Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 24-1092

Measuring the Affective Language of Principals' Evaluation Feedback and Investigating Differences by Principal Gender and Race

Karin Gegenheimer
Research for Action

Ellen Goldring
Vanderbilt University

Over the past decade, reforms to principal evaluation systems have sought to incorporate formal feedback structures as a lever for principal improvement. However, we know little about the feedback that principals receive. Using statewide administrative data from Tennessee, including principals' written feedback from evaluators, we use sentiment analysis to uncover the affective language, or tone, of principals' feedback, and examine differences in affective language based on principal gender and race. We find that the affective language of refinement feedback (constructive feedback) largely resembles that of reinforcement feedback (affirmative feedback) and that female principals receive reinforcement feedback with less positive affective language relative to observably similar male principals. We also find some suggestive evidence that Black principals receive refinement feedback that is less positive in tone than the feedback to their white peers. We conclude with implications for policy and practice and suggestions for future work.

VERSION: November 2024

Suggested citation: Gegenheimer, Karin, and Ellen Goldring. (2024). Measuring the Affective Language of Principals' Evaluation Feedback and Investigating Differences by Principal Gender and Race. (EdWorkingPaper: 24 -1092). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/hb5n-me49

Measuring the Affective Language of Principals' Evaluation Feedback and Investigating Differences by Principal Gender and Race

***

Over the past decade, reforms to principal evaluation systems have sought to incorporate formal feedback structures as a lever for principal improvement. However, we know little about the feedback that principals receive. Using statewide administrative data from Tennessee, including principals' written feedback from evaluators, we use sentiment analysis to uncover the affective language, or tone, of principals' feedback, and examine differences in affective language based on principal gender and race. We find that the affective language of refinement feedback (constructive feedback) largely resembles that of reinforcement feedback (affirmative feedback) and that female principals receive reinforcement feedback with less positive affective language relative to observably similar male principals. We also find some suggestive evidence that Black principals receive refinement feedback that is less positive in tone than the feedback to their white peers. We conclude with implications for policy and practice and suggestions for future work.

***

Introduction

The past decade has seen a renewed focus on principal evaluation in education policy reforms, as almost all states have adopted new principal evaluation policies since 2009 (Donaldson et al., 2021; Donaldson et al., 2020). Many states and districts have restructured principal evaluation to connect more closely with principals' day-to-day responsibilities and to serve as a strategy to support ongoing principal and school improvement (Fuller, Hollingworth, & Liu, 2015). In particular, reforms to principal evaluation have focused on incorporating formal feedback structures based on walk-throughs and observations paired with post-observation performance feedback to support principals' professional growth (Donaldson et al., 2020). The majority of states now require observations and post-observation feedback conferences in their principal evaluation policies (Donaldson et al., 2020).

Despite the policy emphasis on principal evaluation feedback and developmental support, we know little about the feedback that principals receive from their formal evaluations. Much of

the recent research on principal evaluation systems has centered around identifying components of evaluation (Fuller et al., 2015), investigating the validity and reliability of evaluation tools (Grissom, Blissett, & Mitani, 2018), and understanding principals' experiences with these new evaluation systems (DeMatthews, Scheffer, & Kotok, 2020; Hvidston, McKim, & Holmes, 2018). Other research has focused on principals' orientations and reactions to multisource feedback, including self-appraisals and feedback from teachers (Goldring, Mavrogordato, & Haynes, 2015). However, there is very little research on the feedback that principals receive from supervisors, particularly in the context of formal, mandated statewide evaluation systems.

Performance feedback is a critical organizational resource that, when done well, can have substantial and far-reaching effects on individual performance and organizational meritocracy (Cannon & Witherspoon, 2005; Sprick et al., 2010). The theory of action motivating the use of feedback for individual improvement is that feedback tells employees how their performance compares to pre-determined performance expectations (such as principals' instructional leadership standards), giving them information they can use to then adjust their behavior to meet expectations (Carver & Scheier, 1982; Kluger & DeNisi, 1996).

In practice, however, the content and communication of feedback tends to vary across employees. Research in organizational psychology and resource management has found biasing effects of employee gender and race on the tone and language in which feedback is communicated (Biernat, Tocci, & Williams, 2012; Correll et al., 2020; Jampol & Zayas, 2020; Smith et al., 2019), a phenomenon that is particularly concerning considering that employees from marginalized groups often already experience unfair disadvantages in the workplace in practices like hiring, evaluation ratings, compensation, and promotion (Castilla, 2008; Correll, Benard, & Paik, 2007; Correll & Simar, 2016; Fernandez-Mateo & Fernandez 2016; Stauffer &

Buckley, 2005). And while research indicates that bias in feedback can occur regardless of the evaluator's background characteristics (Jampol & Zayas, 2020), prior work on educator evaluation suggests an advantage to having a same-race evaluator in terms of subjective performance ratings (Campbell & Ronfeldt, 2019; Grissom, Blissett, & Mitani, 2018), suggesting the potential for similar patterns related to feedback processes.

Although evidence of gender and racial biases in performance feedback is well-documented in the broader organizational management literature, research has yet to thoroughly examine this issue in the context of principals' evaluation feedback. The K-12 context is a particularly important setting in which to investigate gendered and racialized differences in how performance feedback is communicated, given the documented inequities in the implementation of high-stakes evaluation regimes (e.g., Grissom & Bartanen, 2022).

Our study examines the affective language, or tone, of the formal evaluation feedback that principals receive from their supervisors. Research has documented that the affective tone – defined as the extent to which the evaluative undertone of feedback is positive or negative – can shape the ways employees react to feedback, which in turn can influence the degree to which they understand, accept, and implement the feedback suggestions (Baron, 1993; Fedor et al., 2001; Nelson & Schunn, 2009 Audia & Locke, 2003; Kluger & DeNisi,1996). The nature of affective language in performance feedback has been widely studied in the fields of organizational psychology and resource management (e.g., Chung et al., 2008; Gerull et al., 2019; Jampol & Zayas, 2020) but not in educational systems. The purpose of this paper is thus twofold: first, to describe the affective language of principals' written formal evaluation feedback from their supervisors, and second, to examine variation in affective language according to principal gender and race.

Our study is situated in Tennessee, one of the first states to design and implement a comprehensive, mandated principal evaluation system. The state's principal evaluation system mandates that principals receive a minimum of two site visits and corresponding evaluation ratings, including written feedback, per year. With each site visit, principals should receive detailed written feedback in two areas: (1) an area for improvement, which is called the *refinement* feedback, and (2) an area of effectiveness, which is called the *reinforcement* feedback. Evaluators share the refinement and reinforcement feedback with principals verbally and in writing during post-observation conferences.

We analyze the written refinement and reinforcement feedback text data for all principals in the state of Tennessee across three years, 2014-15 through 2016-17. Our dataset includes 7,292 observations that are unique at the level of principal-evaluation-year, and the unit of analysis is the written feedback that principals receive from their evaluators. Using a combination of sentiment analysis and regression analytic methods, we ask the following research questions:

1. What is the affective language of principals' written evaluation feedback?
2. To what extent does the affective language of principals' written evaluation feedback vary by the gender and race of the principal?
3. To what extent does the affective language of principals' written evaluation feedback vary by principal-evaluator gender and racial similarity?

We find that evaluators use similar affective language to communicate refinement and reinforcement feedback, even though the two types of feedback intend to support contrasting developmental purposes. We also find differences in feedback by leaders' gender, such that female principals tend to receive feedback with less positive affective language relative to male

principals; we find some suggestive evidence of similar differences by principal race. Last, we do not find evidence that gender or racial similarity between principals and evaluators is associated with differences in the affective language of principals' written feedback. Our study seeks to bring attention to the importance of written feedback as part of the ongoing development and support component of principal evaluation systems, and the equity implications of how feedback is communicated to principals.

We structure the remainder of the paper as follows. We first present a theory of action motivating the importance of affective language in principals' evaluation feedback. We then draw on literature largely from organizational psychology and management to summarize the evidence on gender and racial bias in performance feedback. We next provide a brief introduction to the Tennessee context by describing the state's principal evaluation system and performance feedback structure. We then describe the data and outline the empirical strategies used in our analyses. Finally, we discuss our findings and limitations of the study methodology, and conclude with a discussion of implications for policy and practice and suggestions for future work.

The Importance of Affective Language in Performance Feedback

Our study draws on a robust body of work in the fields of industrial-organizational psychology and human resource management that suggests that feedback is important to employee growth and development (Ilgen, Fisher, & Taylor, 1979; Murphy & Cleveland, 1995). Feedback intervention theory (Kluger & DeNisi, 1996) and control theory (Carver & Scheier, 1982) suggest that individuals alter or regulate their behavior after receiving feedback and comparing it with goals or standards. Feedback indicating that an employee meets or exceeds

expectations can signal that performance and behaviors can remain unchanged (e.g., affirmative feedback, reinforcement feedback), whereas feedback suggesting that behavior is below expectations theoretically should prompt changes in practices aimed at closing the performance gap (e.g., constructive feedback, refinement feedback).[1] The fundamental premise underlying this theory of action is that cognitive dissonance induces a psychologically uncomfortable state that motivates the feedback recipient to reduce dissonance by changing behaviors and practices (Festinger, 1957). A discrepancy between behavior and a standard of expected performance, such as an internally or externally defined expectation of principal performance, can increase motivation to reduce the dissonance by acting upon feedback.[2]

Empirical evidence suggests that it is not the feedback itself, but rather how it is communicated, that determines the extent to which it is a useful tool for development (Adler et al., 2016). In particular, prior work highlights the importance of the affective language of feedback (Audia & Locke, 2003; Brutus, 2010). Affective language is the extent to which the evaluative overtone of feedback is positive or negative. Affective language can be expressed via adjectives ("excellent" versus "disappointing") as well as through the feedback's overall message, such as "the principal is unable to complete tasks in a timely manner" (Brutus, 2010).

Research in human resource management has identified affective language as one of the most critical elements of feedback due to its power to influence how recipients hear and digest the information (Audia & Locke, 2003). Because feedback is often accompanied with affective reactions that can distort the recipient's ability to process it, the tone in which feedback is communicated plays an important role in ensuring that recipients accurately interpret the feedback suggestions (Kluger & DeNisi, 1996). Affective language also sends signals about the relative significance and immediacy of the feedback, as individuals make note of linguistic cues

that distinguish between feedback that they surmise they can or should ignore and feedback that they think should act upon (Kluger & DeNisi, 1996).

Research notes the importance of aligning feedback tone with the goals of the feedback (Nelson & Schunn, 2009). The goal of affirmative feedback is to underscore areas of strength, and consequently, it is most effective when communicated in a positive and celebratory tone that cues satisfactory performance. However, when constructive feedback, geared toward highlighting performance gaps and changing behavior, is overly positive or downplays poor performance through hedging, questioning, or other mitigation techniques, it can message that the feedback is inconsequential or minor, thus decreasing the likelihood of prompting changes in behavior (Nelson & Schunn, 2009). Research also suggests that constructive feedback that is coupled with affirmative feedback that includes praise, and other positive language can strengthen the recipient's perception of the supervisor and his or her credibility, resulting in better feedback implementation (Nelson & Schunn, 2009).

Feedback that comes across as overly negative, critical, or controlling – regardless of whether it is affirmative or constructive – can be frustrating or discouraging to recipients, often weakening the relationship between feedback provider and recipient and deterring recipients from engaging in efforts to improve performance (Baron, 1993; Fedor et al., 2001), and, can, in some cases, lead to employee aggression (Barry, Chaplin, & Grafeman, 2006). Careful consideration of affective language is thus critical to ensuring feedback systems are likely to lead to improved performance.

Gender and Racial Bias in Performance Feedback

A key component of most performance evaluation systems is that evaluators, or supervisors, often have considerable discretion in how they communicate performance feedback to employees. This discretion can introduce biases into evaluation and feedback systems (Castilla, 2015; Dobbin et al., 2015; Williams, Muller & Kilanski, 2012). Even when evaluation and feedback policies are clearly defined and closely tied to performance rubrics, evaluators must make decisions about how to construct and communicate evaluation feedback. This degree of discretion opens room for biases to creep into standardized evaluation and feedback procedures – most notably, as research shows, based on employee gender and race.

A rather large literature has documented gender differences in the communication of performance feedback (e.g., Correll et al., 2020; Gerull et al., 2019; Jampol & Zayas, 2020; Smith et al., 2019). One strand of work has focused on understanding language differences in the way women and men are described in written evaluative comments, showing that women are more likely to be described using communal language, such as "helpful" or "nice," while men tend to be described with agentic language and standout adjectives, such as "assertive" and "extraordinary," that highlight professional achievements (Axelson et al., 2010; Correll et al., 2020; Madera, Hebl, and Martin 2009; Schmader, Whitehead, and Wysocki 2007; Trix and Psenka 2003).

Another strand of research has examined gender differences specifically in the affective language of performance feedback. The evidence in this area is decidedly mixed. A study examining law associates' evaluation feedback found that women are more likely to receive comments with more positive words (Biernat et al., 2012), as did an experimental study in which participants were asked to provide written feedback to employee profiles with randomly assigned gendered names (Jampol & Zayas, 2020). The authors found that study participants positively

distorted feedback to women but not to men (Jampol & Zayas, 2020). Two other studies, one of medical residents (Gerull et al., 2019) and another of U.S. naval students (Smith et al., 2019), in contrast, have shown that feedback to men is more likely to be positive in affective language than feedback provided to women, while yet another study of employees in a Fortune 500 company found no gender differences in feedback tone, but did find, in line with prior work, that women were more likely to be described with communal language and men with standout language (Correll et al., 2020).

The more limited research on racial differences in the affective language performance feedback has found little evidence of differences by recipient race. In a study of bank employees' performance evaluations, Wilson (2010) found no racial differences in feedback tone but acknowledged that the very low frequency of negative comments in the sample itself could explain the lack of detectable differences between Black and white employees. Still, another experimental study found similar results. Chung et al. (2008) studied randomly assigned supervisor-employee counseling pairs and found no differences in the tone of feedback by employee race. In the context of teacher feedback to students, research has noted the presence of a positivity bias, where white teachers tend to provide more positively worded comments to Black students than to white students (Harber, 1998; Harber et al., 2012), though this finding may not be aligned with feedback to adults.[3]

Our study seeks to examine whether and to what extent there are differences in the affective language of principals' written evaluation feedback based on principal gender and race. Research on educator evaluation suggest that patterns of gender and racial biases appear in teacher's subjective observation ratings, where male teachers and teachers of color tend to receive lower ratings than female and white teachers, often even after controlling for other

measures of instructional quality (Campbell & Ronfeldt, 2018; Drake, Auletto, & Cowen, 2019; Grissom & Bartanen, 2022; Jiang & Sporte, 2016). Prior work on Tennessee's principal evaluation system found that female principals receive higher numerical performance ratings than their male counterparts, while Black principals tend to score lower than white principals (Grissom, Blissett, & Mitani, 2018), alarming evidence that principals may be advantaged or disadvantaged in the way they are evaluated and supported based on immutable ascriptive characteristics. Some work has also examined whether there is a benefit to having a demographically similar evaluator. One study on teacher evaluation found that teachers score higher performance ratings when observed by a principal of the same race, though there were no benefits to having a same gender evaluator (Grissom & Bartanen, 2022). On the other hand, a recent study on principal evaluation found no discernable benefits to having an evaluator of the same gender or race in terms of principals' attitudes towards evaluation, including perceptions of the specificity and utility of evaluation feedback (Nelson, Grissom, & Cameron, 2021). We thus build on this body of work by exploring whether principals' race and gender, as well as the gender and racial similarity of principal-evaluator pairs, influences the affective language, or tone, of the written feedback that principals receive as part of a statewide evaluation system.

Tennessee's Principal Evaluation System: The TEAM Model

We situate our study in the context of Tennessee, one of the first states to implement a statewide principal evaluation and feedback system. In the 2011-2012 school year, as part of the state's successful Race to the Top bid, Tennessee adopted the Tennessee Educator Acceleration Model (TEAM) as the main school leader evaluation model across the state.

To help principals develop and improve their leadership performance, TEAM requires evaluators to observe principals during site visits and provide post-observation performance feedback using a standards-based rubric. Evaluators are typically the principal's supervisor, such as the superintendent, assistant superintendent, or supervisor of instruction. All individuals who are first-time evaluators attend an annual training held by the Tennessee Department of Education (TDOE) on how to conduct observations and site visits, assign performance ratings, and facilitate post-observation conferences. Following the training, evaluators must take and pass a certification exam.

The intent of the evaluation policy is that feedback aligned to a standards-based rubric will move principal performance toward exemplary practices, and because the rubric captures leadership practices that are linked to schoolwide improvement, feedback should ultimately improve school performance. According to the evaluation policy, principals are required to receive at least two site visits per academic year, once in the fall semester and once in the spring. Principals are evaluated based on a rubric adapted from the Tennessee Instructional Leadership Standards (TILS). The TILS rubric, first developed in 2008 and finalized in its current version in 2013, intends to comprehensively capture the practices of effective instructional leadership (Tennessee State Board of Education, 2015). The rubric includes 17 performance indicators grouped into four distinct domains: (a) Instructional Leadership for Continuous Improvement; (b) Culture for Teaching and Learning; (c) Professional Learning and Growth; and (d) Resource Management.

During each site visit, evaluators assign performance ratings on each of the 17 indicators on the TILS rubric, with scores ranging from one (significantly below expectations) to five (significantly above expectations). In addition to scoring the rubric, evaluators also choose two

indicators to provide in-depth written and oral feedback. Refinement indicators are areas for improvement, and reinforcement indicators are areas of demonstrated effectiveness. Within a week of the site visits, evaluators are required to schedule post-observation conferences to discuss the refinement and reinforcement feedback and to support principals in developing feedback action plans. Evaluators also provide principals with the written feedback and enter it into TDOE's information management system.[4] Our analysis makes use of these written feedback entries.

The nature of the TEAM model presents an interesting context in which to study the affective language of principals' performance feedback. First, Tennessee was an early adopter of a principal evaluation feedback system, which gives us the opportunity to explore feedback implementation over multiple years in a relatively novel policy context. Second, because Tennessee employs a statewide evaluation system, we are able to generalize our results to all principals across a state. Third, the Tennessee context is unique in that TEAM policy requires evaluators to enter their written feedback into the state's data management system, giving us access to unique micro-data on the actual written feedback that principals receive. We acknowledge that we do not observe evaluators' verbal feedback communicated during post-observation conferences, and accordingly, we cannot measure the extent to which the written feedback text mirrors what evaluators discuss with principals in person. However, the use of written feedback data offers the advantage of generalizability, as we are able to observe all principals in the state over multiple years, instead of, for example, focusing on a smaller sample of principals for qualitative observations. Lastly, although the evaluation policy is clearly defined in its implementation processes, there is little guidance and training in developing written

feedback, allowing for the study of how evaluators communicate feedback and how gender and race influence this process.

Data

We use the written refinement and reinforcement feedback text data for principals in the state of Tennessee in academic years 2014-15 through 2016-17, years for which written feedback text data are available. Our sample includes all principals in the state of Tennessee who received written evaluation feedback under the TEAM model. The unit of analysis is the written feedback entry observed at the principal-evaluation level. As such, each individual principal evaluation results in two units of data: a refinement entry and a reinforcement entry.

The feedback entries are linked to unique identifiers for both the principal and evaluator, which we use to merge the feedback text data with TEAM evaluation data and administrative staff files. The evaluation data include refinement and reinforcement indicators, indicator-specific performance ratings, and overall average performance rating. The administrative data include background and demographic information for individual principals and evaluators, including their gender, race/ethnicity, prior education history, age, and years of experience. Although the administrative data does not capture years of experience as a principal or evaluator, respectively, we construct these measures based on educators' job history, which date back to the 2001-2002 academic year.[5] The administrative data also provide information on principals' schools, such as enrollment, school level, and the socio-demographic composition of students with respect to race/ethnicity and eligibility for the free- or reduced-price lunch (FRPL) program. We also construct school-level achievement information to create a standardized summary achievement index for each school, which is the weighted average of all standardized test scores

across grades and subjects in each year.[6] We supplement these data with information on school-level urbanicity from the Common Core of Data. In all, our dataset includes information on principals and their evaluators, and the schools in which they serve.

[INSERT TABLE 1 HERE]

Table 1 displays summary statistics for our analytic sample pooled across years. Just over half of principals are female (56 percent) while only 11 percent are Black. 50 percent hold advanced degrees (Education Specialist or Doctorate) and have on average six years of experience in the principalship. The schools that principals lead tend to be largely white; in the average school, 17 percent of students are Black and seven percent of students are Hispanic. The mean school-level free- and reduced-price lunch eligibility is 59 percent, and the plurality of schools are located in rural districts (34 percent) with the second highest being urban districts (22 percent). In our sample, principals receive just under two evaluations per year, on average, evidencing a degree of policy non-compliance, as TEAM policy requires that principals receive at least two formal evaluations in each academic year.

In comparison to principals, slightly fewer evaluators are female (51 percent) and Black (nine percent), 63 percent of evaluators have advanced degrees and have worked in Tennessee schools for about 27 years on average but have just over two years of experience in the evaluator role. Just under half of all evaluations were conducted by superintendents (40 percent), followed by 29 percent by supervisors of instruction, 12 percent by assistant superintendents, 12 percent by individuals in other central office positions, including federal and special programs, human resources, and school improvement and accountability personnel. The remaining seven percent of evaluations were conducted by principal peers or other school-based staff.[7]

Feedback data represent the written versions of the post-observation feedback for principals' refinement and reinforcement objectives, which evaluators enter into the state management system. The feedback entries tend to be short in length, with most comprising no more than a couple sentences. Figure 1 shows the distribution of word counts for the written feedback text. The total number of words, which is the sum of refinement and reinforcement text, ranges from two to 758, and is right skewed with a median of 55 and mean of 73. As shown in Figure 1, word counts for refinement and reinforcement separately follow almost identical distributions. On average, principals receive slightly more refinement feedback than reinforcement feedback, as the mean ratio of refinement to reinforcement is 1.18. As an example, we provide two feedback entries for refinement and reinforcement feedback below, both of which pertain to the same indicator (Environment) in the Culture for Teaching and Learning domain on the TILS rubric.

Refinement: There is room for improvement around the environment at X Elementary. According to recent survey results and through conversation with an outside consultant, there seems to be a lack of trust and respect among the faculty and staff. It is hard to reach full potential without trust. Mr. X is encouraged to engage in cross grade planning and collaboration in order to strengthen the working relationship among his staff. He is also encouraged to be a more visible presence with his teachers and families in order to build positive relationships with all.

Reinforcement: Mr. X provides the students who are at most risk in the county an environment that gives the students a sense of hope. The students are surrounded by teachers, parents, and community stakeholders who are totally involved in setting a path for success. The principal has a very stringent code of conduct, keeps its surroundings very conducive to learning, and sets the expectation for learning with the goal of returning to their home school.

[INSERT FIGURE 1 HERE]

Analytic Methods

Our analysis proceeds in two main steps. First, we use sentiment analysis to characterize and describe the affective language of principals' refinement and reinforcement feedback.

Sentiment analysis falls within the broader domain of text-as-data methods, which encompass a vast array of computational techniques used to analyze large quantities of textual information, often verbal and written communication transcripts (Ferreira-Mello et al., 2019; Gentzkow, Kelly, & Taddy, 2017). Text-as-data methods allow researchers to identify trends in raw text data using automated processes, facilitating the systematic processing of qualitative data at a scale and speed that would be impossible with traditional qualitative coding methods (Liu & Zhang, 2012). Researchers in education policy have applied text-as-data methods to explore perceptions of student achievement gaps in essays of teacher applicants (Penner et al., 2019), to categorize school improvement reform strategies based on planning and implementation reports (Sun et al., 2019), and to create measures of instructional quality based on teacher-student classroom discourse (Liu & Cohen, 2021).

Within the broad category of text-as-data methods, sentiment analysis relies on the semantic orientation of words and document-level word counts to construct objective measures of the opinions, sentiments, attitudes, and emotions that appear in a given document (Cambria et al., 2017).[8] Though relatively new to education policy research, these methods have already made contributions to the field. For instance, Fesler et al. (2019) use sentiment analysis to examine whether female and male students receive different amounts of positive or negative sentiment in online class discussion forums. There is also some precedent for using sentiment analysis to analyze written evaluations of individuals. Akos & Kretchmar (2017) apply sentiment analysis to examine gendered and racial language patterns in school counselors' letters of recommendations for undergraduate admissions, focusing on differences in the relative frequency of traditionally gendered and racialized word groups, including communal, standout, ability, and grindstone words. Loftus & Tanlu (2018) use sentiment analysis to examine whether

the presence of causal language in employees' feedback improves future performance, and Bludevich et al (2021), in a similar approach to ours, use LIWC-dictionary based sentiment analysis to examine gender differences in letters of recommendation in the medical field.

In our analysis, we use a dictionary-based method that makes use of available dictionaries that classify words into distinct categories. We use the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2015), which is widely used in computational linguistics. LIWC incorporates text samples from a variety of sources that have been coded by experts with extensive reliability and validity checks to create a dictionary that covers 93 linguistic dimensions, such as emotions, cognitive processes, social concerns, and various groups of functional words (see Pennebaker et al., 2015 for the full documentation).[9] From the dictionary, we select three dimensions to measure the affective language of principals' feedback text: tone, positive emotion, and negative emotion. Tone is a summary variable that draws on the percentage of words across multiple linguistic dimensions (beyond positive and negative words) to capture the overall tone of the text on a 100-point scale, where higher values represent a more positive and upbeat tone and lower values represent a more negative tone; this construct was derived from previous work in psychological linguistics (Cohn, Mehl, & Pennebaker, 2004; Pennebaker et al., 2015). Positive and negative emotion represent the percentage of words in a document that reflect positive (happy, nice, great) or negative (hurt, nasty, bad) emotion, respectively.[10]

The LIWC dictionary-based approach emphasizes the role of individual words in measuring the overall affective language of feedback text by categorizing at the word level, then aggregating word-specific categorizations to the feedback entry level. For example, to construct the measures of positive emotion, each word in a feedback entry is categorized as positive or not,

then the algorithm calculates the overall percentage of positive words in the feedback entry. A score of 5 for positive emotion indicates that 5 percent of words in the feedback entry are coded as positive, and the remaining 95 percent are not. A limitation of this approach is that it does not take into account the surrounding context of individual words within paragraphs, or within the overall feedback entry, as is the case in more robust analytic approaches like machine learning based models. However, we adopt the dictionary-based approach here for three main reasons: (1) prior literature on performance feedback that has highlighted the importance of word-specific differences (e.g., Correll et al., 2020); (2) the feedback data are short in length; and (3) the feedback entries represent relatively high-stakes performance documentation, as they are stored in principals' personnel files and thus particular words may carry triggering effects (Heen & Stone, 2014) .

We focus our analysis specifically on affective language as measured by tone, positive emotion, and negative emotion based on prior literature evidencing the importance of affective language to the process of feedback acceptance and implementation (Audia & Locke, 2003; Kluger & DeNisi,1996). We conduct the analysis separately for refinement and reinforcement feedback entries, constructing sentiment scores for each feedback type, which allows us to compare the affective language across the two types of feedback that serve distinctly different purposes.[11]

The second part of our analysis involves examining gender and racial differences in the affective language of feedback. We begin by using simple t-tests to examine bivariate relationships between our outcomes of interest – tone, positive emotion, and negative emotion – and principal gender and race, respectively. Across all analyses, our investigation of racial differences is limited to differences between Black and white principals due to the small number

of non-Black and non-white principals in Tennessee. We next employ these measures of affective language as dependent variables in linear predictive models to examine the extent to which any gender and racial differences persist even after controlling for other observable principal, evaluator, and school characteristics. Again, we run separate models for refinement and reinforcement feedback. Our models take the following form:

$$AffLang_{ijkszt} = \alpha + \beta_1 Female_j + \beta_2 Black_j + \delta \boldsymbol{X_j} + \theta \boldsymbol{W_k} + \pi \boldsymbol{S_s} + \tau_d + \gamma_t + \varepsilon_{ijkszt} \quad (1)$$

where *AffLang* indexes the outcome of interest (tone, positive emotion, negative emotion) for feedback entry *i* for principal *j* with evaluator *k* in school *s* in district *z* in year *t*. The primary coefficients of interest are $\beta_1$and β_2, which represent the difference in affective language for female principals relative to male principals, and for Black principals relative to white principals, respectively. **X** is a vector of observable principal characteristics, including age, level of education, years of experience, average performance rating, and number of evaluations. **W** is a vector of evaluator characteristics, including gender, race/ethnicity, age, level of education, years of experience in Tennessee schools, years of experience as an evaluator, role, and span of control measured as the number of principals evaluated. **S** is a vector of school characteristics, including enrollment, school-level achievement index, proportion FRPL-eligible students, proportion Black students, proportion Hispanic students, school level. To account for the non-random sorting of principals and evaluators across districts, as well as district-specific differences in feedback implementation that may affect the affective language in which evaluators communicate feedback, we include a district fixed effect, τ. We also include a year fixed effect, γ to account for any potential year-over-year changes in feedback communication practices that affect all districts. Across all models, we cluster standard errors by district.[12, 13]

In a final analysis, we explore the association between affective language and principal-evaluator gender and racial similarity. In the first set of models, we modify equation (1) to include interaction terms for *principal is female* and *evaluator is female*. In the second set of models, we replace the gender interaction terms with interactions for *principal is Black* and *evaluator is Black.* We are not able to consider other racial or ethnic groups due to very small sample sizes of non-white and non-Black principals and evaluators in Tennessee.

Results

To address the first research question, what is the affective language of principals' evaluation feedback, we first provide qualitative examples to contextualize the measures of affective language (tone, positive emotion, and negative emotion); we next describe the distributions of these measures to characterize the affective language of principals' refinement and reinforcement feedback.

We start by providing examples of the range of affective language in the refinement and reinforcement feedback data to contextualize the scale of our affective language measures within the written text. Below are two examples of *refinement* feedback entries to highlight differences between positive and negative affective language. The first feedback entry has positive affective language (tone = 99, positive emotion = 8.82, negative emotion = 0), the second has negative affective language (tone = 1, positive emotion = 0, negative emotion = 3.45).

> **Positive Affective Language:** Data Analysis and Use is an indicator of significant strength for you. It is important for you to continue to support both assistant principals in this work. Allowing them the latitude to approach data analysis differently will be beneficial to all. Ensuring teachers are engaged in these discussions will allow conversations to focus on proven instructional practices that make positive differences for children in their performance and progress. *(tone = 99, positive emotion = 8.82, negative emotion = 0)*

> **Negative Affective Language**: This continues to be a focal area as related to both the Teacher Perception Survey and observations. The area is associated with your willingness to make difficult decisions in the face of adversity. Teachers need to know you have the will to make those tough decisions even if those decisions are opposed by a few long-time employees. *(tone = 1, positive emotion = 0, negative emotion = 3.45)*

These examples demonstrate a stark difference in feedback affective language. In the first example, the evaluator asserts that the refinement indicator is in fact an "area of strength" while offering and explaining the importance of suggestions for improvement. On the other hand, in the second example, the evaluator states that the refinement indicator "continues" to be a focal area, suggesting that the principal has failed to make improvements in this area and implying that the need for improvement is due to a lack of willingness on the part of the principal.

We see similar differences between positive and negative affective language in reinforcement feedback. Below are two *reinforcement* feedback entries, where the first demonstrates positive affective language (tone = 99, positive emotion = 8.82, negative emotion = 0), and the second demonstrates negative affective language (tone = 1, positive emotion = 0, negative emotion = 3.57).

> **Positive Affective Language:** You make your school a great place to work! You provide clear expectations and you are fair and consistent. You collaborate with others in your building to find ways to support teachers' professional growth. (*tone = 99, positive emotion = 8.82, negative emotion = 0)*

> **Negative Affective Language**: Concern of the teacher evaluations being very high. Mostly 5's. We need to be very objective in order to help our teachers with any possible weak areas. *(tone = 1, positive emotion = 0, negative emotion = 3.57)*

In the first example, the evaluator highlights areas of success for the principal (expectations, consistency, collaboration) and commends his/her efforts in creating a positive work environment. In the second example, the evaluator does not in fact highlight an area of effectiveness – the purpose of reinforcement feedback – but rather raises concerns about skewed

teacher evaluation scores, suggesting that the principal does not evaluate teachers in an objective manner.

Table 2 shows the distributions of each affective language measure for refinement and reinforcement feedback, and Figure 2 displays this information graphically via histograms. We find that despite being designed for different developmental purposes, both refinement (areas that need improvement) and reinforcement (areas that meet expectations) feedback are generally neutral to moderately positive in affective language; refinement feedback is less positive than reinforcement feedback. The measure of tone presents an almost bimodal distribution, with concentrations of feedback entries coded as neutral and positive.[14] For positive emotion, about six percent of words in the average reinforcement feedback are positive words, while just over four percent of words in the average refinement feedback text are positive words; and both feedback types have around 30 percent of entries with no words classified as positive. Both refinement and reinforcement feedback exhibit very limited negative emotion; upwards of 80 percent of entries have no words classified as negative, and the average percent of negative words at less than one percent in both feedback types with little variation around the means. Feedback that is communicated with high levels of negative emotion may be demotivating or even demoralizing and may contribute to low levels of feedback take up (Baron, 1993; Fedor et al., 2001).

[INSERT TABLE 2, FIGURE 2 HERE]

Although the relative affective language of refinement and reinforcement feedback aligns to their competing developmental purposes (i.e., the affective language of refinement feedback is slightly less positive than that of reinforcement feedback), both types of feedback exhibit similar overall affective language. These descriptive similarities capture parallels in affective language

that are not in fact driven by the length of the feedback that principals receive, as principals tend to receive the same number of words of refinement and reinforcement feedback, on average (see Figure 1).

The general lack of distinction in affective language between refinement and reinforcement is qualitatively apparent in the written text. Here, we provide a comparative example of refinement and reinforcement feedback that are both scored as a zero for positive emotion:

> **Refinement:** Continue to develop procedures to ensure that all teachers are accountable for student performance.
>
> **Reinforcement:** Continue to leverage resources through community partners.

Both feedback entries exhibit similarities in their overall affective language and in their lexicon (i.e., the word "continue" appears as the first word in both entries). From the perspective of a feedback recipient, it is difficult to distinguish the developmental differences between the two pieces of feedback – which one encourages changes in behavior and which one supports continued effective practices. This finding raises questions as to whether principals are receiving pointed feedback that adequately signals a need for improvement versus support for existing practices.

We now turn to the second research question of whether there are detectable differences in affective language by principal gender and race. We begin with simple t-tests to show the unadjusted relationship between affective language and principal race and gender, shown in Table 3. Starting with differences by principal gender, we find statistically significant mean differences in some but not all measures of affective language. Among refinement feedback, female principals receive feedback that is slightly more positive in tone (mean difference = 1.37) relative to male principals.We do not find statistically significant differences by gender for

positive or negative emotion. Among reinforcement feedback, the mean difference for tone does not reach levels of statistical significance. We do, however, note a statistically significant and negative relationship for positive emotion, where female principals tend to receive feedback with a smaller percentage of positive words (mean difference = -0.37) compared to their male counterparts. We do not observe significant differences in negative emotion. Taken together, female principals see a more positive affective language in their refinement feedback but a more negative affective language in their reinforcement feedback.

[INSERT TABLE 3 HERE]

Turning now to differences by principal race, we find descriptive evidence that feedback to Black principals is communicated differently than feedback to their white peers. For refinement, Black principals receive feedback that is less positive in tone (mean difference = -2.54) and with less positive emotion (mean difference = -0.63) than the feedback that white principals receive. In other words, refinement feedback to Black principals' takes a harsher tone with fewer positive words than refinement feedback to white principals. We do not find statistically significant differences for negative emotion. For reinforcement feedback, similar patterns emerge. Feedback to Black principals is both more negative in tone (mean difference = -4.73) and includes a smaller percentage of positive words (mean difference = -1.09), but does not differ in the percentage of negative words. These findings suggest that Black principals receive feedback – regardless of whether it is for improvement purposes or affirmative aims – that is delivered in a less positive and upbeat manner than the feedback that white principals receive.

Importantly, the differences in Table 3 represent unadjusted mean differences, which may be attributable to the gender or race of the principal, or which may be attributable to other confounding factors. For instance, if Black principals tend to work in districts where feedback

processes are practiced relatively sub-optimally, then the observed differences may reflect principal sorting patterns and differential feedback implementation across district contexts. As another example, if female principals are more likely to be evaluated by experienced evaluators who tend to give more positively worded refinement feedback, then the gender differences in Table 3 would reflect evaluator preference or training rather than gender biases. To further investigate the source of variation in how feedback is communicated to principals, we turn to results from within-district models that seek to isolate the role of principal gender and race in affective language.

[INSERT TABLE 4 HERE]

Table 4 presents results from linear probability models that include observable principal, evaluator, and school covariates, as well as district and year fixed effects. We find that, in these models, the observed gender difference in refinement feedback tone in Table 3 is no longer statistically significant at conventional levels, indicating that evaluators communicate refinement feedback to female and male principals using similar affective language. We do, however, find evidence of gendered differences in the affective language of reinforcement feedback. All else equal, female principals tend to receive reinforcement feedback that is less positive in tone and positive emotion relative to male principals (Panel B, Columns 1 and 2). The coefficient for tone is negative and statistically significant, and represents a difference of 0.08 standard deviations, while the coefficient on positive emotion is negative, marginally significant, and represents a difference of 0.05 standard deviations.

We find suggestive evidence that Black principals tend to receive refinement feedback that is less positive in tone and positive emotion than the feedback that white principals receive (Panel A, Columns 1 and 2), though there are no discernable differences in negative emotion.

The coefficients on tone and positive emotion are marginally significant, and show differences of 0.09 (tone) and 0.10 (positive emotion) standard deviations, suggesting that the affective language of feedback to Black principals is less positive than the affective language of feedback to their observably similar white principal peers. We do not find similar patterns among reinforcement feedback, as the coefficient on Black principal is not statistically significant across all three measures of affective language.

[INSERT TABLE 5 HERE]

Last, to address our third research question, we investigate whether differences in feedback affective language are associated with the gender and racial similarity of the principal and evaluator. In our sample, 54 percent of principals identify as the same gender as their evaluator, and 86 percent identify as the same race as their evaluator. All models include the same set of principal, evaluator, and school covariates as in previous models, and include district and year fixed effects as well.

In Table 5, we show associations for principal and evaluator gender as well as the interaction between the two variables. We do not find evidence of a principal and evaluator gender interaction. For both refinement and reinforcement feedback, the interaction between female principal and female evaluator is not statistically significant across all three measures of affective language, indicating that female principals see no differences in the affective language of their written feedback when they have a female evaluator. Instead, we observe that female evaluators tend to write reinforcement feedback with more positive affective language when they are evaluating male principals (Panel B, Column 1).

[INSERT TABLE 6 HERE]

We next turn to differences by principal-evaluator racial similarity. In Table 6, we show associations for principal and evaluator race, as well as the interaction between principal and evaluator race. The interaction terms that focus on principal and evaluator race show no evidence of a "race matching" effect, as coefficients on the interaction terms (and main effects) are not statistically significant for refinement or reinforcement feedback.

Discussion

This study documents and describes patterns in the affective language of principals' evaluation feedback. Many states and districts rely on principal evaluation and feedback systems as a primary tool for principals' leadership development (Davis et al., 2011), yet there is little information about what this feedback looks like, especially in the context of a mandated, statewide evaluation system. As research on principal evaluation continues to highlight the need for effective feedback structures to support principal improvement (e.g., Nelson et al., 2021, Donaldson et al., 2021), we extend this prior work by focusing specifically on principals' written feedback in the context of a statewide evaluation system that requires evaluators to observe and provide performance feedback to principals for the purpose of ongoing instructional leadership improvement. We draw on a substantial body of work in resource management and organizational psychology that highlights the importance of affective language to the feedback implementation process and that documents variation in affective language according to the gender and race of the feedback recipients. Using a combination of dictionary-based sentiment analysis and standard regression techniques, we uncover and examine the affective language of principals' written evaluation feedback, for all principals in the state as well as with particular attention to those from groups traditionally underrepresented in leadership.

Our results describe the affective language of principals' written evaluation feedback, drawing attention to the similarities in affective language between refinement and reinforcement feedback. Despite being designed for different developmental purposes, we find that refinement and reinforcement feedback both tend to take a moderately positive tone. These results suggest that evaluators communicate refinement feedback – feedback that intends to identify performance gaps and areas for improvement – in a tone that is more consistent with what we might expect for reinforcement feedback – feedback that celebrates successes and areas of strength. This finding is consistent with prior work that has shown that evaluators are often reluctant to provide critical feedback in a serious (not sugar-coated) tone because they anticipate hostile reactions, and in anticipation of a heightened emotional response, may increase the amount of praise in their feedback in an attempt to temper the reaction (Cleveland et al., 2007). Research on teacher evaluation has found that principals tend to avoid providing feedback to teachers on areas of growth altogether (Kraft & Gilmour, 2016).

The similarities in tone between refinement and reinforcement feedback may hold important implications for feedback take-up. While it appears encouraging that principals' feedback generally takes a neutral and moderately positive tone - as feedback that is communicated in a negative or harsh tone may be demoralizing to recipients, leading to low feedback take up and perhaps, in some cases, employee aggression (Barry, Chaplin, & Grafeman, 2006; Fedor et al., 2001) – how evaluators communicate feedback signals the relative importance of feedback suggestions. The fact that the affective language of refinement feedback largely resembles that of reinforcement feedback may lessen the perceived seriousness of immediacy of refinement suggestions, thus decreasing principals' likelihood of feedback implementation (Nelson & Schunn, 2009). The language used to communicate feedback helps

recipients place the feedback against their existing behavior and identify the appropriate response. Without clear linguistic clues in the lexicon and affective language, principals must engage in higher levels of information processing to distinguish between feedback that communicates areas for improvement and feedback that highlights areas of strength, as the feedback itself does not present a clear signal to motivate a change, or not, in behavior. A reluctance to provide tough but honest refinement feedback in a tone that conveys its seriousness may limit the degree to which the evaluation feedback actually serves its intended purpose of improving principal performance (Adler et al., 2016). If differences in affective language do not adequately distinguish between the distinct goals of refinement and reinforcement, principals may not be receiving the critical and targeted communication signals that they need to improve.

Our results also point to differences in feedback communication by characteristics of the principal. Female principals receive reinforcement feedback that is, on average, less positive in tone than the feedback communicated to their male peers. Female principals may be disadvantaged in the way their reinforcement feedback is communicated, as the very purpose of reinforcement feedback is to provide affirmative support by acknowledging what a principals are already doing well and encouraging them to continue that behavior. While the broader research base on gendered differences in feedback is mixed in terms of the direction of the bias, we note that our findings align with prior research on feedback to women in leadership positions that shows that women receive less positive feedback (Gerull et al., 2019; Smith et al., 2019); while research on feedback to women who are not in leadership positions has generally shown that feedback tends to be more positive in tone (Biernat et al., 2012; Jampol & Zayas, 2020).

We find some evidence that Black principals are more likely to receive refinement feedback that is communicated in a less positive tone and with fewer positive words relative to

their white counterparts. That is, the affective language of Black principals' written feedback is overall less positive compared to white principals', though the results are marginally significant. These findings stray somewhat from prior literature that has found no evidence of racial differences in feedback tone (Chung et al., 2008; Wilson, 2010), but are more consistent with research on Tennessee's principal evaluation system that points to similar racial differences in principals' numerical performance ratings (Grissom et al., 2018).

We do not find evidence that female or Black principals necessarily receive feedback that is different in affective language based on the gender or race of their evaluator. However, our findings revealed an interesting pattern in feedback tone in gender-dissimilar principal-evaluators pairs, as we find that female evaluators tend to write reinforcement feedback with more positive affective language when they are evaluating male principals. This finding may suggest that evaluators from groups that are traditionally underrepresented in leadership – like female evaluators – tend to write feedback that is more positive when evaluating a principal whose gender identity places them in a historically advantaged group in leadership, males. While there is limited research on cross-gender differences in feedback communication, prior work in organizational psychology has found that male employees receive more job support (Maume, 2011) and have more favorable attitudes towards evaluation and management (Johansson & Wennblomm, 2017) when they have a female supervisor, which may suggest that female evaluators adopt different approaches to evaluation and feedback communication to male employees. The nature of cross-gender differences in feedback communication in the K-12 setting is an area for future work.

Our findings are important because the tone, or affective language, of feedback communication carries implications for how principals might respond to feedback and the extent

to which they may learn and improve from the evaluative process. If female principals systematically receive feedback that has a more negative affective language male principals, their experience with evaluation feedback may discourage them from taking advantage of evaluative processes to improve their practice. A more concerning interpretation is that affective language is in fact intentional on the part of the evaluator and is meant to send informal messages to principals about career opportunity that are otherwise not possible within formal bureaucratic dismissal policies. Our study cannot speak to the relative plausibility or either scenario – where feedback exhibits an unfortunate sub-optimal use of affective language or where it is used as a mechanism to counsel out principals – and we present this as an important area for future work to consider.

Overall, our study shows that even when feedback practices are standardized and clearly defined, there is still room for evaluators to determine how they communicate feedback to principals. Recent studies have highlighted the use of evaluator discretion in principal evaluation (e.g., Donaldson et al., 2021), drawing attention to district and state efforts to reach consistency in various aspects of the evaluation process, including performance ratings and the use of evaluation results in formal structures for principal improvement (Anderson & Turnbull, 2016; Kimball et al., 2015). Our work highlights another important consideration regarding the implementation of evaluation practices: how evaluators communicate written performance feedback.

To this last point, we offer two main policy implications. District and state policymakers may wish to consider evaluator training focused specifically on how to communicate evaluation feedback, both in-person and in writing. While evaluators in the TEAM system are trained on how to conduct observations and assign performance ratings, they do not receive required

trainings on how to write and communicate feedback to principals. State policymakers could provide training that brings awareness to the importance of affective language in feedback communication, as well as on the use of affective language to distinguish between the separate developmental goals of refinement and reinforcement feedback. For evaluation feedback to contribute to principal improvement, principals must be able to discern between feedback that calls for changed behavior and feedback that signals continued use of effective practices. One of the main ways that principals (and employees, more generally) make these distinctions is from signals in the way feedback is communicated – most often, in the affective language.

Second, district and state efforts should focus on addressing gender and racial disparities in how feedback is communicated. Evaluation feedback may be particularly important for Black principals, who continue to be underrepresented, both nationally and in Tennessee schools (NCES, 2020), and who already experience other disadvantages in their role, including challenging school contexts and potentially biased evaluation ratings (Grissom et al., 2018; Taie & Goldring, 2017). Researchers have suggested that training evaluators to be aware of and self-reflect on their feedback communication can address gendered or racial disparities in how feedback is communicated (Jampol, Rattan, & Wolf, 2023). Other research has pointed to the role of accountability – such as systemwide feedback audits – in reducing the effects of employee characteristics on feedback and other evaluation outcomes (Castilla, 2015; Jampol et al., 2023).

We also note important limitations of our study. Importantly, as noted previously, we acknowledge the LIWC-dictionary approach does not take into account context or relationships between words to measure sentiment, but rather draws on word-specific categorizations to construct these measures. We consider this study to be a first step in understanding principals'

evaluation feedback and urge future research to adopt other methods that take into account broader written context, including machine learning models like transformers, to further investigate feedback tone and other dimensions of feedback that may vary across lines of gender and race. Another limitation is that we do not observe the feedback that evaluators communicate to principals in person. It is possible that evaluators intentionally submit written feedback that is different from what they communicate to principals in person, if, for instance, evaluators intentionally seek (or do not seek) to leave a paper trail of their comments. If these differences are associated with principal gender or race, then our results may be biased in that they under or overestimate the extent of the gender or racial biasing effect on feedback. To inform on the extent of potential differences between verbal and written feedback, future qualitative inquiry could make use of observations of principal and evaluator feedback conferences. In addition, while our analysis measures the affective language of principals' feedback, we do not identify the appropriate affective language for refinement and reinforcement feedback and suggest this as an avenue for future work. Last, we note limits to the generalizability of our study. Our study focuses on the feedback that principals receive from their evaluators in one specific state policy context, and thus may not extend to other feedback systems or contexts. We caution against generalizing our results to other state or district contexts, particularly those in which the principal evaluation system is notably different in design. We encourage future work to extend our initial analysis in other states or districts with different evaluation and feedback contexts.

Our results provide some initial evidence on the nature of principals' written evaluation feedback from their evaluators in a statewide evaluation system. Our study discusses one quality of written feedback – affective language – and we encourage future research to explore other dimensions. For example, future work could also explore other qualities that have been linked to

increased take-up and improved performance, such as the extent to which feedback is actionable, is based on objective evidence, and is specific to the employee's job responsibilities (Aguinis, Gottfredson, & Joo, 2012; Ilgen et al., 1979; Park, Johnson, Moon, & Lee, 2019). Future work could also explore how principals respond to feedback, examining whether differences in affective language are associated with differences in how principals respond to and incorporate feedback.

Notes

1. We use the terms affirmative and constructive to describe types of feedback, where affirmative feedback is feedback on areas of strength (reinforcement feedback) and constructive feedback is feedback on areas of improvement (refinement feedback). We use the terms positive and negative to describe the affect, or tone, of the feedback. Feedback type is theoretically distinct from feedback tone: affirmative (reinforcement) feedback could be communicated in a positive or negative tone and constructive (refinement) feedback could also be communicated in a positive or negative tone.
2. Although recent research has put forth several competing theories to explain dissonance effects, theorists remain in agreement on the basic tenet of cognitive dissonance theory, which is that dissonance induces psychological discomfort, which then motivates individuals to change behavior (Harmon-Jones & Mills, 2019).
3. Another line of research has considered the nature of racial bias in students' letters of recommendation (e.g., Akos & Kretchmar, 2017; Polanco-Santana et al., 2021, Ross et al., 2017). These studies find differences that largely resemble patterns of gendered language differences in performance feedback, such that students from minoritized groups are more likely to be described using communal words compared to white students who are more likely to be described using agentic terms and standout language.
4. During the post-observation feedback conferences, evaluators discuss their written feedback with principals, and both evaluators and principals are required to sign off on the completed feedback form before evaluators submit the written feedback and numerical performance ratings into the state's management system.
5. We are not able to identify years of experience as principal for principals who were already in a principal role in the first year of the administrative dataset. This variable is top-coded, as we only know a minimum value for this group of principals.
6. We observe a higher percent missingness in test score data in the 2015-16 school year due to statewide testing issues. For schools with missing achievement information, we impute the average achievement index based on values from the prior and/or successive years.
7. During this study's time frame, some districts, such as Davidson County, piloted alternative leadership models in which principals took on district-level leadership responsibilities, including principal evaluations. As the principal peer model represents a slight deviation from the traditional TEAM model and is no longer employed in any

districts, we run analyses both with and without principal peer evaluators and find qualitatively similar results.

8. In the text-as-data literature, the term "document" refers to one observation of text data (e.g., one feedback entry).
9. The idea behind the LIWC dictionary is that the types (and frequencies) of words that appear in a given document reveal underlying psychological, social, and emotional constructs. For example, as the LIWC documentation describes, a document that is concerned with power is more likely to include words such as "boss, underlying, president, Dr., strong, and poor" compared to a document that does not exhibit an interest in or focus on power (see https://liwc.wpengine.com/how-it-works/). This lexical-based approach guides the development of each of the 93 linguistic dimensions in the dictionary.
10. In our data, tone and positive emotion are highly correlated (0.75 for refinement and 0.68 for reinforcement), while tone and negative emotion show weak negative correlations (-0.20 for refinement and -0.15 for reinforcement). Similarly, inter-item correlations between positive and negative emotion indicate weak and negative relationships, at -0.15 for refinement and -0.04 for reinforcement.
11. Prior to conducting our analyses, we begin by pre-processing the text data. This process includes converting the text to lowercase, removing all punctuation and numbers, and identifying and editing common misspellings. Pre-processing also includes removing stopwords, which are words that appear frequently but do not contain much information (like "the" or "a"), and stemming words, which removes suffixes such that the words "evaluating" and "evaluate" become "evaluat" and are thus treated as the same word in the analysis (Hull, 1996; Manning, Raghavan, & Shutze, 2008; Porter, 1980). In addition, we manually remove two words that appear frequently given the context of our data but that do not carry sentiment: "refinement" and "reinforcement."
12. We observe a moderate degree of missing data across all observable characteristics (see Table 1). As the primary approach to account for missing data, we chose to include indicators for missing values as additional covariates in our models for the sake of analytic simplicity. We include missing indicators only for variables that are not of primary interest (i.e., variables other than principal race and gender, and evaluator race and gender). Our results are qualitatively similar with and without the inclusion of missing indicators.
13. Because our outcomes of interest are constructed from percentage of words in a given feedback entry, we do not include the number of words as a covariate in our models. As an additional check, we run supplementary analyses in which we substitute feedback word count (separately for refinement and reinforcement) as the outcome variable in our specified models to examine associations between principal gender and race and feedback word count that could bias our analyses. Both principal gender and race are not statistically significant predictors of refinement or reinforcement word count.
14. The concentration of observations with a score just above 20 (25.77) for tone are entries coded as 0 for both positive emotion and negative emotion. There is not a similarly defined pattern for the concentration of entries scored as 99 for tone.

References

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.

Adler, S., Campion, M., Colquitt, A., Grubb, A., Murphy, K., Ollander-Krane, R., & Pulakos, E. D. (2016). Getting Rid of Performance Ratings: Genius or Folly? A Debate. *Industrial and Organizational Psychology*, *9*(2), 219–252. https://doi.org/10.1017/iop.2015.106

Aguinis, H., Gottfredson, R. K., & Joo, H. (2012). Delivering effective performance feedback: The strengths-based approach. *Business Horizons*, *55*(2), 105–111.

Akos, P., & Kretchmar, J. (2016). Gender and ethnic bias in letters of recommendation: Considerations for school counselors. *Professional School Counseling*, *20*(1), 102–114.

Anderson, L. M., & Turnbull, B. J. (2016). Evaluating and supporting principals. Policy Studies Associates. https://files.eric.ed.gov/fulltext/ED570471.pdf

Audia, P. G., & Locke, E. A. (2003). Benefiting from negative feedback. *Human Resource Management Review*, *13*(4), 631–646. https://doi.org/10.1016/j.hrmr.2003.11.006

Axelson RD, Solow CM, Ferguson KJ, Cohen MB. Assessing implicit gender bias in medical student performance evaluations. Eval Health Prof. 33(3):365-385. doi:10.1177/0163278710375097.

Bandura, A., & Cervone, D. (1983). Self-evaluative and self-efficacy mechanisms governing the motivational effects of goal systems. *Journal of personality and social psychology*, *45*(5), 1017.

Baron, R. A. (1993). Criticism (informal negative feedback) as a source of perceived unfairness in organizations: Effects, mechanisms, and countermeasures. In *Justice in the workplace: Approaching fairness in human resource management* (pp. 155–170). Lawrence Erlbaum Associates, Inc.

Barry, C. T., Chaplin, W. F., & Grafeman, S. J. (2006). Aggression following performance feedback: The influences of narcissism, feedback valence, and comparative standard. *Personality and Individual Differences*, *41*(1), 177-187.

Bass, Bernard M., Bruce J. Avolio, and Leanne Atwater. 1996. The transformational and transactional leadership of men and women. *Applied Psychology* 45(1):5-34.

Biernat, M., Tocci, M. J., & Williams, J. C. (2012). The language of performance evaluations: Gender-based shifts in content and consistency of judgment. Social Psychological & Personality Science, 3(2), 186–192. https://doi.org/10.1177/1948550611415693

Bludevich, B., Irby, I., Chang, H., Danielson, P. D., Gonzalez, R., Snyder, C. W., & Chandler, N. M. (2021). Letters of recommendation for pediatric surgery fellowship: analysis of linguistic differences based on gender of the applicant. *Journal of pediatric surgery*, *56*(8), 1299-1304.

Brutus, S. (2010). Words versus numbers: A theoretical exploration of giving and receiving narrative comments in performance appraisal. *Human Resource Management Review*, *20*(2), 144–157. https://doi.org/10.1016/j.hrmr.2009.06.003

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (Eds.). (2017). *A practical guide to sentiment analysis*. Cham, Switzerland: Springer International Publishing.

Campbell, S. L., and M. Ronfeldt. 2018. Observational Evaluation of Teachers: Measuring More than We Bargained For? *American Educational Research Journal 55*(6): 1233–1267. doi:10.3102/0002831218776216.

Cannon, M. D., & Witherspoon, R. (2005). Actionable feedback: Unlocking the power of learning and performance improvement. *Academy of Management Perspectives*, *19*(2), 120–134. https://doi.org/10.5465/ame.2005.16965107

Carver, C. S., & Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality–social, clinical, and health psychology. *Psychological bulletin*, *92*(1), 111.

Castilla, E. J. (2015). Accounting for the Gap: A Firm Study Manipulating Organizational Accountability and Transparency in Pay Decisions. *Organization Science*, *26*(2), 311–333. https://doi.org/10.1287/orsc.2014.0950

Chawla, N., Gabriel, A. S., Dahling, J. J., & Patel, K. (2016). Feedback Dynamics Are Critical to Improving Performance Management Systems. *Industrial and Organizational Psychology*, *9*(2), 260–266. https://doi.org/10.1017/iop.2016.8

Chung, Y. B., Marshall, J. A., & Gordon, L. L. (2001). Racial and Gender Biases in Supervisory Evaluation and Feedback. *The Clinical Supervisor*, *20*(1), 99–111. https://doi.org/10.1300/J001v20n01_08

Cleveland, J. N., Lim, A. S., & Murphy, K. R. (2007). 11 Feedback phobia? Why employees do not want to give or receive performance feedback. *Research Companion to the Dysfunctional Workplace*, 168.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W .(2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science*, *15*, 687-93.

Correll, S., & Simard, C. (2016). Vague Feedback is holding women back. Harvard Business Review. https://hbr.org/2016/04/research-vague-feedback-is-holding-women-back

Correll, S. J., Weisshaar, K. R., Wynn, A. T., & Wehner, J. D. (2020). Inside the Black Box of Organizational Life: The Gendered Language of Performance Assessment. *American Sociological Review*, *85*(6), 1022–1050. https://doi.org/10.1177/0003122420962080

DeMatthews, D. E., Scheffer, M., & Kotok, S. (2020). Useful or Useless? Principal Perceptions of the Texas Principal Evaluation and Support System. *Journal of Research on Leadership Education*, 1942775120933920. https://doi.org/10.1177/1942775120933920

Dobbin, F., Schrage, D., & Kalev, A. (2015). Rage against the Iron Cage: The Varied Effects of Bureaucratic Personnel Reforms on Diversity. *American Sociological Review*, *80*(5), 1014–1044. https://doi.org/10.1177/0003122415596416

Donaldson, M., Mavrogordato, M., Dougherty, S., Ghanem, R. A., & Youngs, P. (2020). Principal Evaluation under Elementary and Secondary Every Student Succeeds Act: A Comprehensive Policy Review. *Education Finance and Policy*, 1–15. https://doi.org/10.1162/edfp_a_00332

Donaldson, M. L., Mavrogordato, M., Youngs, P., Dougherty, S., & Al Ghanem, R. (2021). "Doing the 'Real' Work": How Superintendents' Sensemaking Shapes Principal Evaluation Policies and Practices in School Districts. *AERA Open*, *7*, 2332858420986177. https://doi.org/10.1177/2332858420986177

Drake, S., A. Auletto, and J. M. Cowen. 2019. "Grading Teachers: Race and Gender Differences in Low Evaluation Ratings and Teacher Employment Outcomes." American Educational Research Journal 56 (5): 1800–1833. doi:10.3102/0002831219835776.

Eagly, A. H., Johannesen-Schmidt, M. C., Van Engen, M. L. (2003). Transformational, transactional, and laissez-faire leadership styles: A meta-analysis comparing women and men. *Psychological Bulletin* 129(4):569-591.

Eagly, A. H., & Johnson, B. T. (1990). Gender and leadership style: A meta-analysis. *Psychological bulletin* 108(2):233.

Fedor, D. B., Davis, W. D., Maslyn, J. M., & Mathieson, K. (2001). Performance improvement efforts in response to negative feedback: The roles of source power and recipient self-esteem. *Journal of Management*, *27*(1), 79–97. https://doi.org/10.1177/014920630102700105

Fesler, L., Dee, T., Baker, R., & Evans, B. (2019). Text as Data Methods for Education Research. *Journal of Research on Educational Effectiveness*, *12*(4), 707–727. https://doi.org/10.1080/19345747.2019.1634168

Fernandez-Mateo, I., & Fernandez, R. M. (2016). Bending the pipeline? Executive search and gender inequality in hiring for top management jobs. *Management Science*, *62*(12), 3636-3655.

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2019). Text mining in education. *WIREs Data Mining and Knowledge Discovery*, *9*(6), e1332.

Festinger, L. (1957). A theory of cognitive dissonance. Evanston, IL: Row Peterson

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. https://doi.org/10.1177/1942775115614291

Gentzkow, M., Kelly, B., & Taddy, M. (2017). Text as data (NBER Working Paper No. 23276). Cambridge, MA: National Bureau of Economic Research.

Greve, H. (2003). Organizational learning from performance feedback: A behavioral perspective on innovation and change. Cambridge, UK: Cambridge University Pres

Grissom, J. A., & Bartanen, B. (2019). Principal effectiveness and principal turnover. *Education Finance and Policy*, *14*(3), 355-382.

Grissom, J. A., & Bartanen, B. (2022). Potential Race and Gender Biases in High-Stakes Teacher Observations. *Journal of Policy Analysis and Management*, *41*(1), 131-161.

Grissom, J. A., Blissett, R. S. L., & Mitani, H. (2018). Evaluating School Principals: Supervisor Ratings of Principal Practice and Principal Job Performance. *Educational Evaluation and Policy Analysis*, *40*(3), 446–472. https://doi.org/10.3102/0162373718783883

Gerull, K. M., Loe, M., Seiler, K., McAllister, J., & Salles, A. (2019). Assessing gender bias in qualitative evaluations of surgical residents. *The American Journal of Surgery*, *217*(2), 306–313. https://doi.org/10.1016/j.amjsurg.2018.09.029

Harber, K. D. (1998). Feedback to minorities: evidence of a positive bias. *Journal of personality and social psychology*, *74*(3), 622.

Harber, K. D., Gorman, J. L., Gengaro, F. P., Butisingh, S., Tsang, W., & Ouellette, R. (2012). Students' race and teachers' social support affect the positive feedback bias in public schools. *Journal of Educational Psychology*, *104*(4), 1149.

Harmon-Jones, E.; Mills, J. A introduction to cognitive dissonance theory and an overview of current perspectives on the theory. In *Cognitive Dissonance*, *2nd ed.*; Harmon-Jones, E., Ed.; American Psychological Association: Washington, DC, USA, 2019; pp. 3–24.

Heen, S., & Stone, D. (2014). Find the coaching in criticism. *Harvard Business Review*, *92*(1/2), 108-111.

Hill, J., Ottem, R., & DeRoche, J. (2016). *Trends in public and private school principal demographics and qualifications: 1987-88 to 2011-12*. Stats in Brief. NCES 2016-189. Washington, DC: National Center for Education Statistics.

Hull, D. A. (1996). Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, *47*(1), 70–84. https://doi.org/10.1002/(SICI)1097-4571(199601)47:1<70::AID-ASI7>3.0.CO;2-#

Hvidston, D. J., McKim, C. A., & Holmes, W. T. (2018). What Are Principals' Perceptions? Recommendations for Improving the Supervision and Evaluation of Principals. *NASSP Bulletin*, *102*(3), 214–227. https://doi.org/10.1177/0192636518802033

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, *64*(4), 349–371. http://dx.doi.org/10.1037/0021-9010.64.4.349

Jampol, L., Rattan, A., & Wolf, E. B. (2023). A bias toward kindness goals in performance feedback to women (vs. men). *Personality and Social Psychology Bulletin*, *49*(10), 1423-1438.

Jampol, L., & Zayas, V. (2020). Gendered White Lies: Women Are Given Inflated Performance Feedback Compared With Men. *Personality and Social Psychology Bulletin*, 0146167220916622. https://doi.org/10.1177/0146167220916622

Jiang, J. Y., and S. E. Sporte. 2016. "Teacher Evaluation in Chicago: Differences in Observation and Value-Added Scores by Teacher, Student, and School Characteristics." Research Report. University of Chicago Consortium on School Research.

Johansson, T., & Wennblom, G. (2017). In female supervisors male subordinates trust!? An experiment on supervisor and subordinate gender and the perceptions of tight control. *Journal of Management Control*, *28*(3), 321-345.

Jung, Dong I., and Bruce J. Avolio. 2000. Opening the black box: An experimental investigation of the mediating effects of trust and value congruence on transformational and transactional leadership. *Journal of Organizational Behavior* 21(8):949-964.

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. https://files.eric.ed.gov/fulltext/ED566525.pdf

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Kraft, M. A., & Gilmour, A. F. (2016). Can Principals Promote Teacher Development as Evaluators? A Case Study of Principals' Views and Experiences. *Educational Administration Quarterly*, *52*(5), 711–753. https://doi.org/10.1177/0013161X16653445

Liu, J., & Cohen, J. (2021). Measuring Teaching Practices at Scale: A Novel Application of Text-as-Data Methods. *Educational Evaluation and Policy Analysis*, 01623737211009267. https://doi.org/10.3102/01623737211009267

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415-463). Springer, Boston, MA.

Loftus, S., & Tanlu, L. J. (2018). Because of "Because": Examining the Use of Causal Language in Relative Performance Feedback. *The Accounting Review*, *93*(2), 277–297. https://doi.org/10.2308/accr-51830

Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and letters of recommendation for academia: Agentic and communal differences. The Journal of Applied Psychology, 94(6), 1591–1599. https://doi.org/10.1037/a0016539

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge, England: Cambridge University Press.

Maume, D. J. (2011). Meet the new boss… same as the old boss? Female supervisors and subordinate career prospects. *Social Science Research*, *40*(1), 287-298.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding Performance Appraisal: Social, Organizational, and Goal-Based Perspectives*. Thousand Oaks, CA: Sage Publications.

National Center for Education Statistics. (2020). Characteristics of Public School Principals. Retrieved March 31, 2021 from https://nces.ed.gov/programs/coe/pdf/coe_cls.pdf.

Nelson, J. L., Grissom, J. A., & Cameron, M. L. (2021). Performance, Process, and Interpersonal Relationships: Explaining Principals' Perceptions of Principal Evaluation. *Educational Administration Quarterly*, 0013161X211009295.

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science*, *37*(4), 375–401. https://doi.org/10.1007/s11251-008-9053-x

Park, J. A., Johnson, D. A., Moon, K., & Lee, J. (2019). The Interaction Effects of Frequency and Specificity of Feedback on Work Performance. *Journal of Organizational Behavior Management*, *39*(3–4), 164–178.

Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin.

Penner, R., Liu, Solanki, & Loeb. (2019). Differing Views of Equity: How Prospective Educators Perceive Their Role in Closing Achievement Gaps. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, *5*(3), 103. https://doi.org/10.7758/rsf.2019.5.3.06

Polanco-Santana, J. C., Storino, A., Souza-Mota, L., Gangadharan, S. P., & Kent, T. S. (2021). Ethnic/Racial Bias in Medical School Performance Evaluation of General Surgery Residency Applicants. *Journal of Surgical Education*, S1931720421000489. https://doi.org/10.1016/j.jsurg.2021.02.005

Porter, M. F. (2006). An algorithm for suffix stripping. *Program*, *40*(3), 211–218. https://doi.org/10.1108/00330330610681286

Ross, D. A., Boatright, D., Nunez-Smith, M., Jordan, A., Chekroud, A., & Moore, E. Z. (2017). Differences in words used to describe racial and gender groups in Medical Student Performance Evaluations. *PLOS ONE*, *12*(8), e0181659. https://doi.org/10.1371/journal.pone.0181659

Ruscher, J. B., Wallace, D. L., Walker, K. M., & Bell, L. H. (2010). Constructive feedback in cross-race interactions. *Group Processes & Intergroup Relations*, *13*(5), 603–619. https://doi.org/10.1177/1368430210364629

Schmader, T., Whitehead, J., & Wysocki, V. H. (2007). A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. Sex Roles, 57(7–8), 509–514. https://doi.org/10.1007/s11199-007-9291-4

Smith, D. G., Rosenstein, J. E., Nikolov, M. C., & Chaney, D. A. (2019). The Power of Language: Gender, Status, and Agency in Performance Evaluations. *Sex Roles*, *80*(3–4), 159–171. https://doi.org/10.1007/s11199-018-0923-7

Sprick, R., Knight, J., Reinke, W., Skyles, T., & Barnes, L. (2010). *Coaching classroom management: Strategies and tools for administrators and coaches.* Eugene, OR: Pacific Northwest Publishing.

Stone, D., Heen, S., & Patton, B. (2010). *Difficult conversations: How to discuss what matters most*. Penguin.

Sun, M., Liu, J., Zhu, J., & LeClair, Z. (2019). Using a Text-as-Data Approach to Understand Reform Processes: A Deep Exploration of School Improvement Strategies. *Educational Evaluation and Policy Analysis*, *41*(4), 510–536. https://doi.org/10.3102/0162373719869318

Taie, S., and Goldring, R. (2017). Characteristics of Public Elementary and Secondary School Principals in the United States: Results From the 2015–16 National Teacher and Principal Survey First Look (NCES 2017-070). U.S. Department of Education. Washington, DC:

National Center for Education Statistics. Retrieved March 31, 2021 from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2017070.

Tennessee Board of Education. *Teacher and Principal Evaluation Policy*. 5.201 § (2013).

Tennessee Department of Education (2016). *Administrator Evaluation Rubric.* Available here: https://team-tn.org/wp-content/uploads/2013/08/TEAM-Admin-Evaluation-Rubric-20161.pdf.

Trix, F., & Psenka, C. (2003). Exploring the Color of Glass: Letters of Recommendation for Female and Male Medical Faculty. *Discourse & Society*, *14*(2), 191–220. https://doi.org/10.1177/0957926503014002277

United States Government Accountability Office. (2013). *Race to the Top: States implementing teacher and principal evaluation systems despite challenges.* Washington, D.C.: Author. Retrieved from: https://www.gao.gov/assets/660/657936.pdf

Weick, K. E. 1995. *Sensemaking in Organizations*. Thousand oaks, CA: Sage Publications.

Williams, C. L., Muller, C., & Kilanski, K. (2012). Gendered Organizations in the New Economy. *Gender & Society*, *26*(4), 549–573. https://doi.org/10.1177/0891243212445466

Wilson, K. Y. (2010). An analysis of bias in supervisor narrative comments in performance appraisal. *Human Relations*, *63*(12), 1903–1933. https://doi.org/10.1177/0018726710369396

Tables

Table 1. Sample summary statistics

	Mean	SD	Min	Max	N
Refinement words	36.47	34.01	1	564	7291
Reinforcement words	35.54	32.99	1	454	7291
Principal Characteristics					
Female	0.56	0.5			7250
White	0.86	0.34			7292
Black	0.11	0.33			7292
Age	49.22	8.74	28	79	7201
Years of experience	5.61	4.54	0	17	7292
More than Masters	0.15	0.35			7281
Education Specialist	0.35	0.48			7281
Doctorate	0.15	0.36			7281
Average performance rating	3.88	0.56	1.86	5	7246
Number of observations	1.99	0.32	1	4	7292
Evaluator Characteristics					
Female	0.51	0.5			6631
White	0.88	0.32			6935
Black	0.09	0.28			6935
Age	52.93	8.36	32	72	5923
Years of experience	26.72	9.83	1	48	6960
Years evaluating	2.16	1.33	0	6	7,292
More than Masters	0.12	0.33			6950
Education Specialist	0.33	0.47			6950
Doctorate	0.30	0.46			6950
Superintendent	0.40	0.49			6913
Assistant Superintendent	0.12	0.32			6913
Supervisor of Instruction	0.29	0.45			6913
Other Central Office	0.12	0.4			6913
Principal/School Staff	0.07	0.26			6913
Span of control (# principals)	11.74	10.31	1	46	7292
School Characteristics					
Enrollment (x100)	6.15	3.87	0.02	24.95	7195
Elementary	0.39	0.49			7195
Middle	0.28	0.45			7195
High	0.2	0.4			7195
Mixed	0.13	0.34			7195
Proportion students Black	0.17	0.23	0	1	7195
Proportion students Hispanic	0.07	0.09	0	0.75	7195
Proportion students FRPL	0.59	0.23	0	1	7195
Std. achievement score	-0.02	0.38	-2.11	2.34	6897
Urban	0.24	0.43			7211
Suburban	0.22	0.42			7211
Town	0.19	0.39			7211
Rural	0.34	0.47			7211

Source: Authors' calculations based on Tennessee administrative data pooled across years 2014-15 through 2016-17.

Table 2. Descriptive statistics for measures of feedback affective language

	Mean	SD	Min	Max
Refinement				
Tone	67.51	33.29	1	99
Positive Emotion	4.40	4.78	0	50
Negative Emotion	0.22	1.02	0	18.18
Reinforcement				
Tone	74.93	31.65	1	99
Positive Emotion	6.01	6.06	0	66.67
Negative Emotion	0.15	0.72	0	16.67

Source: Authors' calculations based on measures generated from the LIWC dictionary using Tennessee administrative data pooled across years 2014-15 through 2016-17.

Table 3. T-tests for mean differences in feedback affective language by principal gender and race

	(1) Female	(2) Male	(3) Mean Difference	(4) Black	(5) White	(6) Mean Difference
Panel A: Refinement						
Tone	68.11	66.74	1.37**	65.26	67.8	-2.54**
Positive Emotion	4.43	4.32	0.11	3.85	4.48	-0.63***
Negative Emotion	0.23	0.22	0.01	0.24	0.22	0.02
Panel B: Reinforcement						
Tone	75.14	74.96	0.18	70.86	75.59	-4.73***
Positive Emotion	5.86	6.23	-0.37***	5.07	6.16	-1.09***
Negative Emotion	0.14	0.16	-0.02	0.15	0.14	0.01

Source: Authors' calculations based on measures generated from the LIWC dictionary using Tennessee administrative data pooled across years 2014-15 through 2016-17.
+p < .10. **p < .05. ***p < .01.

Table 4. Regression-adjusted gender and racial differences in feedback affective language

	(1) Tone	(2) Positive Emotion	(3) Negative Emotion
Panel A: Refinement			
Female	-1.12	0.08	0.03
	(1.08)	(0.16)	(0.03)
Black	-3.09+	-0.52+	-0.06
	(1.84)	(0.29)	(0.04)
Constant	76.01***	4.75***	-0.26
	(6.09)	(0.97)	(0.21)
Obs	6,451	6,451	6,451
R^2	0.06	0.04	0.04
Panel B: Reinforcement			
Female	-2.52***	-0.31+	-0.02
	(0.93)	(0.19)	(0.02)
Black	1.16	0.02	-0.03
	(1.81)	(0.34)	(0.03)
Constant	77.38***	6.07***	0.14
	(5.69)	(1.29)	(0.13)
Obs	6,451	6,451	6,451
R^2	0.07	0.05	0.02
Principal Covariates	Y	Y	Y
Evaluator Covariates	Y	Y	Y
School Covariates	Y	Y	Y
District FE*	Y	Y	Y
Year FE	Y	Y	Y

Notes: Standard errors clustered at the district level. Standard errors in parentheses. Principal covariates include age, years of experience as principal, education, average performance rating, and number of observations. Evaluator covariates include gender, age, race, years of experience as evaluator, years of experience in education, education, role, and span of control. School covariates include student enrollment, school level, proportion of Black students, proportion of Hispanic students, proportion of students eligible for FRPL, and standardized achievement score. We do not control for school locale as it is collinear with the district fixed effect.
+p < .10. **p < .05. ***p < .01.

Table 5. Principal-evaluator gender interactions

	(1)	(2)	(3)
	Tone	Positive Emotion	Negative Emotion
Panel A: Refinement			
Female Principal	-1.27	0.06	0.01
	(1.83)	(0.25)	(0.04)
Female Evaluator	-0.25	-0.16	-0.03
	(2.24)	(0.31)	(0.07)
Female Principal*Female Evaluator	0.29	0.03	0.02
	(2.17)	(0.30)	(0.05)
Constant	76.11***	4.76***	-0.25
	(5.99)	(0.97)	(0.21)
Obs	6,451	6,451	6,451
R^2	0.06	0.04	0.04
Panel B: Reinforcement			
Female Principal	-1.98	-0.20	-0.04
	(1.54)	(0.33)	(0.03)
Female Evaluator	3.59**	0.35	-0.03
	(1.39)	(0.26)	(0.04)
Female Principal*Female Evaluator	-1.07	-0.23	0.04
	(1.87)	(0.39)	(0.04)
Constant	77.02***	5.99***	0.16
	(5.77)	(1.29)	(0.13)
Obs	6,451	6,451	6,451
R^2	0.07	0.05	0.02
Principal Covariates	Y	Y	Y
Evaluator Covariates	Y	Y	Y
School Covariates	Y	Y	Y
District FE	Y	Y	Y
Year FE	Y	Y	Y

Notes: Standard errors clustered at the district level. Standard errors in parentheses. Principal covariates include age, years of experience as principal, education, average performance rating, and number of observations. Evaluator covariates include age, race, years of experience as evaluator, years of experience in education, education, role, and span of control. School covariates include student enrollment, school level, proportion of Black students, proportion of Hispanic students, proportion of students eligible for FRPL, and standardized achievement score. We do not control for school locale as it is collinear with the district fixed effect.
+p < .10. **p < .05. ***p < .01.

Table 6. Principal-evaluator race interactions

	(1) Tone	(2) Positive Emotion	(3) Negative Emotion
Panel A: Refinement			
Black Principal	-3.20	-0.54	-0.06
	(2.24)	(0.34)	(0.05)
Black Evaluator	0.97	-0.04	-0.08
	(2.52)	(0.35)	(0.08)
Black Principal*Black Evaluator	0.93	0.08	-0.06
	(3.93)	(0.45)	(0.09)
Constant	77.31**	4.90**	-0.30
	(6.36)	(0.95)	(0.24)
Obs	6,252	6,252	6,252
R^2	0.06	0.04	0.04
Panel B: Reinforcement			
Black Principal	1.15	-0.05	-0.01
	(2.42)	(0.45)	(0.04)
Black Evaluator	2.36	0.63	-0.01
	(1.79)	(0.43)	(0.05)
Black Principal*Black Evaluator	-2.27	-0.10	-0.09
	(4.91)	(0.90)	(0.07)
Constant	77.55**	6.24**	0.18
	(6.17)	(1.32)	(0.13)
Obs	6,252	6,252	6,252
R^2	0.07	0.05	0.03
Principal Covariates	Y	Y	Y
Evaluator Covariates	Y	Y	Y
School Covariates	Y	Y	Y
District FE	Y	Y	Y
Year FE	Y	Y	Y

Notes: Standard errors clustered at the district level. Standard errors in parentheses. Principal covariates include years of experience as principal, education, average performance rating, and number of observations. Evaluator covariates include race, years of experience as evaluator, education, role, and span of control. School covariates include student enrollment, school level, proportion of Black students, proportion of Hispanic students, proportion of students eligible for FRPL, and standardized achievement score. We do not control for school locale as it is collinear with the district fixed effect.
+p < .10. **p < .05. ***p < .01.

Figures

Figure 1. Distribution of feedback word count

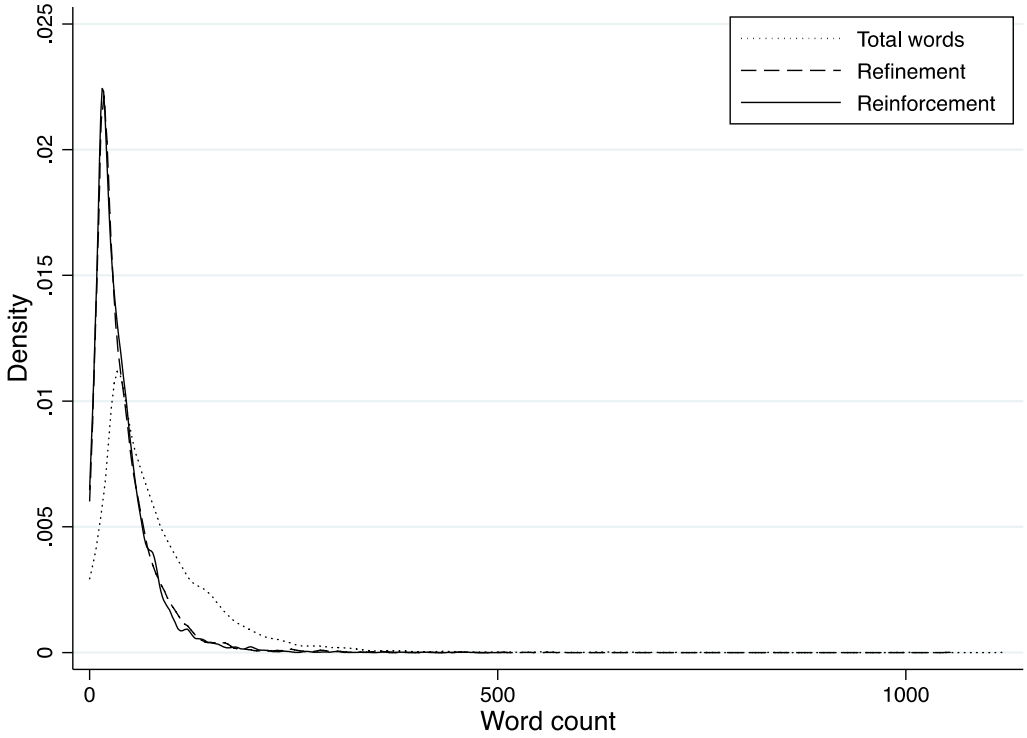

Figure 2. Distribution of affective language measures