Performance, Process, and Interpersonal Relationships: Explaining Principals' Perceptions of Principal Evaluation

Jennifer L. Nelson, Jason A. Grissom, Margaux L. Cameron

Published in *Educational Administration Quarterly*, First Published April 24, 2021, doi: https://doi.org/10.1177/0013161X211009295

This study was supported by a grant from the Institute of Education Sciences, U.S. Department of Education, through Grant R305B170009 to Peabody College at Vanderbilt University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract:

Purpose: Multiple-measure principal evaluation systems have become commonplace in the last decade, but we do not know how principals perceive their evaluations under these regimes. This study analyzes how principals perceive evaluation in a state that was an early adopter of such a system. It describes how attitudes are explained by individual and contextual factors, performance ratings, and elements of the evaluation process.

Research Methods: Using data from a statewide survey of Tennessee principals in three consecutive school years, we create an index of principal evaluation perceptions of evaluation, then employ regression analysis to predict principals' attitudes with measures gleaned from survey and administrative data sources.

Findings: High school and veteran principals have more negative views of their evaluations. Practice ratings from the principal's supervisor, though not the overall evaluation score, are positively correlated with attitudes. Principals assigned ratings more often view evaluation more positively, even accounting for their rating, as do principals who have worked longer with their evaluator. We find no evidence that racial or gender matching between principals and raters leads to more positive perceptions, and in fact Black principals may perceive evaluation more negatively when their evaluator is Black.

Implications: Our results suggest some directions for states and districts seeking to make evaluation more meaningful for principals. Principals appear to value both frequency of feedback and consistency in raters over time. These factors may be especially important for low-rated principals, veteran principals, and those in secondary schools, who may perceive less value from principal evaluation.

Performance, Process, and Interpersonal Relationships: Explaining Principals' Perceptions of Principal Evaluation

Educator evaluation serves multiple purposes. Evaluation results can be used to focus efforts for school improvement (Archer, Kerr, & Pianta, 2014), make human capital decisions (Goldring et al., 2015), allocate professional development (Ruzek, Hafen, Hamre, & Pianta, 2014), and meet school accountability mandates (Lavigne, 2018). These potential uses have motivated widespread adoption of multiple-measures educator evaluation systems in nearly all states in the past decade, particularly in the wake of Race to the Top (RTTT) (Author, 2016; Steinberg & Donaldson, 2016). Importantly, RTTT—and, later, No Child Left Behind (NCLB) waivers—pushed states to reform not just *teacher* evaluation but *leader* evaluation as well (McGuinn, 2012). Nearly 40 states have reformed their leader evaluation in recent years to base practice ratings on standardized rubrics and tie evaluation scores explicitly to school outcomes (Author, 2018; Superville, 2014).

Yet despite this substantial paradigm shift in school leader evaluation, we know little about how principals are experiencing new systems that collect and provide feedback on performance information. Historically, principals have felt that the implementation of evaluation systems lacked transparency and clarity of purpose and gave them little incentive or guidance to improve (Condon & Clifford, 2012; Reeves, 2004). Principals often viewed systems as lacking fairness, accuracy, or face validity, which scholars have suggested led them to ignore evaluation feedback or divest from the evaluation process (Fuller, Hollingworth, & Liu, 2015). Yet changes in the new era of principal evaluation suggest the potential for principals' views to have shifted. These new systems have incorporated multiple measures of performance, including established procedures for rating principal practice, increased frequency of observation and feedback, and

feedback from multiple stakeholders (Author, 2018; Brown-Sims, 2010; Fuller, Hollingworth, & Liu, 2015; Rothman, 2017). Evidence suggests that these new practices have made evaluation feedback a more accurate reflection of principals' work (Author, 2018). Research has not explored, however, the degree to which these large changes to the structures of evaluation have translated into changes in how principals perceive performance evaluation systems and how those systems treat them.

We help to fill this gap by investigating principals' attitudes about evaluation in the context of an established statewide multiple-measure principal evaluation system. These attitudes likely matter in multiple ways. First, research from outside education shows how attitudes toward evaluation, such as how useful or fair it is perceived to be, affect what employees learn from it (Bell, Tannenbaum, Ford, Noe, & Kraiger, 2017; Warr, Allan, & Birdi, 1999). Employees are more likely to be motivated by performance feedback that they view as helpful, which can improve their subsequent performance (Bell & Ford, 2007; Northcraft, Schmidt, & Ashford, 2011). In addition, feelings about evaluation can inform how they perceive the organization's treatment of them, which in turn influences their job satisfaction, stress, and organizational commitment (Ford, Truxillo, & Bauer, 2009; Lind, 2001; Wexley, Singh, & Yukl, 1973).

We draw on organizational justice literature to identify three aspects of performance evaluation that are likely to shape principals' attitudes about their evaluations: what scores they receive, features of the evaluation process (such as how frequently they are observed), and the nature of the relationship between the principal and the evaluator. We investigate these factors using data from Tennessee, an early adopter of multiple-measure evaluation for principals, beginning in 2011-12. We make use of data from the Tennessee Educator Survey (TES), a

statewide survey of educators that includes responses from approximately half of the principals in the state each year. Across multiple years, the TES asked principals about their experiences with the state's evaluation system and their assessment of its fairness and value to their work.

Merging three years of these attitudinal data with state administrative records, we investigate two main questions. First, descriptively, what are principals' attitudes toward the evaluation system, and how do they vary over time and by individual and school characteristics? Second, to what extent do attitudes depend on how principals experience the implementation of evaluation, including the scores it assigns them, how frequently they are observed, and characteristics of the rater? On this latter point, we specifically investigate racial and gender similarity between the principals and their evaluator, given evidence that teacher evaluation outcomes vary when they share these characteristics with their raters (Author, 2019; Drake, Auletto, & Cowen, 2019). Results of this inquiry can not only illuminate the factors that determine how principals experience multiple-measure leader evaluation systems but provide insights for efforts to improve implementation of those systems.

Principal Evaluation and Organizational Justice

Research on principal evaluation systems lags far behind the parallel literature on teacher evaluation (Author, 2018). Prior to the last decade, studies documented a lack of alignment between content and structure of the evaluation instrument on the one hand, and professional standards and best practices for school leadership on the other (Condon & Clifford, 2012; Goldring et al., 2009). For example, in a content analysis of 100 principal evaluation instruments in the state of Virginia, many instruments failed to capture the full range of principals' roles (Catano & Stronge, 2007). Principals subject to these misaligned systems reported that they did not feel they learned much from evaluations that they perceived as unhelpful checklists that did

not accurately capture their performance (Davis & Hensley, 1999; Lashway, 2003). Reflecting the absence of standardization, principals subject to such systems varied widely in their perceptions of their evaluations' purpose and focus, variation that may have explained whether evaluation informed their leadership practice (Sun & Youngs, 2009).

More recent studies of post-Race to the Top principal evaluation systems highlight various improvements to their construction. For example, rubrics have become more elaborated and tied more closely to professional standards. These rubrics better describe expectations for principal performance with suggestions for evidence raters can examine to provide scores on the rubric. Refined evaluation processes also feature clearer expectations for evaluators around evaluation feedback and its communication (Kimball et al., 2015). Moreover, newer evaluation systems incorporate multiple measures to capture principal performance, including measures based on student achievement. While these changes address some of the criticisms leveled by Lashway (2003) and others, they also introduce other issues. For instance, the use of student performance data, including school value-added, in evaluating principals raises concerns about fairness, as research on these measures questions whether they indeed reflect principal performance (e.g., Author, 2015; Fuller & Hollingworth, 2014). Other evidence suggests that raters may face challenges in differentiating principals' performance with complex rubrics or provide lower ratings to principals based on factors beyond their control (Author, 2018).

Research has just begun to document how principals perceive evaluation under this new paradigm. In an experiment in one large district, principals randomly assigned to be evaluated with a standards-based rubric were more satisfied with the process and rated feedback quality higher than principals assigned to business as usual, though the authors concluded that overall effects were muted due to variation in implementation across raters (Kimball, Milanowski, &

McKinney, 2009). In a study of principal evaluation in six urban districts in the process of reforming evaluation as part of a broader initiative, Anderson and Turnbull (2016) find that most principals believed that their evaluations reflected the complexity of their role and accurately reflected their performance, but interviews with district leaders and principals raised concerns about consistency of evaluation practice across raters and whether every rater could use the process to drive authentic feedback conversations.

Studies of performance appraisal outside education have documented different facets of employee perceptions, including assessments of fairness, usefulness, and clarity, which often are related to one another (DeNisi & Murphy, 2017). Indeed, studies grounded in the organizational justice tradition suggest that fairness may be central. *Organizational justice* refers to how employees of an organization judge the correctness or rightness of their treatment by the organization (Cropanzano & Ambrose, 2015). Perceptions of unjustness can lead to lower job satisfaction, burnout, and turnover (Colquitt, Conlon, Wesson, Porter, & Ng, 2001).

Multiple models of organizational justice have been proposed, but scholars agree on at least two dimensions: distributive justice and procedural justice. *Distributive justice* refers to fairness in the distribution of outcomes, such as recognition or pay, among employees (Jost & Kay, 2010; Moorman, 1991). *Procedural justice* refers to the fairness of the organizational processes, including the process's consistency, representativeness, and capacity to be corrected, that lead to those outcomes (Leventhal, Karuza, & Fry, 1980). Applied to performance evaluation, a distributive justice perspective suggests that employees are more likely to have negative views when they receive lower scores, especially in relation to their peers whom they assume have put forth similar efforts. Consistent with this prediction, teachers rated less effective describe teacher evaluation as less helpful to their teaching (Ravenell, 2019).

From a procedural justice perspective, how evaluation is implemented will also affect evaluation attitudes, potentially independently of employees' ratings (Folger, Konovsky, & Cropanzano, 1992). Research on principal evaluation has emphasized the importance of the quality of processes used (Davis, Kearney, Sanders, Thomas, & Leon, 2011). Frequency of observing performance is an important example. Under earlier evaluation systems, teachers and administrators perceived evaluations based on infrequent observations as less effective and that increasing opportunities to sample performance would improve evaluation processes (Xu & Sinclair, 2002; Doherty, 2009). Frequent visits benefit the sense of guidance and understanding the principal perceives (Thomas, Holdaway, & Ward, 2000) and frames evaluation as an ongoing, yearlong process that is better aligned with the school and district's goals (Parylo, Zepeda, & Bengtson, 2012).

Characteristics of evaluators may also contribute to whether employees see evaluation outcomes and processes as just. An evaluator with more years conducting evaluations, for example, may be viewed as better able to rate quality of practice. More experienced teacher evaluators provide ratings that are more consistent with external measures of teacher performance (Rockoff, Staiger, Kane, & Taylor, 2012), perhaps because they better understand observation protocols and how to use them (Bell et al., 2014). Raters who have worked with the employee longer may also provide more accurate ratings (Jacob & Lefgren, 2008). They may also have had time to develop a closer working relationship, helping principals feel more comfortable with receiving feedback and facilitating ongoing conversations about performance (Stronge, 1991; Wilson & Natriello, 1989). What position the evaluator holds may also matter. Ratings may be viewed as more authoritative coming from the district superintendent, for example. Superintendents may be better positioned to mentor principals around evaluation than

other district leaders (France & Thompson, 2015). On the other hand, in larger districts, evaluations from superintendents may be viewed as less accurate or complete, given competition from other aspects of the role that can impede the quality of evaluation implementation (Kimball & Pautsch, 2008). In more general terms, beyond objective observer characteristics, the quality of the relationship between principal and evaluator can vary widely. Evaluators who are consistent, who engage the principal in reflective conversation, and approach goal setting as a partnership rather than dictating goals to principals all engender more genuine, trusting relationships (DeMatthews, Scheffer, & Kotok, 2020).

Organizational justice can extend beyond narrow conceptions of fairness to include the interrelated concepts of helpfulness (Folger & Konovsky, 1989) and specificity of feedback (Levy, Cavanaugh, Frantz, & Borden, 2015), as well as clarity of the evaluation process (Folger & Konovsky, 1989). Ratings and feedback are only helpful insofar as what is being evaluated is transparent. Similarly, specificity signals transparency of criteria and the evidence base for ratings, and makes clearer the ways in which employees can improve their performance. Later, we test the degree to which evaluation helpfulness, specificity, process clarity, and fairness are related in principals' minds.

Of course, principals' formulations of perceptions of evaluation processes and outcomes also draw on their own lived experiences, which are a function of their individual characteristics and those of their local context. Educators' background characteristics, including gender, race, experience, and education level, are important predictors of work-related attitudes, broadly (e.g., Mueller, Finley, Iverson, & Price, 1989; Renzulli, Parrott, & Beattie, 2011). Gender and race, in particular, can determine experience with disparate treatment within organizations (DeNisi & Murphy, 2017; Dipboye, 1985). In the broader literature on personnel evaluation, there is some

evidence in experimental vignette studies and large-scale studies of better ratings for performance capacity being given to younger and white employees (Rosen & Jerdee, 1976; Pulakos, White, Oppler, & Borman, 1989). Research on teacher evaluation identifies advantages for female and white teachers (Author, 2019; Campbell & Ronfeldt, 2018), as well as novice and veteran teachers, as compared to those in the middle range of experience (Author, 2017). This research also suggests that teachers score higher when they are observed by a rater with the same demographic characteristics, especially the same race (Author, 2019; Drake et al., 2019), suggesting that characteristics of both principals and their evaluators may factor into their perceptions of how the evaluation system treats them.

School contextual factors may also be relevant. School context can inform what ratings a principal receives. For example, Author (2018) finds that the fraction of low-income students in a school predicts lower principal evaluation ratings, even when comparing ratings for the same principal in different contexts. These findings echo similar conclusions about classroom composition determining teacher observation ratings (Campbell & Ronfeldt, 2018). School context may impact employee attitudes for other reasons as well. Conditions that create job stress, such as having too few resources to meet the demands of the student population an educator serves, can negatively impact perceptions of the workplace (Renzulli et al., 2011). Similarly, a lack of resources at the district level for evaluator visits, mentoring, and feedback can also negatively influence principals' views on their evaluation (Thomas et al., 2000).

Resource issues impacting efficacy of evaluations, and quality of evaluator-principal relationship, may be related to district size and school level as well. For instance, principals' critiques of a new principal evaluation system in a medium-sized urban district included shaky district supports, such as principals feeling overburdened by bureaucratic tasks (including

elements of the evaluation process itself) and not knowing where to turn for support when their evaluator identified need for improvement (DeMatthews, Scheffer, & Kotok, 2020). DeMatthews and colleagues (2020) found that principals viewed their evaluation positively when their relationship with the evaluator was somewhat personalized, enough so that the evaluator took into account nuances about the school and principal and allowed their feedback sessions to be genuine and not formulaic. It could be that such relationships are less likely to develop in larger school systems on account of the scale of the bureaucracy. District size also affects principals' perceptions of what their evaluator cares most about, such that evaluators in smaller districts focus on daily operations and community communications more than those in larger districts (Muenich 2014). While no prior studies definitively addresses the role of school level (i.e., elementary, middle, high) in shaping principal perceptions of evaluation, DeMatthews and colleagues included only elementary principals in their study, and Muenich surveyed only secondary principals. Contrasting the studies, the kinds of student data the principals could or desired to include in their evaluations differed; only secondary principals mentioned graduation rates, college entrance exams, GPA, and discipline/suspension data. This difference suggests that secondary principals' evaluations may be oversimplified in their view, compared to the range of criteria they could be evaluated on.

Data and Methods

Setting

We make use of data from Tennessee, an early adopter of multiple-measure principal evaluation. Beginning in the 2011-12 school year, Tennessee mandated that teachers and leaders statewide be evaluated each year using multiple measures as part of the First to the Top Act,

passed in pursuance of the state's successful Race to the Top (RTTT) application.[1] This requirement has persisted through the state's successful application in the NCLB waiver process and into the present-day Every Student Succeeds Act (ESSA) era, despite ESSA's loosening of federal expectations around educator evaluation. The statewide system is called the Tennessee Educator Acceleration Model (TEAM). The state also approved alternative rating systems in a few districts, though those systems share many similarities with the statewide system, so we focus here on TEAM. Under TEAM, each principal is assigned an overall Level of Effectiveness (LOE) score at the end of each school year. LOE is a five-point scale spanning *significantly below expectations* (1) to *significantly above expectations* (5). Principals scoring at least a 3 are considered to be meeting expectations. The state has no requirements for how principals' scores are used, beyond provision of performance feedback, so their implications vary according to local district policy; in some districts, scores lower than 3 may subject the principal to an improvement plan or other personnel action, or a high score may entitle the principal to a pay increase. Prior research shows that low-rated principals are more likely to be moved into non-leadership positions or to exit the system (Author, 2018).

LOE is comprised of three components. Fifty percent comes from practice ratings given to the principal by their supervisor (the superintendent or a designee) using the TEAM rubric, 35% from the school's value-added score (referred to as "TVAAS"),[2] and 15% from additional measures of achievement as determined by mutual agreement of the principal and their

[1] The state's Race to the Top win secured $501 million to support education reform in the state.
[2] TVAAS stands for "Tennessee Value-Added Assessment System." TVAAS scores, calculated by the SAS Institute, are based on student growth on end-of-grade tests in grades 3–8 and end-of-course tests in high school. For each test, schools are scored according to their students' performance relative to predictions from schools across the state whose students have similar achievement trajectories. Schools are then given a single score based on the average of growth on each test, weighted by the number of students who took each test.

evaluator.[3] The TEAM rubric is based on the Tennessee Instructional Leadership Standards, which define effective leadership practice for the state. The rubric groups principal practices into four categories: instructional leadership for continuous improvement, culture for teaching and learning, professional learning and growth, and resource management. Within these categories, principals are scored on 17 indicators on a scale from 1 to 5 (as with LOE, *significantly below expectations* to *significantly above expectations*). The TEAM rubric suggests possible sources of evidence to the rater in scoring each indicator. According to State Board of Education guidance, administrators are expected to be rated at least twice per academic year, once in fall and once in spring, with the average of all completed practice ratings used for that portion of the LOE. Raters must complete a training led by state officials to be certified to evaluate principals.

Prior research on the TEAM principal evaluation system found that raters do not differentiate among indicators, on average—that is, the 17 scores capture a single underlying performance construct—but that the average ratings predict other plausible principal performance measures, such as teachers' survey-based ratings of leadership quality in the school, providing evidence of concurrent validity (Author, 2018). Research also finds evidence of predictive validity; for example, principals who receive higher practice ratings see higher rates of teacher retention the next year, particularly among effective teachers (Author, 2019).

Survey and Administrative Data

We make use of deidentified survey and administrative data accessed through the Tennessee Education Research Alliance (TERA), a research-practice partnership between

[3] The achievement measure that comprises 15% of the final score is chosen locally from an approved list of possible achievement metrics. School districts decide the rating criteria associated with the metrics. Our data, however, do not include those criteria—only the metric used (e.g., standardized test score composite) and the score assigned. Most principals were scored from a from some form of standardized assessment—either a statewide one or a test chosen locally—though scores for 25% of principals were derived from either their school's or district's graduation rate. Because we do not have information about the metrics themselves or the processes used to choose or score them, we exclude them from our analysis.

Vanderbilt University and the Tennessee Department of Education (TDOE). The Tennessee Educator Survey (TES), jointly administered by TERA and TDOE, is an annual web-based survey of all teachers and leaders in the state each spring. TES responses are confidential but not anonymous; numerical identifiers permit linkage of survey information to other data sources for research purposes. We make use of data from the springs of 2015, 2016, and 2017, in which comprehensive questions about evaluation attitudes were included on the principal survey. Across these years, the average response rate was 59 percent. Administrators' perceptions of the evaluation process were measured by nine survey items,[4] each with a 4-point response scale.[5] Respondents indicated their level of agreement regarding: the specificity of feedback, the degree to which evaluation feedback offered guidance for improvement; the fairness and helpfulness of the evaluation process; the clarity of the rubric used; whether the respondent made changes in their practices as a result of evaluation; and their overall satisfaction with the process. The specific items are shown in Table 1. Principals also reported the number of times they recalled an evaluator had observed them that year, which ranged from 0 (N = 97) to 3 or more (N = 694). The median was 2.

We merge survey responses with staff and school context information from administrative data files. These data include information on the individual characteristics of principals (e.g., gender, race) and their schools (e.g., school size, grade span, student demographic characteristics). Experience as a principal is not captured in the administrative data, but we construct this measure using educators' job history information, which is available from

[4] The TES contains two additional attitudinal measures capturing how useful principals perceive evaluations and how "negatively focused" they think evaluation is. However, exploratory analysis of these items showed that the distributions for these items changed substantially from year to year, suggesting potential problems with how the items were coded. Relatedly, these items had very low correlations with the other evaluation items. We thus dropped them from our analysis.

[5] Participants could also respond "N/A" to the first three items. Respondents rarely chose this response. We recoded N/A responses as missing for the purposes of this analysis.

2001-02 forward.[6] We use achievement level information to create a summary achievement index for each school that is the weighted average of all state standardized test scores (across grades and subjects) each year.[7] Table 2 describes the characteristics of the analytic sample, pooled across years. Principals in the sample are 59% female and 9% Black. About half hold advanced specialist or PhD degrees, and about half have five or more years of experience as a principal. The mean school enrollment is 626 students, and the mean free and reduced price lunch eligibility at the school level is 60%. About 60% of principals lead elementary schools, and 44% lead rural schools. The final column of Table 2 shows means for the full population of principals and schools in the state, beyond the survey sample. These numbers suggest that the survey sample generally is similar to the underlying population for most characteristics, with the exception that Black and urban principals and principals in schools with larger populations of Black students are somewhat underrepresented, reflecting lower TES participation in Memphis and Nashville, two urban districts with larger numbers of Black principals and students.

Staff data also include information about principals' evaluation, including their summative LOE rating and its components, the TEAM practice rating and school TVAAS.[8] In the models we describe below, we lag these values by one year on the assumption that principals do not know their current-year ratings at the time they complete the TES, so their perceptions are more likely to be informed by last year's ratings. As shown in Table 2, the mean of lagged LOE is 3.85, while mean lagged practice ratings and TVAAS were 3.91 and 3.14, respectively.[9] These

[6] We cannot specify years as a principal for principals who were already working as a principal in the first year of the data set; we only know a minimum. This variable is therefore top-coded.
[7] Because test score data was sparser in 2016 due to testing problems statewide that year, we imputed the average of the achievement index for 88 cases based on index value(s) for prior and/or successive years.

[8] Files also include information on the achievement measure that makes up 15% of the LOE, but we do not consider this component, as choice of this metric varies across the state.
[9] Supplemental correlation analyses show that lagged LOE and TVAAS are highly correlated ($r = 0.80$), lagged LOE and practice ratings are only moderately correlated ($r = 0.43$), and lagged practice ratings and TVAAS are weakly

means closely align with state means. Because only 62 respondents received an LOE score of 1, we collapsed LOE scores of 1 and 2 into the lowest category (N = 277 total).

We can also glean some evaluation process information from the observation files. We can see the number of practice ratings or observations—we use these terms interchangeably—conducted during the school year (172 principals were observed once, 2,062 were observed twice, and 30 were observed 3 or 4 times); these counts are distributed less evenly than the survey-reported totals, which may reflect the timing of the survey, respondents' recall challenges, or principal perceptions of what counted as a formal observation that differed from how they were entered into the performance evaluation system. Each rating also identifies a rater, who we can link to their own personnel information. Evaluators were mostly superintendents (44%) or supervisors (32%), with 9% classified as assistant superintendents and 16% as "other," a category including central office, federal and special programs, human resources, assessment personnel, and school improvement and accountability employees. The typical rater had 5.4 years of experience in their current position but just less than two years of experience assigning ratings under the TEAM system in the years of our data. Comparing principals' and raters' job histories, we see that principals had a mean of 2.06 years being paired with their rater in the evaluation system, but a mean of 10 years working in the same district as the rater. Reflecting somewhat lower representation of women and people of color in district than in school leadership, raters are less likely to be female (50%) or Black (7%) than are the principals they evaluate.

Methods

correlated (r = 0.14). These correlations are very similar to current-year score correlations. They suggest that variation in the LOE is driven more by variation in TVAAS than variation in practice ratings. Indeed, practice ratings tend to be more stable from year-to-year—the correlation between the current and lagged rating is 0.67—than LOE and TVAAS, each of which has a year-to-year correlation of approximately 0.25.

Our research questions are largely descriptive. We first describe principals' attitudes about their evaluations overall and over time, then explore the dimensions of those attitudes using factor analysis. Next, we ask what principal and school characteristics predict perceptions of evaluations. Formally, we estimate the following model:

$$Y_{isdt} = \beta_0 + P_{isdt}\,\beta_1 + S_{sdt}\,\beta_2 + \tau_d + \gamma_t + \varepsilon_{isdt} \quad (1)$$

where attitude of a principal i at school s in district d is a function of principal characteristics P (gender, race, highest degree, years of experience as a principal), school characteristics S (student enrollment, fraction of students eligible for the federal free and reduced price lunch program, fraction of Black and Hispanic students, the school's student achievement index score, level of school, district enrollment size, and school locale type), year fixed effects γ_t, and a random error term ε_{isdt}. In some models, we include district fixed-effects τ_d to account for unobserved district-specific factors that may affect principals' perceptions of their evaluations. Models are estimated via ordinary least squares. We cluster standard errors at the district level.

We then turn to how evaluation outcomes and processes are associated with perceptions. We add measures of principal performance to equation (1). These measures include lagged-year LOE, practice rating averages, and TVAAS.[10] In subsequent models we add measures of principal's self-reported frequency of observation or the frequency recorded in administrative data in that year. We also investigate evaluator characteristics by adding measures of the evaluator's years of experience as an evaluator and in their current position, years the evaluator has evaluated that principal, years the evaluator and principal have worked in the same district

[10] As previously noted, current-year LOE and TVAAS clearly are unknown to the principal at the time the survey is taken. Principals may already have received one ore more practice ratings, though our investigation of the timing of practice ratings showed that most final ratings came after the survey window.

together, an indicator for change in evaluator in that year from the prior year, and an indicator for whether the evaluator is a superintendent.

In a final analysis, we explore the association between attitudes and evaluator/principal demographic matching. We augment equation (1) with an indicator for whether the evaluator is female and an indicator for the evaluator is Black, plus interaction terms for *principal is female x evaluator is female* and *principal is Black x evaluator is Black.* Note that we cannot consider other racial or ethnic groups because of the very small sample sizes of non-white, non-Black leaders and evaluators in Tennessee.

Principals' Attitudes about Evaluation in Tennessee

Table 1 describes principals' responses to the evaluation attitude questions on the TES, pooled across years. Two items measured quality of feedback (whether it identified specific areas of practice for improvement and whether it included guidance for improvement). Principals rated these items as 3.5 and 3.3, respectively, on average, suggesting that most principals viewed these statements as true or mostly true. The remaining seven items capture other aspects of the evaluation process, including its fairness and whether the rubric clearly defines expectations, on a four-point Likert agreement scale. Means in this set cluster between 3.0 and 3.2, suggesting generally positive reviews of the evaluation system, though responses vary, with standard deviations in the neighborhood of 0.7 for these items. Figure 1 shows that these perceptions have become somewhat more positive over time.

Next, we explored the relationships among items. An initial correlation matrix showed that two of the nine survey items (regarding how useful or negatively focused principals perceived evaluation to be) had very low correlations with other items (r = 0.09–0.30). Exploratory factor analysis (EFA) using promax rotation also suggested that these two items did

not relate closely to the others, so we dropped them from the analysis. With these items eliminated, we used EFA to explore whether how many factors the remaining seven items identified. We found that a single-factor solution fit the data best, following the confirmation procedure recommended by Ferguson and Cox (1993).[11] This model had both strong internal consistency (Cronbach's $\alpha = 0.87$) and good statistical fit ($\chi 2$ (13) = 248.09, SRMR = 0.037, CFI = 0.974, RMSEA = 0.092).[12] In other words, principals' survey responses about evaluation fairness, specificity, and so forth (i.e., all seven dimensions listed in Table 1) appear to measure a single underlying construct. We use factor scores from this model as a global measure of how positively principals view their evaluations. For interpretability, we standardize these scores to have a mean of 0 and a standard deviation of 1.

Predicting Evaluation Attitudes

Individual and School Characteristics

Table 3 examines how individual characteristics and local context relate to principals' evaluation attitudes. Column 1 shows individual characteristics only. Column 2 adds contextual characteristics, and column 3 adds district fixed effects, dropping district size and school locale type, which vary minimally within districts. Results show that female and Black principals perceive their evaluations more positively, though sorting appears to account for these

[11] To follow this procedure, we first conducted EFA on the first half of a split random sample, followed by confirmatory factor analysis (CFA) on the second half of the sample, then CFA on the full sample.

[12] As shown in Table 1, the seven items appeared in two different question blocks. This artifact of survey construction could have influenced the measurement structure. For this reason, we also explored a two-factor solution that grouped the feedback items separately. Fit statistics for this solution were the same as for the one-factor solution, a methodological phenomenon documented in the SEM literature (MacCallum, Wegener, Uchino, & Fabrigar, 1993). For reasons of parsimony and theoretical interrelatedness of the items, we opted to use the single-factor solution in our models. Nevertheless, in supplemental models using the two-factor solutions, we found parallel patterns across models; in four cases of differences, patterns were driven by process items and not feedback items. These results suggest that future work should investigate the potential multidimensional nature of principals' attitudes on evaluation.

relationships, as they become statistically insignificant once contextual factors are controlled. Results also show that, across models, first-year principals have the most positive views of evaluation, and the most experienced principals have the least positive views. We also find consistent evidence that principals working in high schools and "other" schools (i.e., those serving grade ranges outside of the traditional levels, such as alternative schools) perceive evaluation significantly more negatively than principals who work in elementary and middle schools.[13] Principals in larger districts appear also to have more negative views (β = -0.22, p = 0.11). More generally, the increase in R^2 when district fixed effects are added suggests that the school district context is an important factor in principals' attitudes about evaluation.

Principal Performance Ratings and Observation Frequency

Table 4 reports the results of models that include principal performance metrics. All models include district fixed effects; school-level covariates also are included but omitted from the table for parsimony.[14]

Column 1 in Table 4 indicates that lagged LOE scores are not significantly related to principals' evaluation attitudes. We cannot reject the null hypothesis that principals at all LOE levels have the same attitudes, conditional on other factors. In other words, the final evaluation rating a principal received last year does not predict attitudes about evaluation.

In contrast, lagged practice ratings, which contribute to LOE, positively and significantly predict principal attitudes, as shown in column 2. Each additional point is associated with an increase in the attitudinal measure of 0.16 SD. TVAAS scores, however, are uncorrelated with

[13] Given that high school principals are majority male and elementary school principals are overwhelmingly female (Taie & Goldring, 2019), we tested whether interactions between gender and school level would show differences from main effects. In a supplemental analysis with district fixed effects, no interactions were statistically significant at conventional levels.

[14] Coefficients for covariates were similar to those shown in Table 3.

attitudes (column 3). The positive relationship between practice ratings and attitudes maintains even controlling for TVAAS scores (column 4).[15] In other words, principals' attitudes appear more responsive to the ratings of their practices assigned by their supervisors than to the value-added component of their evaluation or the overall evaluation rating.

Next, we examined whether this score mattered more than how frequently principals were observed by their evaluators for purposes of assigning a practice rating. Column 1 in Table 5 indicates that principals' self-reported number of observations have a positive and significant relationship with attitudes. A test of equality (not shown here) confirmed that the three coefficients are statistically different from one another ($F = 58.73$, $p < 0.001$ comparing one observation and more than 2 observations; $F = 11.61$, $p < 0.001$ comparing two and more than two observations). In comparison, the objective number of observations a principal received, as documented in the administrative data and used to capture observation frequency in column 2, showed a directionally similar but weaker relationship with attitudes; here the reference category is principals who were recorded as only having received one rating or observation.[16] [17] Perceptions of frequency appear more predictive of attitudes than the number of recorded observations.

[15] One might expect principal attitudes are a function of what a principal scored in the current year. In a supplemental analysis, we used current-year measures of performance and found that current LOE or TVAAS were only significantly associated with attitudes at the higher (LOE) or lower (TVAAS) ranges. The direction of coefficients differed; principals with higher LOE scores, and principals scoring 3 or 4 on their current-year TVAAS, perceived evaluation more positively than those who scored 1 (LOE) or 1, 2, or 5 (TVAAS). In contrast, current observation score has a stronger association in magnitude and significance than lagged observation score, and remains significant in model 4. Thus, the lagged observation score is a more conservative estimate of the performance-attitude relationship than using the current score.

[16] A test of equality confirmed that the two and three or more categories in column 2 were not statistically different ($F = 0.32$, $p < 0.57$).

[17] We attempted to estimated the model in column 2 of Table 5 including cases where principals had zero observations in the administrative data, but only 22 principals fell into this category across years, giving us little power to differentiate this group. Given the small sample and to maintain comparability with column 4, which conditions on observation score (and thus by construction requires at least one observation), we leave these principals as missing.

In the remainder of the table, we examined the relative importance of frequency of observation and performance ratings in shaping attitudes. Column 3 in Table 5 shows that both principal-reported frequency of observation and practice rating each have a positive, significant relationship with attitudes. In column 4, the pattern is similar, and in fact one of the frequency categories is statistically significant at the 0.10 level when the practice rating is included. These findings suggest that both evaluation outcomes and at least one aspect of process correlate with attitudes, irrespective of one another. Accounting instead for LOE (columns 5 and 6) results in a pattern whereby frequency of observation is significant but LOE is not, as does substituting LOE with TVAAS (not shown).[18]

Evaluator Characteristics

Table 6 explores evaluator characteristics. All models include the number of years the evaluator has with conducting observations for practice ratings in the TEAM system. Across models, we generally find a nonlinear pattern, with principals expressing less positive attitudes with raters with 1–2 years of experience with the system than with brand-new or more veteran raters.

Columns 2 through 6 add other characteristics of evaluators. We add them one at a time to avoid potential multicollinearity. In column 2, we find some evidence that principals view the evaluation more favorably when they have been paired longer with the same evaluator, even controlling for how long the evaluator has been active. Each additional year is associated with an increase of 0.06 SD. In column 3, we include the number of years the principal and the evaluator have worked in the same district since 2002 (the first year of the job history data), a proxy for

[18] In supplemental models, we find that using current-year performance scores gives mostly similar results. Of note is that the magnitude and significance of current-year observation score is greater than the lagged score (models 3 and 4), and in a saturated model with current-year LOE, frequency of observations from administrative data loses significance (model 6).

how long they have known one another. Each additional year is associated with an increase of 0.02 SD.

Column 4 includes an indicator for whether the principal experienced a change in their evaluator in a given year. The point estimate is negative but not statistically significant at conventional levels. Similarly, we find no significant association with having multiple different raters in the same year.

The final column considers whether the principals view evaluation more positively when it is conducted by the highest district leader, the superintendent, as compared to other leaders (e.g., an assistant superintendent, a supervisor). The coefficient is positive but again not statistically significant at conventional levels; this was true even in supplemental analyses separating districts above and below the median district size. Data show that superintendents provide ratings much more often in rural districts (50% of the time) than in urban (23%) or suburban (29%) districts, so in supplemental models we tested for heterogeneity in this correlation by locale type. That analysis (not shown) found suggestive evidence that having the superintendent as evaluator was associated with more positive attitudes in suburban districts.

Table 7 tests for associations with evaluator gender and race and the interaction of these characteristics with those of the principal. Columns 1 and 2 examine gender, and columns 3 and 4 examine race, with the even-numbered columns including the average lagged practice rating as a covariate. Descriptively, half of the evaluators are female and 7 percent are Black (see Table 2). Fifty-three percent of principal respondents matched the gender of their evaluator, while 88 percent matched by race.

We find no evidence of an interaction between principal and evaluator gender in columns 1 and 2. The interaction term is not statistically significant at conventional levels, and the

patterns in the coefficients are inconsistent with a hypothesis that principals are more positive when evaluated by someone of the same gender.[19] The interaction terms in the models that focus on race similarly are not statistically significant at conventional levels and produce patterns that are inconsistent with a "race matching" effect on principal evaluation attitudes. Black principals appear to be more satisfied when rated by a white evaluator than by a Black evaluator. A supplemental test of equality comparing whether a Black principal with a Black observer has equivalent attitudes to a Black principal with a white observer showed that the two were significantly different from one another ($F = 4.39$, $p = 0.04$).

In further explorations of the relationship between demographic matching and principal attitudes, we used an evaluator fixed effects strategy to isolate variation in principal attitudes to be among principals with the same evaluator.[20] This approach holds all time-invariant characteristics of evaluators (e.g., communication style, leadership capacity) constant so may help better identify the gender or race interaction. Obviously, evaluator gender and race are fixed characteristics and so are accounted for by the evaluator fixed effect; however, we can still estimate the interaction between evaluator gender or race and principal gender or race. Table A1 in the appendix shows the results. As in Table 7, no clear pattern emerges for gender. In the race models, however, the interaction remained negative, doubled in magnitude, and became statistically significant (model 6). This finding corroborates the finding in Table 7 that Black principals perceive their evaluations less positively when their rater is Black. This pattern suggests a fruitful avenue for further investigation.[21]

---

[19] Predicted attitudes from this model would suggest that male principals evaluated by male raters are the least positive, while women evaluated by female raters have similar attitudes to principals in the off-diagonals.

[20] We fix the evaluator who conducted the summative end-of-year evaluation.

[21] In further exploratory analyses, we tested whether attitudes and actual outcomes (that is, practice ratings) were lower for Black principals with Black evaluators. In a model with practice ratings as the dependent variable, we found that, compared to Black principals with white evaluators, Black principals with Black evaluators receive significantly lower scores ($\beta = -0.22$, $p < 0.05$). Although this finding deserves further unpacking, it does suggest

In a final analysis, we re-estimated the main models from Tables 6 and 7 with controls for both the number of principal-reported observations and lagged practice ratings. The results, shown in Appendix Table A2, show that the associations between the evaluator characteristics and principal attitudes are not substantively affected by accounting for these other factors. It also demonstrates the relative strengths of the various relationships we test. Among these, the number of times the principal is evaluated appears to have the strongest association with attitudes; principals who recall being rated more than two times have attitudes that are half a standard deviation more positive than principals rated just once. This difference is approximately the same as the difference between principals receiving a practice rating of 1 (the lowest possible rating) and those receiving a 5 (the highest possible rating). For comparison, differences this large would require a principal to be paired for 10 years with their evaluator (which is outside the range in these data) or to work in the same district for 16 years.

Discussion and Conclusions

This study identifies how principals' views of their evaluations vary with their own characteristics and the characteristics of their local context, as well as by measures of how they experience the evaluation. We extend prior research on principals' attitudes (e.g., Davis & Hensley, 1999; Kimball, Milanowski, & McKinney, 2009; Sun & Youngs, 2009) by examining them statewide in the context of a principal evaluation system reformed in the post-RTTT era to standardize practice ratings and link them to student achievement measures. We apply an organizational justice framework, which focuses us in particular on the outcomes of evaluation (distributive justice) and its process (procedural justice). We find descriptively that women and

---

that the relationship between Black principals' attitudes about evaluation when rated by Black raters parallels the ratings they receive.

Black principals perceive their evaluations more positively, though this pattern may be more about where such principals work, as these relationships weaken when local context is accounted for. We also find that novice principals have more positive attitudes, while high school principals are less positive. Our analysis of evaluation outcomes shows some nuance. Overall evaluation ratings are uncorrelated with attitudes, as are the school value-added scores that comprise 35% of the final rating. In contrast, practice ratings assigned by supervisors are important for how principals view evaluation. In particular, principals who get more positive feedback from their supervisors about their practices view evaluation more favorably, suggesting that how they understand evaluation is a function more of the individualized, job-specific feedback they receive rather than of more indirect achievement-related metrics.

Beyond the ratings they receive, principals also feel more positively when they are rated more frequently, though this relationship is stronger for how often principals recall being evaluated than when we count the number of observations in the administrative data. We also find that some measures of principal–evaluator relationships predict attitudes, such as how many years the two have been paired for evaluation and how long they have worked together in the same district. These findings are important because principals' attitudes about evaluation likely affect the potential for principals to learn from their evaluations and for it to inform their practices.

We interpret these two predictors of principal attitudes—principals' observation or practice scores and how frequently they are evaluated—as consistent with predictions of distributive and procedural justice, respectively. When considered separately, both distributive and procedural justice measures were significant predictors of principals' attitudes. The finding that practice ratings assigned by supervisors were significantly associated with school leaders

viewing evaluation more positively provides support for the importance of distributive justice in situations where people assess the fairness of a specific outcome they receive (Folger & Konovsky, 1989). More significantly, however, the procedural component of justice predicted attitudes more strongly than did the distributive component. The primacy of procedural justice in predicting employee attitudes is consistent with research on organizational justice in settings outside of education on a range of attitudes, including perception of organizational authorities, their employing institution, and trust in their supervisor (Folger & Konovsky, 1989; Moorman, 1991).

This finding highlights an implication of our results for principal evaluation—namely, that how evaluations are conducted, rather than simply what scores principals receive—are key to how principals view the evaluation process. The importance of implementation underscores the value of efforts documented in recent studies of districts' efforts to ensure consistency of rating practices, the evidence they consider, and how they build into broader systems of principal support (e.g., Anderson & Turnbull, 2016; Kimball et al., 2015). Greater standardization of evaluation processes is likely to lead principals to see evaluation results and feedback as fairer and more accurate, perceptions that promote engaging with feedback rather than dismissing it, rationalizing it, or treating is superficially (Goldring, Mavrogordato, & Haynes, 2015).

For districts, our results also point to the value principals place on frequency of performance feedback and having a consistent supervisor who provides ratings over time. Consistency likely matters for building trust and for raters to observe the nuances in a principal's practice that make their evaluation feedback more useful. Making the rater consistent and the feedback more frequent brings the performance feedback process closer to an ongoing conversation with the supervisor that promotes principal improvement (Anderson & Turnbull,

2016). This shift is consistent with the aims of an emerging movement to reform the role of principal supervisors in many urban districts, in particular, from one that administers principal evaluation with an orientation toward operations and compliance to one that engages evaluation in the context of ongoing coaching and support for instructional leadership (Goldring et al., 2018; Honig & Rainey, 2019; Rubin et al., 2021). At minimum, our findings suggest that districts should provide principals with the minimum number of ratings and feedback conversations required by the state system, as principals viewed the system less favorably where the district was not in compliance. Ensuring frequent feedback should not be a mere "compliance thing" (DeMatthews et al., 2020, p. 12), however, as feedback is likely to be more useful if understood as part of an overall system for supporting principal practice (Honig, 2012). To this point, frequency and rater consistency may be especially important for principals receiving low practice ratings, who tend to have less positive views of evaluation.

Also, the relatively large differences by school level suggest that school districts might take steps to ensure that evaluation is meaningful for high school principals and that the feedback it provides is relevant to their work leading a complex organization. The importance of making evaluation meaningful and useful applies to more experienced principals as well, as leaders' views of the system appear to be somewhat lower for veteran principals. This pattern may reflect diminishing perceived utility of evaluation feedback to principals with many years learning how to enact the role. Addressing the less positive views of high school and more established principals may require more intentional communication about the value of evaluation or steps to reform evaluation to ensure its relevance to job performance and improvement for those leaders. To inform such efforts, future qualitative inquiry into the evaluation experiences of principals in different job contexts—not only by school type and job experience but by school size, district

size, and other key contextual factors that influence principals' work—would be useful in unpacking the reasons for the patterns we observe.

This last point underscores a limitation of our analysis, which is that it is based on survey data. Surveys, while providing the advantage of being able to explore attitudes in a representative set of principals across an entire state, provide necessarily coarse measures of attitudes that may be biased by recall challenges and respondents' interpretations of questions. For example, the relationship between attitudes and practice ratings (but not other evaluation outcomes) may be an artifact of principals interpreting that questions about their evaluations were specifically about the rating portion. Because the data are from a single state, we also face the limitation of external validity. Tennessee was an early adopter of multiple-measure principal evaluation, and the data we used were gathered after principals had several years of experience with the system. We do not know whether our results would generalize to states just adopting a multiple-measure principal evaluation system, or how much our results are driven by the specific characteristics and requirements of Tennessee's system. We recommend extending this initial look at principals' attitudes about evaluation to other state (or district) contexts.

Future work could extend our analysis of race and gender interactions between principals and their evaluators. We do not find evidence of race or gender "matching" relationships on principal attitudes, and in fact find some evidence that Black principals rated by Black evaluators perceive their evaluation experience more negatively than do Black principals rated by white evaluators. We do not know whether this finding reflects something about principal and evaluator relationships by race, about the stringency of performance standards evaluators set for own-race versus other-race employees, about racial differences in how evaluation processes are implemented, or about something else. Further inquiry may help illuminate the mechanism. One

potential avenue for exploration comes from Burt's (1987) work on structural equivalence, which suggests the possibility that evaluators perceive same-race employees as more similar to themselves in status, creating feelings of threat and a competitive dynamic.

Future work could also explore other aspects of organizational justice as applied to principal evaluation. For example, *interpersonal justice*—or dimensions of treatment among actors in an organization such as displaying dignity, respectfulness, and honesty (Bies & Moag, 1986)—may be particularly relevant. Such measures of interpersonal dynamics are likely more relevant to principals' attitudes than the relationship proxies we had available, as the ability of an evaluator to create a fair climate and to communicate honestly with the principal is of central importance to principals in their experience of the evaluation process (Hvidston, Range, & McKim, 2015). Finally, future research could extend our analysis to examine how principal attitudes about evaluation connect to other important outcomes, such as changes to principal practice, job performance, or other work attitudes, such as job satisfaction and organizational commitment. It could also investigate whether prior attitudes on evaluation shape the kinds of evaluation practices principals use on others, either on teachers or on other principals as they advance in their careers (Castilla & Ranganathan, 2020).

References

Anderson, L. M., & Turnbull, B. J. (2016). *Evaluating and supporting principals*. Policy Studies Associates. Retrieved from https://files-eric-ed-gov.proxy.library.vanderbilt.edu/fulltext/ED570471.pdf.

Archer, J., Kerr, K., & Pianta, R.C. (2014). Why measure effective teaching? In T. Kane, K. Kerr, & R.C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project*, 1-6. John Wiley & Sons.

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., Mccaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2014). Improving observational score quality: Challenges in evaluator thinking. In T. Kane, K. Kerr, & R.C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project*, 50-97. John Wiley & Sons.

Bell, B. S., & Ford, J. K. (2007). Reactions to skill assessment: The forgotten factor in explaining motivation to learn. *Human Resource Development Quarterly*, *18*(1), 33-62.

Bell, B. S., Tannenbaum, S. I., Ford, J. K., Noe, R. A., & Kraiger, K. (2017). *100 years of training and development research: What we know and where we should go*. Retrieved from https://digitalcommons.ilr.cornell.edu/articles/1289

Bies, R. J., & Moag, J. S. (1986). Interpersonal justice: Communication criteria of fairness. *Research on Negotiation in Organizations,* 1(1), 43–55.

Brown-Sims, M. (2010). *Evaluating school principals: Tips and tools.* National Comprehensive Center for Teacher Quality. ERIC Number: ED543770.

Burt, R. S. (1987). Social contagion and innovation: Cohesion versus structural equivalence. *American Journal of Sociology*, *92*(6), 1287-1335.

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

Castilla, E. J., & Ranganathan, A. (2020). The Production of Merit: How Managers Understand and Apply Merit in the Workplace. *Organization Science* 31(4), 909-935.

Catano, N., & Stronge, J. H. (2007). What do we expect of school principals? Congruence between principal evaluation and performance standards. *International Journal of Leadership in Education, 10*(4), 379-399. doi:10.1080/13603120701381782

Condon, C., & Clifford, M. (2012). *Measuring principal performance: How rigorous are commonly used principal performance assessment instruments?* Naperville, IL: Learning Point Associates. Retrieved from https://www.air.org/sites/default/files/downloads/report/Measuring_Principal_Performance_0.pdf

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology,* 86(3), 425.

Cropanzano, R. S., & Ambrose, M. L. (2015). Organizational justice: Where we have been and where we are going. In R. S. Cropanzano & M. L. Ambrose (Eds.), *The Oxford handbook of justice.* Oxford University Press.

Davis, S. H., & Hensley, P. A. (1999). The politics of principal evaluation. Journal of Personnel Evaluation in Education, 13(4), 383-403.

Davis, S., Kearney, K., Sanders, N., Thomas, C., & Leon, R. (2011). *The policies and practices of principal evaluation: Executive summary.* San Francisco, CA: WestEd.

DeNisi, A. S., & Murphy, K. R. (2017). Performance appraisal and performance management: 100 years of progress?. *Journal of Applied Psychology*, *102*(3), 421.

Dipboye, R. L. (1985). Some neglected variables in research on discrimination in appraisals. *Academy of Management Review*, *10*(1), 116-127.

Doherty, J. F. (2009). *Perceptions of teachers and administrators in a Massachusetts suburban school district regarding the implementation of a standards-based teacher evaluation system*. Unpublished doctoral dissertation. Seton Hall University.

Drake, S., Auletto, A., & Cowen, J. M. (2019). Grading teachers: Race and gender differences in low evaluation ratings and teacher employment outcomes. *American Educational Research Journal*, *56*(5), 1800-1833.

Ferguson, E., & Cox, T. (1993). Exploratory factor analysis: A users' guide. *International journal of selection and assessment*, *1*(2), 84-94.

France, R.G. & Thompson, E. (2015). Suburban district leadership does matter. *Journal for Leadership and Instruction,* Spring*,* 5-8.

Ford, D. K., Truxillo, D. M., & Bauer, T. N. (2009). Rejected but still there: Shifting the focus in applicant reactions to the promotional context. *International Journal of Selection and Assessment*, *17*(4), 402-416.

Folger, R., & Konovsky, M. A. (1989). Effects of procedural and distributive justice on reactions to pay raises. *Academy of Management Journal, 32, 115–130.*

Folger, R., Konovsky, M. A., & Cropanzano, R. (1992). A due process metaphor for performance appraisal. Research in organizational behavior, 14, 129–129.

Fuller, E. J., Hollingworth, L., & Liu, J. (2015). Evaluating state principal evaluation plans across the United States. *Journal of Research on Leadership Education*, *10*(3), 164-192.

Fuller, E. J., & Hollingworth, L. (2014). A bridge too far? Challenges in evaluating principal effectiveness. *Educational Administration Quarterly*, *50*, 466–499.

Goldring, E., Cravens, X. C., Murphy, J., Porter, A. C., Elliott, S. N., & Carson, B. (2009). The evaluation of principals: what and how do states and urban districts assess leadership? *The Elementary School Journal*, *110*(1), 19-39.

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, *44*(2), 96-104.

Goldring, E.B., Grissom, J.A., Rubin, M., Rogers, L.K., Neel, M., & Clark, M. (2018). *A new role emerges for principal supervisors: Evidence from six districts in the Principal Supervisor Initiative.* New York: The Wallace Foundation.

Goldring, E. B., Mavrogordato, M., & Haynes, K. T. (2015). Multisource principal evaluation data: Principals' orientations and reactions to teacher feedback regarding their leadership effectiveness. *Educational Administration Quarterly, 51*(4), 572-599.

Honig, M. I. (2012). District central office leadership as teaching: How central office administrators support principals' development as instructional leaders. *Educational Administration Quarterly*, *48*(4), 733-774.

Honig, M. I., & Rainey, L. R. (2019). Supporting principal supervisors: what really matters?. *Journal of Educational Administration*, 57(5), 445-462.

Hvidston, D. J., Range, B. G., & McKim, C. A. (2015). Principals' perceptions regarding their supervision and evaluation. The American Association of School Administrators (AASA) *Journal of Scholarship and Practice,* 12(2), 20-33.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-136. doi:10.1086/522974

Jost, J. T., & Kay, A. C. (2010). Social justice: History, theory, and research. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), Handbook of social psychology (Vol. 2, 5th ed., pp. 1122–1165). Hoboken, NJ: John Wiley & Sons Inc.

Kimball, S. M., Arrigoni, J., Clifford, M., Yoder, M., & Milanowski, A. (2015*). District leadership for effective principal evaluation and support*. U.S. Department of Education, Teacher Incentive Fund. Retrieved from: https://files.eric.ed.gov/fulltext/ED566525.pdf.

Kimball, S. M., Milanowski, A., & McKinney, S. A. (2009). Assessing the promise of standards-based performance evaluation for principals: Results from a randomized trial. *Leadership and Policy in Schools, 8*(3), 233-263.

Kimball, S. M., & Pautsch, C. A. (2008). *Principal evaluation and support in two school districts using new leadership standards: A cross-site comparison*. Madison: University of Wisconsin-Madison.

Lashway, L. (2003). *Improving principal evaluation* (ERIC Digest. Access ERIC: Full Text [071 Information Analyses—ERIC IAPs No.EDO-EA-03-09]). Eugene, OR: ERIC Clearinghouse on Educational Management.

Lavigne, A. L. (2018). Examining individual-and school-level predictors of principal adaptation to teacher evaluation reform in the United States: A two-year perspective. *Educational Management Administration & Leadership*, 1741143218807491.

Leventhal, G. G., Karuza, J., Jr., & Fry, W. R. (1980). Beyond fairness: A theory of allocation preferences. In G. Mikula (Ed.), Justice and social interaction (pp. 167–218). New York, NY: Springer–Verlag.

Levy, Paul E., Caitlin M. Cavanaugh, Noelle B. Frantz, and Lauren A. Borden. (2015). "The role of due process in performance appraisal: A 20-year retrospective." In R. S. Cropanzano & M. L. Ambrose (Eds.), *The Oxford handbook of justice.* Oxford University Press.

Lind, E.A. (2001) Fairness Heuristic Theory: Justice judgments as pivotal cognitions in organizational settings. In: Greenberg, J. and Cropanzano, R. (eds), *Advances in organizational justice*. Stanford, CA: Stanford University Press, pp. 56–88.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological bulletin*, *114*(1), 185.

McGuinn, P. (2012). Stimulating reform: Race to the Top, competitive grants and the Obama education agenda. *Educational Policy*, *26*(1), 136-159.

Moorman, R. H. (1991). Relationship between organizational justice and organizational citizenship behaviors: Do fairness perceptions influence employee citizenship? *Journal of Applied Psychology,* 76, 845–855.

Mueller, C. W., Finley, A., Iverson, R. D., & Price, J. L. (1999). The effects of group racial composition on job satisfaction, organizational commitment, and career commitment: The case of teachers. *Work and Occupations*, *26*(2), 187-219.

Northcraft, G. B., Schmidt, A. M., & Ashford, S. J. (2011). Feedback and the rationing of time and effort among competing tasks. *Journal of Applied Psychology*, *96*(5), 1076.

Parylo, O., Zepeda, S. J., & Bengtson, E. (2012). Principals' experiences of being evaluated: A phenomenological study. *Educational Assessment, Evaluation and Accountability*, *24*(3), 215-238.

Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). Examination of race and sex effects on performance ratings. *Journal of Applied Psychology,* 74, 770–780.

Ravenell, A. (2019). Trends in Teacher Perceptions of Educator Evaluation. Tennessee Education Research Alliance. Accessed online 11 December https://peabody.vanderbilt.edu/TERA/files/Survey_Snapshot_Educator_Evaluation_FINAL.pdf

Reeves, D. B. (2004). *Assessing educational leaders: Evaluating performance for improve individual and organizational results* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Renzulli, L. A., Parrott, H. M., & Beattie, I. R. (2011). Racial mismatch and school type: Teacher satisfaction and retention in charter and traditional public schools. *Sociology of Education*, *84*(1), 23-48.

Rockoff, J.E., Staiger, D.O., Kane, T.J., & Taylor, E.S. (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic* Review 102(7):3184–3213. doi:10.1257/aer.102.7.3184.

Rosen, B., & Jerdee, T. H. (1976). The nature of job-related age stereotypes. *Journal of Applied Psychology,* 61(2), 180–183.

Rothman, R. (2017). *Improving School Leadership Under ESSA: Evidence-Based Options for States & Districts.* NISLWhitepaper. Criterion Education, LLC.

Rubin, M., Goldring, E., Neel, M.A, Rogers, L.K., & Grissom, J.A. (2021). Changing principal supervision to develop principals' instructional leadership capacity. In P. Youngs, J. Kim, & M. Mavrogordato (eds.), *Exploring Principal Development and Teacher Outcomes: How Principals Can Strengthen Instruction, Teacher Retention, and Student Achievement*. New York: Routledge, 35–47.

Ruzek, E. A., Hafen, C. A., Hamre, B. K., & Pianta, R. C. (2014). Combining classroom observations and value added for the evaluation and professional development of teachers. In T. Kane, K. Kerr, & R.C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project*, 203-233. John Wiley & Sons.

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*, *11*(3), 340-359.

Stronge, J. H. (1991). The dynamics of effective performance evaluation systems in education: Conceptual, human relations, and technical domains. *Journal of Personnel Evaluation in Education*, *5*(1), 77-83.

Sun, M., & Youngs, P. (2009). How does district principal evaluation affect learning centered principal leadership? Evidence from Michigan school districts. Leadership and Policy in Schools, 8(4), 411-445.

Superville, D. S. (2014, May 20). States forge ahead on principal evaluation. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2014/05/21/32principals_ep.h33.html

Thomas, D. W., Holdaway, E. A., & Ward, K. L. (2000). Policies and practices involved in the evaluation of school principals. *Journal of personnel evaluation in education*, *14*(3), 215-240.

Warr, P., Allan, C., & Birdi, K. (1999). Predicting three levels of training outcome. *Journal of Occupational and Organizational Psychology, 72*, 351-375.

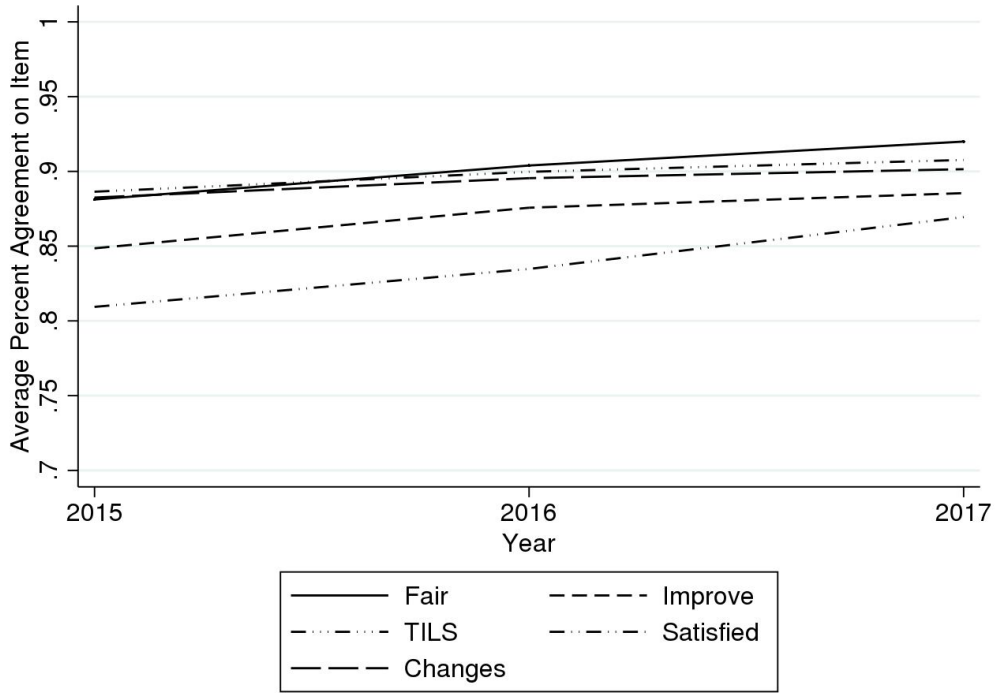
Wexley, K. N., Singh, J. P., & Yukl, G. (1973). Subordinate personality as a moderator of the effects of participation in three types of appraisal interview. *Journal of Applied Psychology,* 58, 54–59.

Wilson, B., & Natriello, G. (1989). *Teacher Evaluation and School Climate*. Washington, DC: Office of Educational Research and Improvement. Retrieved from http://www.eric.ed.gov/PDFS/ED374556.pdf.

Xu, S., & Sinclair, R. L. (2002). *Improving teacher evaluation for increasing student learning.* Paper presented at the annual meeting of the American Education Research Association, New Orleans, LA.

Figure 1. Change in Attitudes Over Time

Panel A. Average percent responding agree and strongly agree



Panel B. Average percent responding mostly true or true

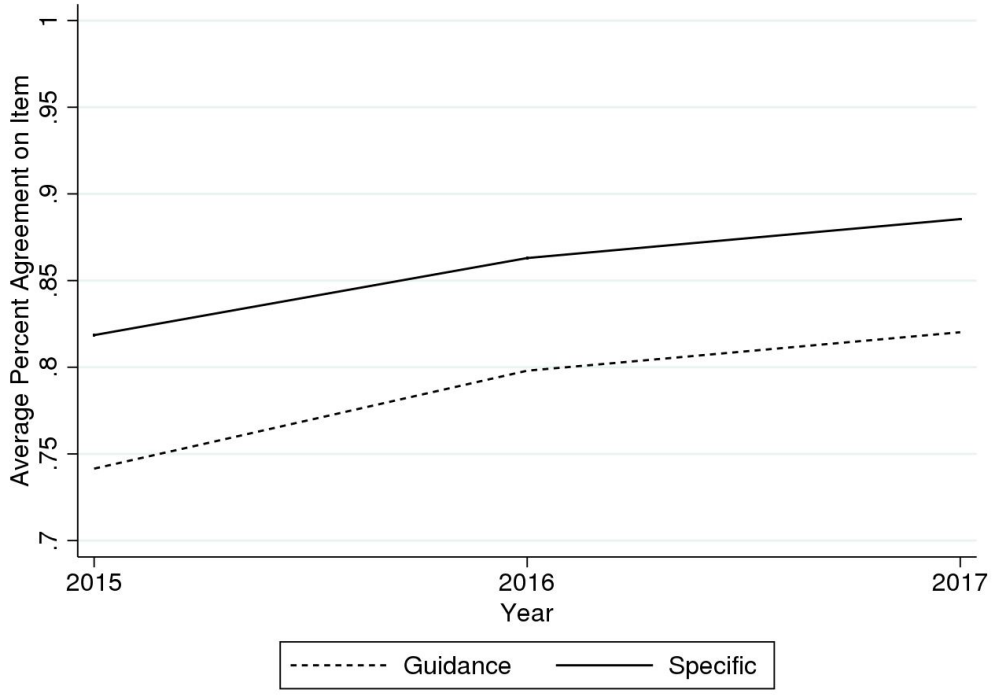


Table 1. Survey Items and Descriptives

	N	Mean (SD)	Min	Max	Factor Loading
How true is each of the following statements about the feedback you have received? The feedback I received so far this year: ***(on scale of 1-4, Not at all True, Somewhat True, Mostly True, True; N/A category has been dropped)***					
SPECIFIC: The feedback I received so far this year identified specific areas of my practice that could be improved.	2,156	3.49 (0.80)	1	4	0.63
GUIDANCE: The feedback I received so far this year included guidance on how to make improvements in my practice.	2,153	3.29 (0.94)	1	4	0.65
How strongly do you agree or disagree with each of the following statements about the administrator evaluation process during this school year? ***(on scale of 1-4, Strongly disagree, Disagree, Agree,* and *Strongly Agree)***					
FAIR: The processes used to conduct my administrator evaluation are fair to me.	2,260	3.16 (0.66)	1	4	0.83
IMPROVE: The administrator evaluation process helps me improve as a professional.	2,253	3.12 (0.68)	1	4	0.87
TILS: The Tennessee Instructional Leadership Standards (TILS) and corresponding rubric clearly define what is expected of me as an administrator.	2,256	3.15 (0.64)	1	4	0.72
CHANGES: I have made changes in my leadership practice as a result of the evaluation.	2,260	3.16 (0.65)	1	4	0.81
SATISFIED: Overall, I am satisfied with Tennessee's administrator evaluation process.	2,246	3.00 (0.70)	1	4	0.83

Table 2. Sample Descriptives

	N	Mean (SD)		Min	Max	State Mean
Female	2,286	0.59	(0.49)	0	1	0.56
Black	2,286	0.09	(0.29)	0	1	0.15
PhD/EdS	2,281	0.49	(0.50)	0	1	0.49
1 year prior experience	2,286	0.12	(0.33)	0	1	0.12
2 years prior experience	2,286	0.11	(0.32)	0	1	0.11
3-4 years prior experience	2,286	0.19	(0.39)	0	1	0.19
5-7 years prior experience	2,286	0.19	(0.39)	0	1	0.18
8+ years prior experience	2,286	0.27	(0.44)	0	1	0.28
Enrollment, x100	2,278	6.26	(3.84)	0.12	28.10	6.52
Fraction free and reduced price lunch	2,272	0.60	(0.22)	0	1	0.60
Fraction Black students	2,272	0.15	(0.21)	0	1	0.19
Fraction Hispanic students	2,272	0.08	(0.10)	0	0.71	0.08
Achievement index	2,109	0.16	(0.80)	-5.22	3.49	0.11
Middle school	2,261	0.18	(0.38)	0	1	0.18
High school	2,261	0.16	(0.37)	0	1	0.19
Other school	2,261	0.05	(0.21)	0	1	0.04
Urban	2,261	0.20	(0.40)	0	1	0.26
Town	2,261	0.22	(0.42)	0	1	0.19
Rural	2,261	0.44	(0.50)	0	1	0.40
Suburban	2,261	0.14	(0.35)	0	1	0.15
District size, x100,000	2,286	0.08	(0.15)	0	1.43	0.07
Level of Effectiveness, lagged	2,051	3.85	(1.04)	1	5	3.85
Score of 1 or 2	2,051	0.14	(0.34)	0	1	0.13
Score of 3	2,051	0.21	(0.41)	0	1	0.21
Score of 4	2,051	0.33	(0.47)	0	1	0.33
Score of 5	2,051	0.33	(0.47)	0	1	0.32
Average practice rating, lagged	2,145	3.91	(0.53)	2	5	3.90
TVAAS, lagged	1,725	3.14	(1.66)	1	5	3.14
Number of observations (principal-reported)	2,267	1.94	(0.87)	0	3	1.94
Number of observations (administrative data)	2,264	1.94	(0.29)	1	3	1.92
Evaluator's experience with observing, years	2,286	1.84	(1.28)	0	4	1.83
0 years experience	2,286	0.21	(0.41)	0	1	0.20
1-2 years experience	2,286	0.47	(0.50)	0	1	0.49
3 or more years experience	2,286	0.32	(0.47)	0	1	0.31
Evaluator experience in current position, years	2,104	5.41	(4.80)	0	15	5.48
Years paired with evaluator	2,179	2.06	(1.18)	1	5	1.98
Years working in same district as evaluator	2,286	9.99	(5.34)	0	16	9.91
Change in evaluator	2,067	0.41	(0.49)	0	1	0.44
Multiple raters	2,005	0.15	(0.35)	0	1	0.15
Evaluator's role						
Superintendent	1,985	0.44	(0.50)	0	1	0.38
Assistant superintendent	1,985	0.09	(0.29)	0	1	0.12
Supervisors	1,985	0.32	(0.47)	0	1	0.29
Other	1,985	0.15	(0.36)	0	1	0.21
Evaluator is female	1,996	0.50	(0.50)	0	1	0.49
Evaluator is Black	1,994	0.07	(0.26)	0	1	0.11
Attitudes towards evaluation (global/overall)	2,286	0	(1.00)	-3.70	1.54	--

Table 3. Individual and School Characteristics as Predictors of Principal Evaluation Attitudes

	(1) Individual Characteristics	(2) School and District Characteristics	(3) District Fixed Effects
Female	0.12* (0.05)	0.08 (0.06)	0.10 (0.06)
Black	0.22+ (0.11)	0.14 (0.13)	0.11 (0.14)
PhD/EdS	-0.04 (0.06)	-0.04 (0.06)	-0.05 (0.06)
1 year experience	-0.20* (0.08)	-0.22** (0.08)	-0.18* (0.09)
2 years experience	-0.18* (0.08)	-0.24** (0.08)	-0.23** (0.08)
3-4 years experience	-0.19** (0.07)	-0.19** (0.07)	-0.20** (0.07)
5-7 years experience	-0.20* (0.08)	-0.21** (0.08)	-0.21* (0.08)
8+ years experience	-0.26** (0.08)	-0.27** (0.08)	-0.31** (0.08)
Enrollment, x100		0.01 (0.01)	0.01 (0.01)
Fraction free and reduced price lunch		0.15 (0.14)	0.07 (0.16)
Fraction Black students		0.36+ (0.19)	-0.20 (0.26)
Fraction Hispanic students		-0.06 (0.43)	-0.69+ (0.37)
Achievement index		0.08 (0.05)	-0.07 (0.05)
Middle school		0.02 (0.08)	-0.02 (0.08)
High school		-0.34** (0.11)	-0.29* (0.12)
Other school		-0.34* (0.14)	-0.38* (0.17)
District size, x100,000		-0.22 (0.14)	
Urban		0.00 (0.13)	
Town		-0.07 (0.12)	
Rural		-0.05 (0.11)	
Suburban		--	
Constant	0.00 (0.07)	-0.05 (0.17)	0.11 (0.17)
Observations	2281	2084	2084
R^2	0.02	0.04	0.18

Standard errors clustered at the district level.
Standard errors in parentheses
All models include year fixed effects.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 4. Principal Performance Measures and Evaluation Attitudes

	(1) Level of Effectiveness (LOE)	(2) Observation Score	(3) School-Level TVAAS	(4) Observation Score & School-Level TVAAS
Female	0.08	0.07	0.14^{+}	0.12
	(0.07)	(0.07)	(0.07)	(0.07)
Black	0.17	0.16	0.17	0.18
	(0.14)	(0.15)	(0.17)	(0.17)
PhD/EdS	-0.06	-0.07	-0.09	-0.09
	(0.07)	(0.07)	(0.07)	(0.07)
1 year prior experience	-0.14	-0.08	-0.08	-0.04
	(0.09)	(0.09)	(0.09)	(0.10)
2 years prior experience	-0.18*	-0.16^{+}	-0.15^{+}	-0.13
	(0.09)	(0.08)	(0.09)	(0.09)
3-4 years prior experience	-0.16^{+}	-0.15^{+}	-0.16^{+}	-0.14
	(0.08)	(0.08)	(0.09)	(0.10)
5-7 years prior experience	-0.21*	-0.20*	-0.17	-0.18^{+}
	(0.10)	(0.10)	(0.11)	(0.11)
8+ years prior experience	-0.29**	-0.29**	-0.31**	-0.31**
	(0.10)	(0.10)	(0.11)	(0.11)
LOE, lagged (=3)	-0.13			
	(0.09)			
LOE, lagged (=4)	-0.10			
	(0.10)			
LOE, lagged (=5)	0.02			
	(0.10)			
Average practice rating, lagged		0.16**		0.12^{+}
		(0.06)		(0.07)
TVAAS, lagged (=2)			0.01	0.01
			(0.10)	(0.10)
TVAAS, lagged (=3)			-0.06	-0.05
			(0.09)	(0.10)
TVAAS, lagged (=4)			-0.06	-0.06
			(0.10)	(0.11)
TVAAS, lagged (=5)			0.07	0.07
			(0.06)	(0.06)
Constant	0.18	-0.49	0.11	-0.36
	(0.18)	(0.30)	(0.21)	(0.35)
Observations	1876	1962	1566	1549
R^2	0.19	0.18	0.20	0.20

Standard errors clustered at the district level.
Standard errors in parentheses
All models include school level covariates and district and year fixed effects.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 5. Perceived and Actual Number of Times Observed and Principal Attitudes

	(1) Self-Reported Observations	(2) Observed Observations	(3) Self-Reported Observations & Observation Score	(4) Observed Observations & Observation Score	(5) Self-Reported Observations & LOE	(6) Observed Observations & LOE
Number of observations (Principal-reported=1)	0.47**		0.46**		0.44*	
	(0.15)		(0.16)		(0.17)	
Number of observations (Principal-reported=2)	0.80**		0.80**		0.77**	
	(0.15)		(0.16)		(0.17)	
Number of observations (Principal-reported>2)	1.02**		1.02**		1.00**	
	(0.16)		(0.17)		(0.19)	
Number of observations (administrative data=2)		0.28		0.33+		0.40*
		(0.17)		(0.19)		(0.20)
Number of observations (administrative data=3)		0.40		0.45		0.57*
		(0.25)		(0.28)		(0.26)
Average practice rating, lagged			0.15*	0.16*		
			(0.06)	(0.06)		
Level of Effectiveness, lagged (=3)					-0.11	-0.13
					(0.08)	(0.08)
Level of Effectiveness, lagged (=4)					-0.13	-0.11
					(0.09)	(0.10)
Level of Effectiveness, lagged (=5)					0.01	0.02
					(0.10)	(0.10)
Constant	-0.70**	-0.17	-1.28**	-0.83*	-0.61*	-0.22
	(0.22)	(0.21)	(0.33)	(0.32)	(0.25)	(0.23)
Observations	2069	2067	1950	1950	1865	1865
R^2	0.23	0.18	0.24	0.19	0.23	0.19

Standard errors clustered at the district level.
Standard errors in parentheses
All models include individual and school level covariates and district and year fixed effects.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 6. Evaluator Characteristics and Principal Evaluation Attitudes

	(1) Evaluator experience	(2) Years paired with evaluator	(3) Years worked with evaluator	(4) Change of evaluator	(5) Multiple raters	(6) Evaluator role
Evaluator experience with observing (0 years omitted)	--	--	--	--	--	--
1-2 years	-0.14	-0.19*	-0.14	-0.14	-0.10	-0.23**
	(0.09)	(0.09)	(0.09)	(0.10)	(0.10)	(0.09)
At least 3 years	-0.01	-0.10	-0.01	-0.01	0.03	-0.13
	(0.12)	(0.12)	(0.12)	(0.13)	(0.13)	(0.11)
Years principal paired with evaluator		0.06+				
		(0.03)				
Years principal has worked in same district with evaluator			0.02**			
			(0.01)			
Change of evaluator from previous year				-0.08		
				(0.07)		
Multiple (>1) evaluators within-year					0.03	
					(0.08)	
Evaluator is superintendent						0.13
						(0.11)
Evaluator experience in current position, years	0.00	0.00	-0.00	0.00	0.00	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.28	0.27	0.08	0.28	0.23	0.30
	(0.20)	(0.20)	(0.21)	(0.23)	(0.22)	(0.21)
Observations	1922	1922	1922	1798	1770	1815
R^2	0.19	0.19	0.19	0.20	0.19	0.19

Standard errors clustered at the district level
Standard errors in parentheses
All models include individual and school level covariates and district and year fixed effects.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 7. Gender and Race Match between Principal and Evaluator and Principal Evaluation Attitudes

	(1) Gender interaction	(2) Gender interaction with lag observation score	(3) Race interaction	(4) Race interaction with lag observation score
Female Principal	0.14	0.13		
	(0.09)	(0.10)		
Female Evaluator	0.11	0.12		
	(0.12)	(0.12)		
Female Principal*Female Evaluator	-0.12	-0.14		
	(0.13)	(0.13)		
Black Principal			0.25+	0.27+
			(0.14)	(0.14)
Black Evaluator			0.10	0.05
			(0.32)	(0.33)
Black Principal*Black Evaluator			-0.49	-0.44
			(0.32)	(0.32)
Average practice rating, lagged		0.13+		0.14+
		(0.07)		(0.07)
Constant	0.30	-0.20	0.35	-0.16
	(0.21)	(0.36)	(0.21)	(0.37)
Observations	1826	1717	1806	1697
R^2	0.19	0.19	0.20	0.20

Standard errors clustered at the district level.
Standard errors in parentheses
All models include individual and school level covariates, evaluator years of experience and years in position, and year and district fixed effects.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Appendix to *EAQ* article "Performance, Process, and Interpersonal Relationships" (Nelson, Grissom, & Cameron, 2021)

Table A1. Gender and Race Match between Principal and Evaluator and Principal Evaluation Attitudes, Evaluator Fixed Effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female Principal	0.04	0.07	-0.03	0.12				
	(0.07)	(0.10)	(0.10)	(0.10)				
Female Evaluator		--						
Female Principal*Female Evaluator		-0.07						
		(0.14)						
Black Principal					0.07	0.25^{+}	-0.77	0.24
					(0.15)	(0.15)	(0.58)	(0.15)
Black Evaluator						--		
Black Principal*Black Evaluator						-0.89**		
						(0.31)		
Constant	0.18	0.20	0.43	-0.25	0.19	0.22	0.98	0.09
	(0.26)	(0.27)	(0.38)	(0.45)	(0.26)	(0.27)	(1.37)	(0.30)
Observations	1922	1826	907	919	1913	1800	135	1665
R^2	0.30	0.30	0.30	0.31	0.30	0.30	0.45	0.29

Standard errors clustered at the district level.
Standard errors in parentheses
All models include individual and school level covariates, observer years of experience and years in position, and year fixed effects.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table A2. Joint Analysis of Performance, Process, and Interpersonal Relationships and Evaluation Attitudes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Average practice rating, lagged	0.12^{+}	0.13^{+}	0.14*	0.11	0.11	0.13^{+}	0.13^{+}
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
Number of observations	0.53**	0.49**	0.55**	0.46*	0.53**	0.53**	0.53**
(Principal-reported=1)	(0.18)	(0.17)	(0.19)	(0.19)	(0.18)	(0.18)	(0.18)
Number of observations	0.89**	0.85**	0.91**	0.82**	0.89**	0.88**	0.89**
(Principal-reported=2)	(0.17)	(0.17)	(0.18)	(0.18)	(0.17)	(0.17)	(0.17)
Number of observations	1.08**	1.05**	1.10**	1.03**	1.08**	1.08**	1.09**
(Principal-reported>2)	(0.19)	(0.19)	(0.20)	(0.19)	(0.19)	(0.19)	(0.19)
Years principal paired with evaluator	0.05						
	(0.03)						
Years principal has worked in same district with evaluator		0.03**					
		(0.01)					
Change of evaluator from previous year			-0.10				
			(0.07)				
Multiple (>1) evaluators within-year				0.01			
				(0.09)			
Evaluator is superintendent					0.14		
					(0.11)		
Evaluator and principals' genders match						-0.05	
						(0.06)	
Evaluator and principals' races match							-0.25^{+}
							(0.14)
Constant	-1.05**	-1.29**	-1.10*	-1.05*	-1.03*	-1.04**	-0.86*
	(0.37)	(0.37)	(0.43)	(0.41)	(0.40)	(0.37)	(0.37)
Observations	1790	1790	1684	1651	1688	1790	1790
R^2	0.25	0.26	0.26	0.25	0.25	0.25	0.25

Standard errors clustered at the district level.
Standard errors in parentheses
All models include individual and school level covariates, observer years of experience and years in position, and year fixed effects.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$