons given the small school-level sample sizes. For example, at schools where there are few teachers to split across ratings categories 2, 3, 4 and 5 (e.g., at the 25th-percentile school with 7 teachers), the within-school comparisons at each threshold will be either sparse or unidentified. It is also noteworthy that Dee and Wyckoff (2013) estimate related RD models in their study with and without school fixed effects, and obtain substantively similar results regardless of whether school fixed effects are included in their models.²²

²² Dee and Wyckoff (2013) perform their evaluation in Washington DC, which is a large, urban school district. The Common Core of Data reports that as of the 2011-2012 school year, Washington DC had 3,472 FTE teachers and 127 schools, which translates to a teacher-to-school ratio of roughly 27. Furthermore, given the outcomes of interest in Dee and Wyckoff (2013), they do not need to restrict their analytic sample in a way that reduces school-level sample sizes like our restriction to include only individuals who submitted a survey.

Table E.1. Replication of Table 6 with Additional Controls for Schooling Context.

Dependent Variable	Level 2-3	Level 3-4	Level 4-5
<i>General Satisfaction with Teaching</i>			
Question 1	1.147 (0.96)	0.969 (-0.31)	1.097 (1.24)
Question 2	1.012 (0.09)	0.944 (-0.59)	1.104 (1.40)
Question 3	1.109 (0.77)	1.014 (0.14)	1.149 (2.00)**
Question 4	1.128 (0.89)	0.955 (-0.45)	1.036 (0.49)
<i>Satisfaction at School</i>			
Question 5	1.348 (2.06)**	0.902 (-0.93)	1.115 (1.31)
Question 6	1.288 (1.88)*	1.081 (0.75)	1.173 (1.98)**
Question 7	1.226 (1.51)	0.989 (-0.11)	1.327 (3.64)***
Question 8	1.309 (1.76)*	0.860 (-1.41)	1.129 (1.49)
Question 9	0.981 (-0.13)	1.011 (0.10)	1.057 (0.74)
Question 10	1.141 (0.71)	0.972 (-0.22)	1.200 (1.84)*
Overall p-value	0.11	0.81	0.03**
N	5443	10560	15000

***/**/* 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 in the preceding analysis, standard errors are clustered at the school level. P-values indicate the probability in each column of the observed number of relative risk ratios greater than one occurring by chance (based on 3,000 bootstrap repetitions).

Robustness to Alternative Bandwidths

Tables E.2 through E.7 show results from our main specification for each question at each threshold with narrower bandwidths around the discontinuity. Our primary results in Table 6 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. Our estimates based on the full bandwidths, as reported in Table 6, are displayed in the first row of each table for comparison.

Tables E.2 and E.3 show results for the general and school-specific satisfaction questions, respectively, for all three alternative bandwidths for the 2/3 threshold. Tables E.4 and E.5 present analogous information at the 3/4 threshold, and Tables E.6 and E.7 at the 4/5 threshold. The most notable deviations from our primary findings in Table 6 are in Tables E.2 and E.4. In Table E.2, all coefficients go from being nominally above 1.0 to nominally below 1.0 as we shrink the bandwidth from 75 to 40, and a reverse pattern is observed in Table E.4. However, even in these cases none of the differences imply statistically significant changes to our findings as reported in the text. The other results in this section indicate even less sensitivity of our findings to bandwidth choice.

Table E.2. Robustness of Findings to Alternative Bandwidths. Cutoff for Ratings 2/3, Job Satisfaction Questions 1-4.

Bandwidth	N	Q1	Q2	Q3	Q4
Full	5443	1.139 (0.90)	1.007 (0.05)	1.095 (0.68)	1.125 (0.87)
$ S_i \leq 60$	4481	1.096 (0.60)	0.981 (-0.12)	1.102 (0.69)	1.017 (0.12)
$ S_i \leq 50$	3822	1.035 (0.20)	0.867 (-0.86)	0.968 (-0.21)	0.945 (-0.35)
$ S_i \leq 40$	3104	0.934 (-0.37)	0.898 (-0.60)	0.942 (-0.35)	0.929 (-0.42)

***/**/* 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 6 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.

Table E.3. Robustness of Findings to Alternative Bandwidths. Cutoff for Ratings 2/3, Job Satisfaction Questions 5-10.

Bandwidth	N	Q5	Q6	Q7	Q8	Q9	Q10
Full	5443	1.360 (2.13)**	1.303 (2.00)**	1.242 (1.62)	1.338 (1.91)*	0.984 (-0.11)	1.172 (0.86)
$ S_i \leq 60$	4481	1.336 (1.86)*	1.329 (1.99)**	1.283 (1.71)*	1.320 (1.69)*	0.903 (-0.64)	1.109 (0.51)
$ S_i \leq 50$	3822	1.281 (1.47)	1.320 (1.76)*	1.156 (0.90)	1.338 (1.63)	0.792 (-1.34)	1.225 (0.93)
$ S_i \leq 40$	3104	1.382 (1.71)*	1.352 (1.75)*	1.209 (1.08)	1.426 (1.82)*	0.751 (-1.50)	1.253 (0.96)

Notes: See notes from Table E.2.

Table E.4. Robustness of Findings to Alternative Bandwidths. Cutoff for Ratings 3/4, Job Satisfaction Questions 1-4.

Bandwidth	N	Q1	Q2	Q3	Q4
Full	10560	0.966 (-0.33)	0.944 (-0.60)	1.015 (0.15)	0.956 (-0.45)
$ S_i \leq 60$	8034	0.998 (-0.02)	0.988 (-0.11)	1.049 (0.44)	1.050 (0.43)
$ S_i \leq 50$	6546	1.077 (0.61)	1.011 (0.09)	1.112 (0.92)	1.168 (1.27)
$ S_i \leq 40$	5113	1.134 (0.92)	1.026 (0.19)	1.156 (1.13)	1.134 (0.94)

Notes: See notes from Table E.2.

Table E.5. Robustness of Findings to Alternative Bandwidths. Cutoff for Ratings 3/4, Job Satisfaction Questions 5-10.

Bandwidth	N	Q5	Q6	Q7	Q8	Q9	Q10
Full	10560	0.896 (-1.00)	1.070 (0.64)	0.981 (-0.18)	0.853 (-1.48)	1.009 (0.09)	0.967 (-0.26)
$ S_i \leq 60$	8034	0.938 (-0.53)	1.142 (1.12)	1.028 (0.24)	0.846 (-1.39)	1.111 (0.91)	0.966 (-0.24)
$ S_i \leq 50$	6546	0.875 (-1.00)	1.158 (1.14)	1.015 (0.12)	0.842 (-1.31)	1.167 (1.20)	0.941 (-0.39)
$ S_i \leq 40$	5113	0.919 (-0.55)	1.199 (1.24)	1.086 (0.59)	0.845 (-1.14)	1.227 (1.43)	0.917 (-0.48)

Notes: See notes from Table E.2.

Table E.6. Robustness of Findings to Alternative Bandwidths. Cutoff for Ratings 4/5, Job Satisfaction Questions 1-4.

Bandwidth	N	Q1	Q2	Q3	Q4
Full	15000	1.095 (1.22)	1.104 (1.40)	1.146 (1.98)**	1.030 (0.43)
$ S_i \leq 60$	13604	1.120 (1.41)	1.131 (1.63)	1.146 (1.83)*	1.047 (0.61)
$ S_i \leq 50$	12245	1.142 (1.55)	1.112 (1.30)	1.177 (2.10)**	1.044 (0.53)
$ S_i \leq 40$	10417	1.158 (1.53)	1.165 (1.68)*	1.181 (1.95)*	1.024 (0.27)

Notes: See notes from Table E.2.

Table E.7. Robustness of Findings to Alternative Bandwidths. Cutoff for Ratings 4/5, Job Satisfaction Questions 5-10.

Bandwidth	N	Q5	Q6	Q7	Q8	Q9	Q10
Full	15000	1.116 (1.32)	1.176 (2.01)**	1.325 (3.36)***	1.138 (1.59)	1.059 (0.76)	1.207 (1.90)*
$ S_i \leq 60$	13604	1.101 (1.08)	1.180 (1.92)*	1.362 (3.79)***	1.223 (2.29)**	1.069 (0.82)	1.238 (1.98)**
$ S_i \leq 50$	12245	1.128 (1.25)	1.167 (1.69)*	1.362 (3.57)***	1.181 (1.77)*	1.059 (0.66)	1.256 (1.99)**
$ S_i \leq 40$	10417	1.113 (1.02)	1.140 (1.33)	1.319 (2.94)***	1.150 (1.34)	1.086 (0.86)	1.242 (1.75)*

Notes: See notes from Table E.2.

Robustness to Excluding Teachers Who Report that They Did Not Receive a Rating

Table E.8 reports results analogous to Table 6 after restricting the sample to exclude teachers who reported that they did not receive their evaluation ratings from the system. As noted in the text, approximately ten percent of teachers in the analytic dataset (n=2,174) reported that they did not receive their ratings. Also note that some teachers did not answer the question about whether they received a rating – we only exclude teachers in Table E.8 who explicitly indicated that they did not receive their ratings.

Table E.8. Effects of Ratings on Job Satisfaction for Teachers Who Report Receiving their Ratings.

Dependent variable	Level 2-3	Level 3-4	Level 4-5
<i>General Satisfaction with Teaching</i>			
Question 1	1.249 (1.45)	1.015 (0.13)	1.113 (1.35)
Question 2	1.068 (0.42)	1.005 (0.05)	1.153 (1.88)*
Question 3	1.156 (0.99)	1.106 (0.95)	1.214 (2.63)***
Question 4	1.200 (1.25)	0.961 (-0.37)	1.078 (1.01)
<i>Satisfaction at School</i>			
Question 5	1.374 (2.12)**	0.960 (-0.34)	1.212 (2.11)**
Question 6	1.294 (1.82)*	1.162 (1.30)	1.252 (2.52)***
Question 7	1.202 (1.28)	1.067 (0.57)	1.376 (3.83)***
Question 8	1.276 (1.56)	0.953 (-0.41)	1.189 (2.00)**
Question 9	1.125 (0.73)	1.078 (0.67)	1.093 (1.10)
Question 10	1.189 (0.89)	1.034 (0.24)	1.174 (1.47)
Overall p-value	0.04**	0.31	0.04**
N	4697	9274	13589

***/**/* 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. P-values indicate the probability in each column of the observed number of relative risk ratio greater than one occurring by chance (based on 3,000 bootstrap repetitions).