



The Valence of Teacher Performance Feedback and Its Consequences: Examining a Critical Mechanism of Reformed Teacher Evaluation Systems

Seth B. Hunter
George Mason University

Matthew P. Steinberg
Accelerate

Districts nationwide have increased the frequency of teacher evaluations. Yet, we know little about the role of evaluator feedback for teacher improvement. Using unique classroom observation-level data, we use evaluator ratings and teacher self-assessments of teacher performance to rigorously examine (positive and negative) feedback valence from the teacher's perspective and its consequences. While teachers and evaluators similarly rate performance, there's significant variability in teacher-evaluator ratings. Teacher performance improves across multiple within-year classroom observations, though evaluator ratings likely overstate improvements among the lowest-performing teachers. While negative feedback from evaluators likely improves within-year teacher performance and may improve their productivity regarding student achievement, statistically insignificant yet practically meaningful evidence suggests it may also push teachers toward schools with more positive feedback.

VERSION: November 2024

Suggested citation: Hunter, Seth B., and Matthew P. Steinberg. (2024). The Valence of Teacher Performance Feedback and Its Consequences: Examining a Critical Mechanism of Reformed Teacher Evaluation Systems. (EdWorkingPaper: 22 -676). Retrieved from Annenberg Institute at Brown University: <https://doi.org/10.26300/97k9-br18>

The Valence of Teacher Performance Feedback and Its Consequences: Examining a Critical Mechanism of Reformed Teacher Evaluation Systems

Seth B. Hunter

George Mason University

Matthew P. Steinberg

Accelerate

November 20, 2024

The authors thank their partners from the unnamed school district for helpful critical and positive feedback, conference participants at the Association for Education Finance and Policy (AEFP) annual conference, participants at the EdPolicy*Forward* Research Workshop at George Mason University, Dan Goldhaber, and Sarah Woulfin. Seth B. Hunter is an associate professor of education at George Mason University (shunte@gmu.edu); Matthew P. Steinberg is the managing director of research and evaluation at Accelerate (matthew@matthewpsteinberg.com).

Abstract

Districts nationwide have increased the frequency of teacher evaluations. Yet, we know little about the role of evaluator feedback for teacher improvement. Using unique classroom observation-level data, we use evaluator ratings and teacher self-assessments of teacher performance to rigorously examine (positive and negative) feedback valence from the teacher's perspective and its consequences. While teachers and evaluators similarly rate performance, there's significant variability in teacher-evaluator ratings. Teacher performance improves across multiple within-year classroom observations, though evaluator ratings likely overstate improvements among the lowest-performing teachers. While negative feedback from evaluators likely improves within-year teacher performance and may improve their productivity regarding student achievement, statistically insignificant yet practically meaningful evidence suggests it may also push teachers toward schools with more positive feedback.

Keywords: Education policy; evaluation; school/teacher effectiveness; supervision; regression analyses; econometric analyses

Introduction

Since the late 2000s, education agencies nationwide have reformed teacher evaluation systems to improve student outcomes. Despite the millions of dollars invested in these systems (Chambers et al., 2013; Stecher et al., 2018), rigorous studies evaluating the impact of their adoption do not detect average effects on student achievement nationwide or across multiple districts (Bleiberg et al., In Press; Hunter & Bowser, 2024; Song et al., 2021), although two districtwide studies detect subject-specific average effects in math or reading (Steinberg & Sartain, 2015; Taylor & Tyler, 2012).ⁱ The lack of detected average effects for these costly at-scale reforms raises essential questions about mechanisms. Inferentially rigorous studies have investigated accountability-driven mechanisms, such as performance-based teacher dismissals or bonuses (Cullen et al., 2021; Dee et al., 2021; Dee & Wyckoff, 2015; Pham et al., 2021), and developmentally focused mechanisms, including policy-assigned classroom observations that allow school administrators to assess and provide feedback to teachers about their performance (de Barros, 2019; Hunter & Kho, 2023; Kraft & Christian, 2021; Song et al., 2021; Youngs et al., 2020). We extend this literature and address the need to better understand teacher evaluation by examining a potentially significant driver of improvements in teaching and teacher effectiveness: a unique feature of feedback teachers receive after each formal classroom observation. Notably, teachers experienced more opportunities to receive such feedback in reformed systems since most increased the frequency of classroom observations and post-observation feedback conferences (Steinberg & Donaldson, 2016).

There are different *feedback* types, each with a rich set of features; we examine a feature of *performance feedback* called *valence*. We follow prior work across disciplines to define and operationalize these terms. *Feedback* is information provided to an individual (i.e., teacher)

about their past behavior or performance to modify their thinking or behavior to maintain performance among the highest performers or improve task performance for others (Church et al., 2019; Cianci et al., 2010; Fong et al., 2019; Goff et al., 2014; Goldring et al., 2015; Kluger & DeNisi, 1996; Vohra & Singh, 2005). *Performance feedback* is information communicated to an individual (i.e., teacher) about their current performance level compared to an expected benchmark; it plays a critical role in human capital and resource management by identifying discrepancies between actual and expected performance and offering encouragement or corrective action to maintain or enhance future performance through increased effort and learning (Holderness et al., 2017; Kraft & Christian, 2021). *Feedback valence* refers to how individual (i.e., teacher) performance compares to a relevant benchmark and is typically classified as positive or negative; positive feedback indicates that performance was higher than expected, while negative feedback suggests that performance was lower than expected (Holderness et al., 2017; Raaijmakers et al., 2017). Valence induces neurological, psychological, and behavioral responses that can affect performance (Bogard et al., 2020; Lechermeier & Fassnacht, 2018; Raaijmakers et al., 2017).

Our study of valence and its consequences makes four contributions. Following Raaijmakers and colleagues (2017), it is among the first to operationalize feedback valence rigorously; we use observation-specific assessments of teacher performance by school administrators and teachers, affording new understandings about feedback. Like nearly every district in the United States, teachers in our setting are evaluated by school administrators during at least one formal classroom observation per year. Unlike any other evaluation setting (to our knowledge), the study district also expects teachers to submit a self-assessment of their performance after each observation. We subtract the mean indicator score a teacher gives their

performance for observation k in year t based on a standards-based teaching rubric from the mean indicator score their formal evaluator gives for the same observation using the same rubric. We view evaluator scores and teacher self-assessment scores as *feedback* sources. In this study, we focus on an operationalization of *performance feedback* in which the “expected benchmark” is teachers’ self-assessments of their teaching after an observation, and evaluator scores for the same observation represent the teacher’s “performance level.” Our operationalization allows us to rigorously investigate positive and negative feedback *valence* from the teacher’s perspective (i.e., self-assessment scores), a rare affordance as most valence studies do not incorporate input from the feedback recipient and those that do typically rely on survey responses (Bogard et al., 2020; Cherasaro et al., 2016; Cutumisu & Schwartz, 2016; Hunter & Springer, 2022; Raaijmakers et al., 2017). Second, our valence measure also allows us to explore valence *intensity* or the extent to which evaluator assessments exceeded or fell short of teacher self-assessments (i.e., the magnitude of the score difference). Third, we rigorously examine the implications of *extensive valence* (i.e., whether feedback is negative) for teacher performance, productivity, and retention; moreover, ours is among the first studies across disciplines to examine relationships with labor market outcomes. Fourth, this study is the first to rigorously examine feedback valence in reformed teacher evaluation systems.

We address the following questions: (1) Do teachers and evaluators similarly rate teacher performance? (2) What are the consequences of negative feedback valence on teacher performance, productivity, and mobility? Negative feedback activates neurological responses, leading to negative emotions that recipients (i.e., teachers) typically want to avoid (Carver & Scheier, 1982; Kim et al., 2018; Lechermeier & Fassnacht, 2018; Wiswede et al., 2009). Whether avoidance prompts (1) disregard and little to no performance improvement, (2)

motivation to improve or learning resulting in improvement, or (3) flight from negative feedback providers depends on several organizational and individual characteristics (Bogard et al., 2020; Lechermeier & Fassnacht, 2018). Our data permits us to examine valence and its consequences independent of the most concerning sources of heterogeneity.

Relying on unique observation-level ratings of teacher performance from teachers and their evaluators from a large urban district in the South, we first examine the magnitude, distribution, and *within-year* patterns of classroom observation scores as these determine valence. Little work examines such within-year patterns. We then examine if negative feedback valence from classroom observation scores affects teachers' subsequent performance, productivity, and mobility. We find that, on average and modally, teachers and their evaluators rate teacher performance similarly at the observation level. However, there is significant variability in valence intensity. We further find that teacher performance improves across multiple within-year classroom observations, reflected by ratings from teachers and evaluators. However, while the rates of growth in teacher self-assessments are homogenous across teachers receiving different numbers of annual classroom observations, evaluator scores exhibit higher growth rates for teachers receiving more annual observations, suggesting that the lowest-performing teachers (i.e., those who receive the most annual observations) experience increasingly positive feedback valence. We also show that evaluators likely inflate the fourth (and final) observation score level for the lowest-performing teachers, resulting in a high dose of positive feedback near the end of school. Finally, when teachers receive negative feedback valence, their performance – as measured by evaluator ratings during subsequent classroom observations – improves, and these improvements are reflected in a teacher's contribution to student achievement growth. Simultaneously, teachers who receive negative feedback may also

seek alternative teaching assignments in schools where administrators provide positive feedback on average; however, we urge caution when interpreting these mobility patterns as they are based on imprecise estimates. Ultimately, the teacher evaluation mechanism examined strongly suggests that performance feedback from evaluation system measures can support the developmental goals of teacher evaluation systems.

Background and Related Literature

Performance feedback is defined and operationalized in various ways across disciplines. In organizational psychology, it is seen as information intended to modify thinking or behavior to improve task performance (Kraft & Christian, 2021). In management control systems, it involves comparing performance to a benchmark to identify discrepancies and encourage improvement (Holderness et al., 2017). Despite differences across disciplines, there is a consensus that performance feedback is a tool for behavioral modification and performance improvement, core goals of reformed teacher evaluation systems (Almy, 2011; Church et al., 2019; Donaldson, 2021; Holderness et al., 2017; Hunter & Springer, 2022; Kluger & DeNisi, 1996; Kraft & Christian, 2021).

Valence in feedback refers to whether the feedback is positive or negative and can significantly impact recipient performance via neural, psychological, and behavioral mechanisms. Positive feedback is associated with increased activation in brain regions linked to reward processing, which individuals are strongly motivated to pursue; therefore, positive feedback can reinforce target behaviors (Kim et al., 2018). Conversely, negative feedback triggers neural areas that process discomfort (Kim et al., 2018). Individuals can avoid such aversive stimuli by reaching target behaviors (improving performance), disregarding negative feedback or its implications for performance improvement (unchanged or worse performance), or

fleeing negative feedback providers (switching schools) (Audia & Locke, 2003; Goldring et al., 2015; Vohra & Singh, 2005).

Emotionally, positive feedback tends to enhance self-esteem and pride, which can also reinforce target behaviors (London & Smither, 2002). Positive feedback recipients typically perceive it as more accurate and useful than negative feedback, increasing perceptions of message quality, usefulness, and fairness (Goff et al., 2014; Jawahar, 2007; Lechermeier & Fassnacht, 2018; Raaijmakers et al., 2017). That teachers in reformed teacher evaluation systems repeatedly report that their feedback is useful and accurate may suggest a prevalence of positive feedback (Cherasaro et al., 2016; Hunter, 2022, 2024). When coupled with positive feedback, performance-based rewards and setting specific performance goals can amplify its effectiveness (Cianci et al., 2010; Ilgen et al., 1979). However, most current teacher evaluation systems eschew performance-based rewards and research examining the descriptive feedback early-career teachers receive in Tennessee suggests that most feedback episodes exclude setting specific goals, suggesting positive feedback in reformed systems may not be effective (Hunter & Springer, 2022; National Council on Teacher Quality, 2019). Moreover, a meta-analysis of rigorous research concerning feedback interventions suggests that positive feedback may not improve performance (Kluger & DeNisi, 1996).

Other work finds that positive feedback can lead to worse performance under certain conditions (Lechermeier & Fassnacht, 2018). Individuals who repeatedly receive positive feedback about tasks requiring little marginal effort can become complacent or reduce their motivation to improve, leading to no improvement or worse performance over time (Cianci et al., 2010; Ilgen et al., 1979). These conditions may exist in teacher evaluation systems as scholars consistently find that evaluators rate teacher performance highly, with little variability in the

ratings that teachers receive (Grissom & Loeb, 2017; Kraft & Gilmour, 2016b; Weisberg et al., 2009). Further K12 evidence indicates that evaluators, most of whom are principals, may issue high teacher performance ratings to avoid what they perceive to be the onerous process of dismissing low performers or the time and effort needed to create and monitor teacher improvement plans (Kraft & Gilmour, 2016a; Rodriguez & Hunter, 2021). Consequently, evaluator ratings may purposefully exceed teacher self-assessments to avoid conflicts with teachers that follow the issuance of low scores (Donaldson, 2021; Halverson et al., 2004; Kraft & Gilmour, 2016a).

Negative feedback can lead to emotions like frustration, anger, and disappointment and is perceived by recipients as less accurate and diminishing self-image (Goff et al., 2014; Jawahar, 2007; Lechermeier & Fassnacht, 2018; Raaijmakers et al., 2017). Indeed, consistent negative feedback may prompt recipients to leave their organization (school) to seek positive feedback elsewhere (Alicke & Sedikides, 2009; Church et al., 2019; Kluger & DeNisi, 1996; Vohra & Singh, 2005). However, in some contexts, these negative emotions can motivate individuals to increase effort, focus on goal attainment, and adjust their strategies for goal attainment (Goldring et al., 2015; Kluger & DeNisi, 1996; Lechermeier & Fassnacht, 2018; Raaijmakers et al., 2017). Negative feedback provides evaluators with opportunities to point teachers to specific professional learning opportunities, such as peer coaching or workshops, to improve performance (Hunter & Springer, 2022; Kraft & Christian, 2021). Additionally, research suggests that negative feedback provided by credible evaluators soon after the observed performance and received by individuals with relatively high self-efficacy can improve performance (Ilgen et al., 1979; Jawahar, 2006; Lechermeier & Fassnacht, 2018). Furthermore, individuals are more likely to increase effort rather than lower standards when performance goals

are clear and help individuals identify errors and avoid problems, thereby enhancing neural rewards processing, psychological self-esteem, and performance (Hunter & Springer, 2022; Kluger & DeNisi, 1996; Van Dijk & Kluger, 2011).

Teachers in reformed evaluation systems report that their evaluators are credible and that they can use post-observation feedback to improve performance (Cherasaro et al., 2016; Hunter, 2024), suggesting that negative feedback might improve performance. Furthermore, policies in our study setting require that teachers receive post-observation feedback within one week of observation, ensuring that their feedback is timely. However, research examining the qualitative feedback received by hundreds of early-career teachers suggests that many receive feedback that does not set clear goals, provide evidence of individual errors, or provide actionable next steps teachers can follow to avoid future performance problems (Hunter & Springer, 2022).

Ultimately, the effects of feedback valence on teacher performance, productivity, and mobility are unclear *a priori*. Prior work finds that positive and negative valence can improve performance under certain conditions, and reformed teacher evaluation systems exhibit conditions that could enable or mitigate such effects.

Study Setting

The setting for this study is a large urban school district located in the Southern United States. We rely on unique administrative data on the performance evaluations of educators collected by the district's central office. Annually, teachers in this district are evaluated during at least one formal classroom observation (hereafter referred to as *observations*). The number of annual observations a teacher receives is based on their prior-year compositive effectiveness score and current-year certification status (i.e., whether the teacher has taught for less than four years). Notably, the district expects teachers to submit a self-assessment of their performance

after each observation and input their self-assessment scores into the district's central data management system prior to the post-observation conference with their evaluators. During these post-observation conferences, the district expects evaluators and teachers to refer to and discuss both the evaluator scores of the teacher's performance as well as the teacher's self-assessment scores.

Classroom Observation System

Beginning in the 2011-12 school year, the district implemented a revised classroom observation system for teacher evaluation. However, the requirement that teachers submit a self-assessment of their instructional performance did not begin until the 2016-17 school year, and teacher self-assessment scores were not recorded in the district's central data system until the 2017-18 school year. Thus, this analysis relies on teacher self-assessment data from the 2017-18 and 2018-19 school years. To evaluate teacher performance, evaluators and teachers rely on a classroom observation rubric resembling Danielson's Framework for Teaching, which includes three evaluation domains: instruction (12 rubric items, or indicators); classroom environment (four indicators); and planning (three indicators) (see Appendix A for more detail on the classroom observation rubric). Each indicator describes specific aspects of standards-based teaching that are mapped onto three proficiency levels: below expectations (=1); at expectations (=3); and above expectations (=5). If evaluators believe the preponderance of evidence is between a 1 and 3, they are advised to issue a 2; similar logic applies to scores of 4. While low observation scores may affect myriad teacher or evaluator behaviors (Adnot et al., 2017; Grissom et al., 2017), they do not trigger formal policy consequences in the study context (e.g., tenure revocation, dismissal).

Training, Certification, and Accountability. Evaluators receive two days of training on the use of the evaluation rubric, facilitating post-observation conferences with teachers, basic knowledge of state evaluation policy, and evaluation-informed teacher improvement. This training culminates in a certification exam participants must pass to conduct classroom observations. Certified evaluators need not be principals or assistant principals; however, less than 20% of study district evaluators are not school-based (i.e., district central office personnel). State policy holds evaluators accountable in two ways. First, teachers can file formal grievances if evaluators do not follow policy expectations (e.g., if teachers do not receive a copy of their observation scores). Second, the state evaluation system also assesses school administrators' teacher evaluation and professional learning skills (see Grissom et al., 2018 for further explanation of the evaluator evaluation system). Although teachers do not participate in formal evaluator training, the study district encourages schools to hold norming sessions during which evaluators and teachers discuss the meaning of the performance rubric to develop a common understanding about how to use it when assessing teacher performance.

Ratings Process. State policy dictates that teachers are annually assigned between one classroom observation – to teachers receiving the highest prior-year composite effectiveness score – and four classroom observations – to teachers with the lowest prior-year composite effectiveness score.ⁱⁱ Teachers in the middle effectiveness categories are assigned four or two observations depending on their certification status, which is effectively determined by years of experience. Although state policy expects the typical classroom observation to last approximately 15 minutes, evidence suggests that observations typically last 30 minutes (Hunter, 2020).

School administrators decide which evaluators observe which teachers and how many different evaluators will rate the same teacher in schools with multiple evaluators; prior evidence indicates that these assignments are nonrandom and strategic (Hunter & Rodriguez, 2021). Hunter and Rodriguez (2021) find that in schools with multiple evaluators, those spending less time per observation conduct more observations than their less efficient peers; evaluators with more years of experience also conduct more observations than their peers.

Evaluators may either announce the timing of classroom observations to teachers in advance or may decide to observe a teacher's classroom instruction unannounced. Although a structured, face-to-face post-observation conference follows every classroom observation, pre-observation conferences only occur for announced observations. Moreover, state policy dictates that post-observation conferences should occur less than one week after the observation.

The timing of score entry into the district's central data management system is clear for teachers but ambiguous for evaluators. Teachers enter self-assessment scores into the data management system after the observation but prior to their post-observation conference with their evaluators. Evaluators can record their observation scores of teachers at any time during or after the observation but must enter their scores prior to the post-observation conference. This is because district leaders expect evaluators and teachers to discuss both teacher self-assessment and evaluator scores during the post-observation conference. Although teachers do not see evaluator scores prior to the post-observation conference, evaluators can review teacher self-assessment scores (via the data management system) before recording their evaluation scores. Thus, teacher self-assessments might influence evaluator scores; in such cases, we assume that evaluators will inflate their evaluation scores, an assumption that is consistent with prior evidence that evaluators prefer to avoid conflicts with teachers during post-observation

conferences (Kraft & Gilmour, 2016a). Finally, during post-observation conferences, evaluators are expected to discuss with teachers their instructional strengths, at least one area for improvement, and plans for improvement. Evaluators might help teachers address an area for improvement directly via feedback or indirectly by pointing teachers to appropriate professional learning opportunities. State and district policies expect evaluators to set improvement timelines and rubric-aligned performance goals with their teachers.

Data & Sample

We rely on administrative data from the 2017-18 and 2018-19 school years. We also incorporate teacher school assignment data from the beginning of the 2019-20 school year for analyses of teacher mobility. From the administrative data, we construct three analytic samples. The first sample is a teacher-observation-year panel; at this level, we link all evaluated teachers in grades K–12 to their evaluators, observation dates, teacher self-assessment scores, and evaluator scores (*full sample*).ⁱⁱⁱ The second sample is at the teacher-year level and includes teacher and evaluator race/ethnicity, gender, education level, years of experience, and summative observation and effectiveness scores (*teacher-by-year sample*).^{iv} The second sample also includes K-12 teachers and, when relying on prior-year measures, excludes first-year teachers. The third sample (*VAM sample*) is restricted to math or reading/English teachers who receive a state-issued value-added measure (VAM) (i.e., teachers in tested grades/subjects). We link these math and reading/English teachers to the following classroom characteristics: the proportion of a teacher's students who are female, economically disadvantaged, and nonwhite, and the number of office referrals received by the teacher's average student. We also obtain the prior-year standardized math and reading achievement scores of students in grades 4 – 8 and prior-year standardized algebra I and II and English I – III end-of-course achievement scores for high

school students. We calculate the student-level average of each student's prior-year math and reading (or algebra and English) scores (which are standardized at the subject-grade-year level), and then aggregate these student-level means to the classroom level to construct a composite measure of the incoming (i.e., prior-year) academic achievement of a teacher's students.

Sample

Table 1 summarizes the demographic and performance characteristics of teachers in our full analytic sample, both overall and by the count of annual classroom observations. The sample includes 5,251 unique K-12 teachers, 9,070 teacher-year observations and 20,045 teacher-year-observation occurrences. On average, teachers receiving more annual observations have fewer years of teaching experience and are less likely to hold an advanced degree (see Panel A). Moreover, teachers receiving more annual observations receive lower observation scores from their evaluators, on average, than teachers who receive fewer annual observations. Similarly, among teachers who receive a state-issued VAM score, those who receive fewer annual observations are more effective (as measured by student achievement growth) than their teacher peers who receive three or four annual observations. These patterns are consistent with the fact that teachers who are less experienced, on average, receive lower performance scores than their more experienced colleagues and more annual observations by school-based evaluators.

[Insert Table 1 about here]

Magnitude, Distribution, and Within-Year Patterns of Classroom Observation Scores

We construct a measure of intensive valence from the teacher's perspective using teacher self-assessments and evaluator scores. Specifically, we define $Score_{jkt}^{evaluator}$ as the evaluator's rating of teacher j during classroom observation k in school year t and $Score_{jkt}^{teacher}$ as teacher j 's

self-assessment of classroom observation k in school year t . We define intensive valence as follows:

$$(1) \textit{Valence}_{jkt} = \textit{Score}_{jkt}^{\textit{evaluator}} - \textit{Score}_{jkt}^{\textit{teacher}}$$

When an evaluator rates a teacher's performance as high (or higher) than the teacher's self-assessment, $\textit{Valence}_{jkt} \geq 0$ and feedback valence is positive. Alternatively, we characterize $\textit{Valence}_{jkt} < 0$ as negative feedback valence.^v

Figure 1 shows the magnitude and distribution of teacher and evaluator observation scores. Across all 20,045 teacher-year-observation occurrences, the mean (standard deviation) teacher self-assessment score is 3.79 (0.62). Notably, the typical evaluator score is nearly identical to the typical teacher self-assessment score, with a mean (standard deviation) of 3.73 (0.64). Thus, while mean intensive valence is modest in magnitude (0.06), indicating that the typical observation is rated similarly by both teachers and their evaluators, we also observe significant variability in intensive valence (SD=0.55).

<Figure 1 about here>

We further illustrate the relationships between teacher and evaluation scores and valence in Figure 2. The top-left panel illustrates the relationship between teacher self-assessments and evaluator ratings; the dotted line represents identical scores, and the solid line is a lowess curve. The data and lowess curve reveal a positive correlation between teacher self-assessment and evaluator scores, though this relationship also demonstrates notable heteroskedasticity, with greater variation in teacher self-assessments occurring at higher evaluator ratings. The intensive valence analyses in the lower panels illuminate the mechanical relationships between observation scores and valence; in these panels, the dotted line is plotted for $\textit{Valence} = 0$ while the solid line represents lowess curves. Low (high) evaluator scores tend to yield greater magnitudes of

negative (positive) intensive valence because high (low) evaluator scores produce a floor (ceiling). Despite the positive mechanical relationship between evaluator scores and intensive valence, the substantial variation in teacher self-assessments in the bottom-left of Figure 2 shows that low evaluator scores can still yield positive valence ($Valence \geq 0$) and high evaluator scores can yield negative valence ($Valence < 0$). Analogous patterns exist between teacher self-assessments and intensive valence in the bottom-right panel of Figure 2.

< Figure 2 about here >

Distributions and Within-Year Patterns

To better understand the underlying patterns of intensive valence, Figure 3 presents the within-year distribution of mean teacher self-assessment and evaluator scores by the total number of observations teacher j was subject to in school year t . Teacher and evaluator scores follow similar patterns across the first three annual observation occurrences. Namely, teacher self-assessment scores are, on average, greater in magnitude than evaluator scores for any given observation; this pattern holds independent of the total number of annual observations a teacher received. Further, there is a downward trend in mean observation scores – both from evaluators and teacher self-assessment scores – as the number of total annual observations received increases, a pattern consistent with the fact that lower-performing teachers are annually assigned more classroom observations. At the same time, teacher performance, as measured by evaluator and teacher scores, rises across multiple observations within each teacher group receiving the same number of annual observations. Yet, for teachers receiving four annual observations – the lowest-performing teachers based on prior-year evaluation ratings – evaluator scores are significantly higher than teacher self-assessment scores for the fourth (and final) annual observation, producing an unexpectedly high dose of positive feedback.

<Figure 3 about here>

To formally examine the within-year patterns of observation scores presented in Figure 3, we estimate variants of the following regression specification:

$$(2) y_{jkt} = \delta \lambda_{jkt} + \phi_{jet} + u_{jkt},$$

where y_{jkt} is, alternatively, $Score_{jkt}^{teacher}$, $Score_{jkt}^{evaluator}$, or $Valence_{jkt}$. The variable λ_{jkt} represents the linear count k of observations received by teacher j in school year t , ϕ_{jet} captures teacher-by-evaluator-by-year fixed effects (FE), and u_{jkt} is the error term. The coefficient δ represents the magnitude by which observation scores or valence intensity change with each additional within-year observation (i.e., the score gradient). Further, the vector ϕ_{jet} effectively compares the change in observation scores across observations within teacher-by-evaluator pairs within each school year. We apply these FE because evaluator characteristics, recipient (teacher) traits, and interactions between these characteristics and traits can affect evaluator and self-assessment scores independent of true performance (Grissom & Bartanen, 2022; Ilgen et al., 1979; Kraft & Gilmour, 2016a; Lechermeier & Fassnacht, 2018; Steinberg & Sartain, 2021). The observation-level data directly contributing to our coefficients come from teachers observed at least three times by the same evaluator in the same year. We also estimate δ for teacher subgroups by total observations received, interacting λ_{jkt} with indicator variables for the total count of annual observations teacher j received in school year t .

Table 2 (Panel A) presents evidence of the within-year score gradient. On average, teacher self-assessment scores increase within a school year by 0.10 points (approximately 0.16 standard deviations of teacher scores) with each additional observation. In comparison, evaluator scores increase by 0.16 points (approximately 0.25 standard deviations of evaluator scores) with each additional observation. Thus, with each additional observation received, the intensive

valence increases by 0.06 points (approximately 0.11 standard deviations of intensive valence scores). This suggests that as teachers receive additional observations, their evaluators increasingly provide more positive feedback. Notably, the teacher self-assessment score gradient is relatively homogeneous across teachers receiving different total annual observations (see column V of Table 2, Panel A). However, the evaluator score gradient is increasing in the count of annual observations received (see column VI), suggesting that evaluators provide increasingly positive feedback to the lowest-performing teachers who receive more annual observations. Thus, the degree of positive feedback received and the magnitude of intensive valence is largest for teachers receiving four annual observations (0.07 points) – the lowest performing teachers, on average – while we detect no differences in intensive valence among higher-performing teachers receiving two annual observations (see column IV).

<Table 2 about here>

To examine the influence of the fourth (and final) evaluator score (i.e., $Score_{j4t}^{evaluator}$) on the measured growth in performance among the lowest-performing teachers, we estimate variants of equation (2) as follows. First, we estimate equation (2) on a subset of the full sample, which excludes the fourth teacher self-assessment score ($Score_{j4t}^{teacher}$) and the fourth evaluator score ($Score_{j4t}^{evaluator}$) for teachers with four annual observations; doing so enables insight into the performance growth of all teachers (and by total observations received) across just the first three classroom observations (see Table 2, Panel B). We then apply the parameter estimates from this regression to predict the fourth-observation evaluator score ($\widehat{Score}_{j4t}^{evaluator}$); i.e., the fourth observation score evaluators should have issued based on the observation score trend across the first three observations (note that we do not extrapolate the teacher self-assessment score for the fourth observation since it does not meaningfully deviate from the observation trend based on the

first three teacher scores; see Figure 4). Then, we create a new variable – $Score_{jkt}^{\widehat{evaluator}}$ – and replace the actual evaluator score from a teacher’s fourth classroom observation ($Score_{j4t}^{evaluator}$) with the predicted fourth-observation evaluator score ($Score_{j4t}^{\widehat{evaluator}}$). Next, we investigate whether (and the extent to which) the performance gradient across observations is influenced by the actual evaluator score from a teacher’s fourth classroom observation. To do so, we compare the parameter estimate δ (from Equation (2)) to the same parameter estimate from a regression in which only the observed fourth-observation evaluator score is replaced by the predicted fourth-observation evaluator score (see Table 2, Panel C); all other scores use the observed evaluator and teacher scores.

Results reveal that the growth in teacher performance – as rated by both teachers and evaluators – based on just the first three observations (see Panel B, columns V and VI) is identical to the estimated performance growth when we include the predicted fourth evaluator score (see Panel C, columns V and VI). Further, the intensive valence gradient is not only small in magnitude (though statistically significant) when excluding the fourth score and when using the predicted fourth evaluator score (0.03, see Panels B and C, column I), but also is substantively different in magnitude than the discordance gradient based on all observed evaluator scores (0.06). Notably, the estimated growth in teacher performance based on evaluator scores – the scores that determine high-stakes teacher evaluation ratings in this context – is significantly smaller in magnitude when excluding the fourth scores and when using the predicted fourth evaluator score (0.13, see Panels B and C, column III) than when based on all observed evaluator scores (0.16, see Panel A, column III). Together, these results suggest that evaluators may likely inflate the fourth and final observation score for the lowest-performing teachers.

As a robustness check on our primary results on the magnitude of intensive valence presented in Table 2, we estimate a nonparametric version of equation (2) in which we replace the linear count of total annual observations (λ_{jkt}) with λ_k , an indicator variable for the k^{th} observation up to the fourth classroom observation (the omitted reference category is the first classroom observation of the school year) (see Table 3). In alternative models, we include either month FE, which controls for the within-year timing of each classroom observation, or domain FE, which controls for potential differences in classroom observation scores by the domain of teacher performance. In all cases, results indicate significant positive feedback valence during the fourth (and final) classroom observation (see Table 3). Equations 1 and 2 and related robustness checks were applied to the full sample.

<Table 3 about here>

The Consequences of Feedback Valence

Methods

Within-Year Performance: Extensive Margin. We first consider relationships between within-year performance, as measured by evaluator scores, and our valence measure, a derivation of prior-observation teacher self-assessment and evaluator scores using the full sample. Omitted variables related to prior-observation teacher self-assessments and serially related to evaluator scores may undo our inferences about these relationships. Recent research from reformed teacher evaluation systems suggests time-invariant evaluator-by-teacher interactions (i.e., race-matching) may serially affect evaluator scores, implying that our research design should control for evaluator-by-teacher interactions that do not change over time (Grissom & Bartanen, 2022; Steinberg & Sartain, 2021). Research also intimates possible time-variant evaluator-by-teacher confounders, such as evaluator anticipation of feedback recipients' reactions to negative

feedback, which may aggravate the work environment or leaving the organization (school), which school administrators seek to avoid and may avoid by issuing inflated ratings (D. J. Campbell & Lee, 1988; Heidemeier & Moser, 2009; Kraft & Gilmour, 2016a; Rodriguez & Hunter, 2021). Additionally, variation in the composition of students taught by a teacher over time affects (summative) evaluator ratings (S. L. Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016). Therefore, we identify covariation between *Valence* and $Score_{jkt}^{evaluator}$ within evaluator-by-teacher-by-years, effectively comparing performance over months within an academic year within evaluator-by-teacher pairs. We first model the extensive margin of negative feedback valence on evaluator scores, estimating variants of the following specification:

$$(4) \text{Score}_{jkt}^{evaluator} = \delta I(\text{Valence}_{jt,k-1} < 0) + \phi_{jet} + u_{jkt}.$$

The indicator function $I(\text{Valence}_{jt,k-1} < 0)$ takes a value of one when the j th teacher receives negative feedback during the observation occurrence ($k-1$) within year t . The variable ϕ_{jet} captures teacher-by-evaluator-by-year FE, and as with equation (2), we estimate equation (4) using the full sample. At the extensive margin, δ represents the difference in $Score_{jkt}^{evaluator}$ after teacher j receives negative instead of positive feedback during their prior classroom observation across observations conducted by the same evaluator in an academic year. Mechanically, information from at least two observation occurrence-level panels within evaluator-by-teacher-by-years contributes to δ directly because ϕ absorbs variation from evaluator-by-teacher-by-years with a single panel. Each panel includes information from observations k and $(k-1)$; therefore, δ relies on information from teachers receiving at least three observations per year, who are among the least experienced or lowest performing regarding prior-year composite effectiveness scores.

We test the rigor and external validity of our inferences based on equation (4) under a separate set of assumptions about data generation and by estimating δ using information from all but the teachers receiving one observation per year (those with the highest prior-year composite effectiveness score), which we argue are teachers of the greatest practical and policy interest. Our inferences based on equation (4) could be undone if post-observation conferences with negative feedback systematically coincide with other performance-improving inputs independent of the observation process, and these coincidental inputs systematically explain significant variation in evaluator scores independent of evaluator-by-teacher-by-year interactions. Equation (5) and its variants account for many sources of dynamic within-year shocks via the lagged outcome:

$$(5) \text{Score}_{jkt}^{\text{evaluator}} = \delta I(\text{Valence}_{jt,k-1} < 0) + \text{Score}_{jt,k-1}^{\text{evaluator}} + u_{jkt}.$$

Equation (5) estimates δ using variation across and within teachers (and across and within evaluators) who, during observation ($k - 1$), received the same evaluator score but negative instead of positive feedback (due to higher-than-their-evaluator teacher self-assessment scores). Consistent with the value-added literature, we assume that the lagged outcome controls for all unobserved differences affecting δ up to the point when the lagged outcome was measured during observation ($k - 1$), which coincides with the receipt of treatment (negative feedback) (Bacher-Hicks & Koedel, 2023; Cowan et al., 2022). Notably, there is an average of four months between observations k and ($k-1$) among teachers receiving two observations per year and a mean of only two months between observations k and ($k - 1$) for teachers receiving four observations per year. Therefore, we assume that equation (5) accounts for all but the most recent pre-treatment differences between treated and untreated cases, leaving some alternative treatments occurring near the time of observation ($k - 1$) as the only source of variation in

$Score_{jkt}^{evaluator}$ capable of undoing our inferences. The alternative treatments equation (5) does not control for would need to have a relationship with $Score_{jkt}^{evaluator}$, but not $Score_{jt,k-1}^{evaluator}$ from 2 to 4 months prior; we assume such alternative treatments are unlikely to exist. In variants of equation (5), we apply evaluator or evaluator-by-year FE to restrict comparisons of $Score_{jkt}^{evaluator}$ to those issued by the same evaluator across months over years or months within academic years for reasons related to our discussion of confounding evaluator characteristics.

The final analysis of the within-year extensive margin examines the sensitivity of our inferences from equation (4) using tests introduced by Rosenbaum and Rubin (1983) and extended by Cinelli and Hazlett (2020). These tests report the maximum bias multiple, non-linear omitted variables (OV) can introduce under plausibly confounding conditions. We contextualize such conditions using $Score_{jt,k-1}^{evaluator}$. First, we add $Score_{jt,k-1}^{evaluator}$ to equation (4) to see if it affects δ since the differenced residual is necessarily correlated with the lagged outcome as both are functions of $u_{jt,k-1}$, which can create mechanical endogeneity. Then, we estimate the explanatory power of $Score_{jt,k-1}^{evaluator}$ regarding $Score_{jkt}^{evaluator}$ and $I(Valence_{jt,k-1} < 0)$ independent of ϕ . The sensitivity test for equation (4) assumes no OV explains more residual outcome or treatment variation (i.e., variation independent of ϕ) than $Score_{jt,k-1}^{evaluator}$. We then see if an OV resembling $Score_{jt,k-1}^{evaluator}$ explains enough residual outcome and treatment to undo our inferences; moreover, we press the limits of plausibility using scaled-up OVs explaining twice and thrice as much residual variation as $Score_{jt,k-1}^{evaluator}$. We then repeat these steps using $Score_{jt,k-1}^{teacher}$, or the teacher self-assessment scores from observation ($k - 1$).

Within-Year Performance: Intensive Margin. To estimate the intensive margin, we interact the indicator function $I(Valence_{jt,k-1} < 0)$ with the degree of valence intensity during

observation ($k - 1$), which we measure as the absolute value of $Valence_{jt,k-1}$ (i.e., $|Valence_{jt,k-1}|$). Including this interaction term enables insight into whether $Score_{jkt}^{evaluator}$ varies based on valence direction (negative or positive) and intensity. We specify the model as follows:

$$(6) \text{ } Score_{jkt}^{evaluator} = \delta I(Valence_{jt,k-1} < 0) + \beta_1 |Valence_{jt,k-1}| + \beta_2 [I(Valence_{jt,k-1} < 0) \cdot |Valence_{jt,k-1}|] + \phi_{jet} + u_{jtk}$$

In equation (6), β_1 represents the change in $Score_{jkt}^{evaluator}$ associated with a unit increase in valence intensity among cases of positive feedback during observation ($k - 1$) across observations within evaluator-by-teacher-by-years.^{vi} β_2 captures the differential change in $Score_{jkt}^{evaluator}$ as a function of valence intensity during observation ($k-1$) among cases of negative feedback relative to cases of positive feedback across observations within evaluator-by-teacher-by-years. The linear combination of $\beta_1 + \beta_2$ represents the total change in $Score_{jkt}^{evaluator}$ associated with a unit change in valence intensity during observation ($k - 1$) for cases of negative feedback across observations within evaluator-by-teacher-by-years. Standard errors in equations (4) – (6) are clustered at the teacher level.

Across-Year Productivity and Mobility: Extensive Margins. We examine changes in annual teacher value-added to student achievement (VAM, using the VAM sample) and teacher mobility across schools (using the teacher-by-year sample) associated with receiving a preponderance of negative feedback across observations within academic years using the following equation

$$(6) y_{jt} = \delta I(Valence_{jt} > 0) + X_{jt}A + W_{jt}B + \lambda_{jt} + \theta_{je} + u_{jt},$$

where y_{jt} is the state-issued VAM score or a binary mobility indicator for teacher j in year t .^{vii}

The mobility measure indicates whether (or not) teacher j remains in their school after the end of year t (*Retain*). Since y_{jt} is measured at the teacher-year level, we aggregate valence measures to the same level: $Valence_{jt}$ is the j th teacher's average valence intensity across all observations within year t and $I(Valence_{jt} > 0)$ takes a value of one when teacher j 's average intensive valence from all observations in year t is less than zero. The vector \mathbf{X} includes the classroom characteristics of students in teacher j 's class during school year t , including: prior-year student achievement, proportion of economically disadvantaged students, proportion of nonwhite students, and prior-year student disciplinary referrals as prior work finds that these aspects of classroom composition affect evaluator scores, and may therefore affect our valence measure and evaluator-provided evaluation-informed supports to improve teacher productivity (S. L. Campbell & Ronfeldt, 2018; Garrett & Steinberg, 2015; Steinberg & Garrett, 2016). The vector \mathbf{W} includes controls for the following teacher-level characteristics: gender, race, education level, years of experience, and prior-year observation scores, which can also affect evaluator scores and provided supports (Grissom & Bartanen, 2022; Steinberg & Sartain, 2021). The vector \mathbf{V} controls for evaluator-level characteristics for evaluator e who conducted observation k (for teacher j) in school year t , including gender, race, education level, years of experience, and prior-year observation score since these characteristics may affect evaluator-by-teacher matching and feedback provided (Hunter & Rodriguez, 2021; Hunter & Springer, 2022). λ_{jt} controls for the number of observations assigned to teacher j in year t because prior work suggests that evaluator knowledge about these assignments may (un)consciously influence evaluator scores and evaluation-informed supports to improve performance or productivity (Hunter, 2020). We also control for teacher-by-evaluator heterogeneity, such as race-matching effects, on teacher

productivity and mobility via θ_{je} . Equation (6) does not account for unobserved differences within evaluator-by-teacher pairs over time unrelated to our controls. Standard errors are clustered at the teacher level.

Findings

Within-Year Performance: Extensive Margin. Table 4 results suggest that teachers experiencing negative (instead of positive) feedback valence about their performance during the prior observation are rated, on average, higher by their evaluators at the extensive margin. Whether we compare evaluator scores within teacher-by-years or evaluator-by-teacher-by-years, the models predict an increase of 0.07 (0.11 standard deviations) in evaluator scores (columns I and II).

[Insert Table 4 about here]

Results from the teacher-by-year and evaluator-by-teacher-by-years FE models in columns I and II may reflect spurious reversion to within-teacher mean performance because we compare a teacher's performance during observation k conditional on a discontinuous function of their performance from observation $(k - 1)$, and because negative extensive valence is associated with low evaluator scores (see Figure 2). We empirically interrogate the possibility of mean reversion as an explanation of results in columns I and II using a falsification test under the assumption that a performance change of approximately x units during observation k within individuals due to mean reversion is preceded by an abnormal performance change of x units in the opposite direction during observation $(k - 1)$ (otherwise, the mean reversion during observation k would over- or under-compensate). If negative valence is effectively an indicator of positive mean reversion and our assumption is reasonable, we should observe heterogeneity within cases of negative and positive valence regarding the magnitude of the abnormal shock (x)

during observation $(k - 1)$ according to the magnitude of x . We implement this test by defining x as $(Score_{jkt}^{evaluator} - \overline{Score_{j*t}^{evaluator}}) / [SD(Score_{j*t}^{evaluator})]$, or evaluator scores standardized within evaluator-by-teacher-by-years, then find the quartiles of x , defined as $xQrt$. Next, we estimate variants of the following model:

$$(7) Score_{jkt}^{evaluator} = \delta I(Valence_{jt,k-1} < 0) + \pi I(Valence_{jt,k-1} < 0) * xQrt_{jet} + \beta_3 xQrt_{jet} + \phi_{jet} + u_{jkt},$$

where the reference group of $xQrt$ is the first quartile. We are interested in the joint significance of π ; if we do not detect joint significance, it suggests none of the heterogeneity needed for a valid positive mean reversion argument. The joint significance of π when standardizing within teacher-by-years is $p > 0.89$ and is $p > 0.60$ when standardizing within evaluator-by-teacher-by-years, casting significant doubt on a positive mean reversion explanation.

Furthermore, results in Table 4 columns III-V from the lagged outcome models without teacher-by-year FEs or interactions with them yield remarkably consistent estimates with the FE models in columns I and II. Comparing changes in $Score_{jkt}^{evaluator}$ conditional on negative valence among observations $(k - 1)$ with the same observation $(k - 1)$ evaluator score suggests a performance improvement of 0.08 units (0.13 standard deviations). Moreover, these relationships hold when comparing within years, evaluators and years, and evaluator-by-years, which would control for the endogenous influence of evaluator traits and characteristics that do not change within years. Notably, positive mean reversion is unlikely to account for estimates in the lagged outcome models because they do not make comparisons within teachers over time.

Within-Year Performance: Intensive Margin. We also find that $Score_{jkt}^{evaluator}$ responds to the intensity of negative feedback valence (see Table 5). A unit increase in positive valence during observation $(k - 1)$ is associated with a 0.25 unit (0.39 standard deviation)

decrease in $Score_{jkt}^{evaluator}$ during observation k .^{viii} In contrast, a unit increase in negative feedback is associated with a 0.07 increase in $Score_{jkt}^{evaluator}$ (0.11 standard deviation). Thus, this evidence suggests that teachers respond to more positive feedback at the intensive margin by reducing effort, reflected in a decrease in $Score_{jkt}^{evaluator}$. Conversely, negative feedback at the intensive margin appears to improve performance.

[Insert Table 5 about here]

Within-Year Performance: Sensitivity Tests. The sensitivity tests strongly suggest that inferences based on evaluator-by-teacher-by-year FE are insensitive to plausible OVs. In unprinted results, we find that adding the lagged outcome to the evaluator-by-teacher-by-year FE model yields estimates of 0.06 units (teacher-clustered standard error (0.03), $p < 0.05$), suggesting that doing so may introduce a negligible amount of mechanical endogeneity. We use the lagged outcome in the evaluator-by-teacher-by-year FE as a benchmark against which we assess the plausibility of inference-undoing OVs. Evaluator scores from observation $(k - 1)$ in these FE models explain only one percent of treatment and outcome variation not explained by the FE; if an OV explained similar residual variation in this model, it would shift the coefficient and 95% confidence interval to 0.04 and (0.02, 0.06) (Table 6 Row I). Moreover, OVs explaining twice the residual variation in treatment and outcome as $Score_{jt,k-1}^{evaluator}$ in the evaluator-by-teacher-by-year FE do not undo our inferences (Row II). Indeed, our inferences are not undone unless an OV explains three times as much residual variation in evaluator scores and prior-observation valence as $Score_{jt,k-1}^{evaluator}$ in the FE model (Row III); we assume that such OVs are unlikely. Rows IV-VI in Table 6 also suggest that our inferences based on the evaluator-by-teacher-by-years FE model are robust to OVs with once, twice, and thrice the explanatory power of $Score_{jt,k-1}^{teacher}$.

[Insert Table 6 about here]

Across-Year Productivity and Mobility: Extensive Margins. Teachers who receive negative feedback, on average, across observations within an academic year have higher end-of-year VAM scores than teachers who receive positive feedback, on average, by 1.30 units (0.20 standard deviations of VAM; column I, Table 7). These results provide additional support for our inference that negative feedback valence promotes growth. However, the relationship between annualized negative feedback and teacher retention suggests that growth may occur at the expense of teacher turnover (column II, Table 7). Teachers who receive annualized negative feedback, on average, are five percentage points less likely to remain in their school the following year; while substantively large (baseline teacher turnover is 12%), this estimate is not statistically significant. Results from additional mobility analyses on the probability that teachers exit the district or switch to a new school in the same district are consistent with these retention results (see Appendix C).

[Insert Table 7 about here]

If the teachers who exit their schools for another do so to avoid negative feedback, we expect them to switch to schools that are less likely to provide negative feedback. We tested this hypothesis by comparing the year t school-level average valence of school-switching teachers' sending and receiving schools (see Appendix C). Among the sample of teachers who switched schools, those who received extensive negative feedback from evaluators, on average, switched into receiving schools with 0.05 units (0.09 standard deviations) less discordance than their sending school. Stated differently, teachers who received negative feedback and switched schools within the district typically entered schools where teachers experienced, on average, more positive feedback valence.

Discussion

Recent evidence has yielded mixed findings about the impact of teacher evaluation reforms on student achievement, underscoring the importance of understanding the mechanisms through which evaluation systems might improve teaching and teacher effectiveness. Our study provides unique insight into one potentially critical mechanism - performance feedback - by leveraging novel data that captures both evaluator ratings and teacher self-assessments of classroom performance.

Our findings make several significant contributions to understanding how feedback functions within teacher evaluation systems. By examining both evaluator ratings and teacher self-assessments, we learned that while teachers and evaluators generally align in their assessments of classroom performance, there is substantial observation-level variability in these scores, yielding meaningful variation in valence intensity. The alignment in typical ratings differs notably from research outside K-12 education, where employee self-assessments tend to exceed evaluator ratings (Church et al., 2019; Heidemeier & Moser, 2009).

Most significantly, we infer that negative feedback valence improves teacher performance. The positive effects of negative feedback are even more striking when considering feedback intensity. These improvements in teacher performance following negative feedback are corroborated by across-year analyses showing higher annual VAM scores for teachers receiving negative feedback, on average. The differential responses to negative and positive feedback valence align with theoretical predictions about how valence affects performance (Kim et al., 2018; Lechermeier & Fassnacht, 2018); the performance decline following intense positive feedback suggests that in this context, positive feedback valence may reduce motivation to improve, resulting in performance declines (Cianci et al., 2010; Ilgen et al., 1979).

Several features of our research designs strengthen inferential rigor about the effects of negative feedback valence. The consistency of results across evaluator-by-teacher-by-year fixed effects and lagged outcome models, which rely on different identification assumptions, suggests findings are not driven by dynamic selection based on teacher performance two to four months prior to observation k , unobserved factors associated with that performance, or annually varying interactions between teacher-evaluator pairs. Furthermore, formal sensitivity tests indicate that our inferences from FE models are robust to (arguably implausible) omitted variables.

Limitations and Future Research

While our study enables rigorous examination of feedback valence, it faces three broad limitations. First, we do not assume that our findings generalize to all reformed settings or to the highest-performing teachers in a setting. Second, we cannot definitively rule out all alternative explanations for the relationship between negative feedback and improved performance. Finally, we do not directly observe the specific micro-mechanisms (e.g., increased effort, evaluation-informed professional development) through which negative feedback leads to improvement.

Future research might examine generalizability, micro-mechanisms, how specific features of feedback delivery moderate its effects, and whether and how evaluators use feedback or evaluator scores for multiple purposes. Our findings also raise questions about performance assessments by evaluators, who may accurately assess performance in general but inflate assessments in cases furthering the attainment of another goal, such as teacher retention (Kraft & Gilmour, 2016a; Rodriguez & Hunter, 2021).

Implications

Several features of reformed teacher evaluation systems may create conditions enabling negative feedback to improve rather than inhibit performance. Standards-based rubrics provide

clear performance expectations and help teachers identify specific areas for improvement. These rubrics also allow evaluators and teachers to develop a shared understanding of performance expectations and rubric applications. That teachers in the study district are expected to self-assess their performance using a standards-based rubric for each formal observation is a unique feature of the system examined and one that may encourage productive post-observation conferences. Self-assessments promote reflection, enable teachers to prepare for post-observation conferences by collecting evidence and examples that could promote more rigorous discussion of performance, and support teacher agency by involving them in the evaluation process and performance improvement, all of which can enable positive effects for negative feedback (Church et al., 2019; Heidemeier & Moser, 2009; Hunter, 2023b, 2024). Additionally, evaluator training and certification requirements may enhance their credibility - a factor that prior work identifies as critical for negative feedback to motivate improvement rather than disengagement (Ilgen et al., 1979; Jawahar, 2006). Policies requiring frequent post-observation conferences within one week of assessment could also facilitate improve by ensuring feedback timeliness, which research suggests is important for translating negative feedback into performance improvements (Lechermeier & Fassnacht, 2018).

Successfully leveraging feedback mechanisms requires carefully cultivating conditions enabling them to serve developmental purposes. One of those conditions reflects principals' willingness in the study district to issue negative feedback, something principals in other settings eschew (Kraft & Gilmour, 2016a). In addition to encouraging schoolwide norming sessions to establish a shared understanding of rubric applications, study district leaders might have facilitated the creation of work environments where teachers accepted feedback productively, and principals believed they could deliver it without recourse (Quintelier et al., 2020b, 2020a).

While negative feedback seems to improve performance, an unintended consequence could be that teachers do what school administrators report is a significant concern regarding teacher evaluation—that it prompts teachers to leave the school (Kraft & Gilmour, 2016a). These dynamics underscore the difficulty in creating conditions where the comprehensive effects of negative feedback yield school and individual teacher improvement.

Bibliography

- Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher Turnover, Teacher Quality, and Student Achievement in DCPS. *Educational Evaluation and Policy Analysis*, 39(1), 54–76.
- Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology*, 20(1), 1–48.
- Almy, S. (2011). *Fair to Everyone: Building the Balanced Teacher Evaluations that Educators and Students Deserve* (Teacher Quality). The Education Trust.
<https://edtrust.org/resource/fair-to-everyone-building-the-balanced-teacher-evaluations-that-educators-and-students-deserve/>
- 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>
- Bacher-Hicks, A., & Koedel, C. (2023). Estimation and interpretation of teacher value added in research applications. In *Handbook of the Economics of Education* (Vol. 6).
- Bleiberg, J., Brunner, E., Harbatkin, E., Kraft, M. A., & Springer, M. G. (In Press). The Effect of Teacher Evaluation on Achievement and Attainment: Evidence from Statewide Reforms. *Journal of Political Economy Microeconomics*, In Press. <https://doi.org/10.1086/732837>
- Bogard, J. E., Delmas, M. A., Goldstein, N. J., & Vezich, I. S. (2020). Target, distance, and valence: Unpacking the effects of normative feedback. *Organizational Behavior and Human Decision Processes*, 161, 61–73. <https://doi.org/10.1016/j.obhdp.2020.10.003>
- Campbell, D. J., & Lee, C. (1988). Self-Appraisal in Performance Evaluation: Development Versus Evaluation. *Academy of Management Review*, 13(2), 302–314.

- Campbell, S. L., & Ronfeldt, M. (2018). Observational Evaluation of Teachers: Measuring More Than We Bargained for? *American Educational Research Journal*, 000283121877621. <https://doi.org/10.3102/0002831218776216>
- Carver, C. S., & Scheier, M. F. (1982). Control Theory: A Useful Conceptual Framework for Personality-Social, Clinical, and Health Psychology. *Psychological Bulletin*, 92(1), 25.
- Chambers, J., Brodziak de los Reyes, I., & O'Neil, C. (2013). *How Much Are Districts Spending to Implement Teacher Evaluation Systems? Case Studies of Hillsborough County Public Schools, Memphis City Schools, and Pittsburgh Public Schools* (Working Paper No. WR-989-BMGF; RAND Working Paper). RAND Corporation.
- Cherasaro, T. L., Brodersen, R. M., Reale, M. L., & Yanoski, D. C. (2016). *Teachers' responses to feedback from evaluators: What feedback characteristics matter?* (No. REL 2017-190; Making Connections, pp. 1–29). REL Central.
- Church, A. H., Bracket, D. W., Fleenor, J. W., & Rose, D. S. (2019). *The Handbook of Strategic 360 Feedback*. Oxford University Press.
- Cianci, A. M., Schaubroeck, J. M., & McGill, G. A. (2010). Achievement Goals, Feedback, and Task Performance. *Human Performance*, 23(2), 131–154. <https://doi.org/10.1080/08959281003621687>
- Cowan, J., Goldhaber, D., & Theobald, R. (2022). Performance Evaluations as a Measure of Teacher Effectiveness When Implementation Differs: Accounting for Variation across Classrooms, Schools, and Districts. *Journal of Research on Educational Effectiveness*, 15(3), 510–531. <https://doi.org/10.1080/19345747.2021.2018747>

- Cullen, J. B., Koedel, C., & Parsons, E. (2021). The Compositional Effect of Rigorous Teacher Evaluation on Workforce Quality. *Education Finance and Policy*, 16(1), 7–41.
https://doi.org/10.1162/edfp_a_00292
- Cutumisu, M., & Schwartz, D. L. (2016). Choosing versus Receiving Feedback: The Impact of Feedback Valence on Learning in an Assessment Game. *Choosing versus Receiving Feedback: The Impact of Feedback Valence on Learning in an Assessment Game*, 9th, 341–346.
- de Barros, A. (2019). Evaluating Teacher Evaluation: Evidence from Chile. *Organization of Schools and Systems & Education in Global Contexts*. Society for Research in Educational Effectiveness, Washington, DC.
- Dee, T. S., James, J., & Wyckoff, J. (2021). Is Effective Teacher Evaluation Sustainable? Evidence from District of Columbia Public Schools. *Education Finance and Policy*, 16(2), 313–346. https://doi.org/10.1162/edfp_a_00303
- Dee, T. S., & Wyckoff, J. (2015). Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, 34(2).
<https://doi.org/10.1002/pam>
- Donaldson, M. L. (2021). *Multidisciplinary Perspectives on Teacher Evaluation: Understanding the Research and Theory* (1st ed.). Routledge.
- Fong, C. J., Patall, E. A., Vasquez, A. C., & Stautberg, S. (2019). A Meta-Analysis of Negative Feedback on Intrinsic Motivation. *Educational Psychology Review*, 31(1), 121–162.
<https://doi.org/10.1007/s10648-018-9446-6>
- Garrett, R., & Steinberg, M. P. (2015). Examining Teacher Effectiveness Using Classroom Observation Scores: Evidence From the Randomization of Teachers to Students.

Educational Evaluation and Policy Analysis, 37(2), 224–242.

<https://doi.org/10.3102/0162373714537551>

Goff, P. T., Guthrie, J. E., Goldring, E., & Bickman, L. (2014). Changing principals' leadership through feedback and coaching. *Journal of Educational Administration*, 52(5), 682–704.

<https://doi.org/10.1108/JEA-10-2013-0113>

Goldring, E., 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.

<https://doi.org/10.1177/0013161X14556152>

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.

<https://doi.org/10.1002/pam.22352>

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>

Grissom, J. A., Kalogrides, D., & Loeb, S. (2017). Strategic Staffing? How Performance Pressures Affect the Distribution of Teachers Within Schools and Resulting Student Achievement. *American Educational Research Journal*, 54(6), 1079–1116.

<https://doi.org/10.3102/0002831217716301>

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

Halverson, R., Kelley, C., & Kimball, S. M. (2004). Implementing Teacher Evaluation Systems: How Principals Make Sense of Complex Artifacts to Shape Local Instructional Practice. In W. K. Hoy & C. G. Miskel (Eds.), *Educational Administration, Policy, and Reform: Research and Measurement*. Information Age Publishing.

Heidemeier, H., & Moser, K. (2009). Self–other agreement in job performance ratings: A meta-analytic test of a process model. *Journal of Applied Psychology, 94*(2), 353–370.
<https://doi.org/10.1037/0021-9010.94.2.353>

Ho, A. D., & Kane, T. J. (2013). *The Reliability of Classroom Observations by School Personnel. Research Paper. MET Project.*
http://pitt.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwY2BQAB2JnphobmqYlJSSbGSRaJlmZJgErCfMDBKNTc0sUIBWVCKV5m5CDEypeaIMMm6uIc4euqAJi_gCyJkL8a4uwKaFpam5oRgDC7BfnCrOwJoGjB8gDSwzxYH6xRk4IizDQy0ivf0gXCEYV68YvH9Jr7BEHFhEg6NX11jPAACqsie0

Holderness, D. K., Olsen, K. J., & Thornock, T. A. (2017). Who Are *You* to Tell Me *That?! The Moderating Effect of Performance Feedback Source and Psychological Entitlement on Individual Performance. Journal of Management Accounting Research, 29*(2), 33–46.
<https://doi.org/10.2308/jmar-51538>

Hunter, S. B. (2020). The Unintended Effects of Policy-Assigned Teacher Observations: Examining the Validity of Observation Scores. *AERA Open, 6*(2).
<https://doi.org/10.1177/2332858420929276>

Hunter, S. B. (2022). *The (In)Consistency of Teacher Survey Responses About Teacher Evaluation Implementation: Implications for Policymaking* [Working Paper].

- Hunter, S. B. (2023a). Do You Mean What I Mean? Comparing Teacher Performance Self-Scores and Evaluator-Generated Scores. *Journal of Education Human Resources*, 41(2), 210–250. <https://doi.org/10.3138/jehr-2020-0026>
- Hunter, S. B. (2023b). Explaining variation in the implementation of observational processes by school leaders: Evidence from a Tennessee-based researcher–practitioner partnership. *Journal of Educational Administration*, 61(1), 16–40. <https://doi.org/10.1108/JEA-03-2022-0049>
- Hunter, S. B. (2024). High-leverage teacher evaluation practices for instructional improvement. *Educational Management Administration & Leadership*, 52(4), 991–1013. <https://doi.org/10.1177/17411432221112995>
- Hunter, S. B., & Kho, A. (2023). The Effects of Teacher Evaluation Policy on Student Achievement and Teacher Turnover: Leveraging Teacher Accountability and Teacher Development. *Journal of Education Human Resources*, Advance Access. <https://doi.org/10.3138/jehr-2023-0040>
- Hunter, S. B., Kho, A., & Bowser, K. (2023). Policy-Assigned Teacher Observations, Their Implementation, and Student Discipline Outcomes: Main, Mediated, and Moderated Relationships. *Tennessee Education Research Alliance Working Paper*, 2023(December).
- Hunter, S. B., & Rodriguez, L. A. (2021). Examining the demands of teacher evaluation: Time use, strain and turnover among Tennessee school administrators. *Journal of Educational Administration*, 59(6), 739–758. <https://doi.org/10.1108/JEA-07-2020-0165>
- Hunter, S. B., & Springer, M. G. (2022). Performance Feedback, Human Capital, and Teacher Performance: A Mixed-Methods Analysis. *Educational Evaluation and Policy Analysis*, 44(3), 380–403. <https://doi.org/10.3102/01623737211062913>

- Hunter, Seth B. & Bowser, Katherine M. (2024). Next-Generation Teacher Evaluation in Rural Missouri: Main and Moderated Effects on Student Achievement and Effects-to-Expenditure Ratios. *EdWorkingPaper*, 24(935), 1–63.
<https://doi.org/doi.org/10.26300/x36v-vs97>
- Ilggen, 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.
<https://doi.org/10.1037/0021-9010.64.4.349>
- Jawahar, I. M. (2006). Correlates of satisfaction with performance appraisal feedback. *Journal of Labor Research*, 27(2), 213–236. <https://doi.org/10.1007/s12122-006-1004-1>
- Jawahar, I. M. (2007). The Influence of Perceptions of Fairness on Performance Appraisal Reactions. *Journal of Labor Research*, 28(4), 735–744. <https://doi.org/10.1007/s12122-007-9014-1>
- Kim, S., Hwang, S., & Lee, M. (2018). The benefits of negative yet informative feedback. *PLOS ONE*, 13(10), e0205183. <https://doi.org/10.1371/journal.pone.0205183>
- 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., & Christian, A. (2021). Can Teacher Evaluation Systems Produce High-Quality Feedback? An Administrator Training Field Experiment. *American Educational Research Journal*, 00028312211024603. <https://doi.org/10.3102/00028312211024603>
- Kraft, M. A., & Gilmour, A. F. (2016a). 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>

- Kraft, M. A., & Gilmour, A. F. (2016b). Revisiting the Widget Effect: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness. *Association of Education Finance and Policy*, 1–31.
- Lechermeier, J., & Fassnacht, M. (2018). How do performance feedback characteristics influence recipients' reactions? A state-of-the-art review on feedback source, timing, and valence effects. *Management Review Quarterly*, 68(2), 145–193.
<https://doi.org/10.1007/s11301-018-0136-8>
- Liebowitz, D. D., Porter, L., & Bragg, D. (2022). The Effects of Higher-Stakes Teacher Evaluation on Office Disciplinary Referrals. *Journal of Research on Educational Effectiveness*, 15(3), 475–509. <https://doi.org/10.1080/19345747.2021.2015496>
- London, M., & Smither, J. W. (2002). Feedback orientation, feedback culture, and the longitudinal performance management process. *Human Resource Management Review*, 12(1), 81–100. [https://doi.org/10.1016/S1053-4822\(01\)00043-2](https://doi.org/10.1016/S1053-4822(01)00043-2)
- Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A Composite Estimator of Effective Teaching* (pp. 1–51). http://www.nbexcellence.org/cms_files/resources/Jan_2013_A_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf
- National Council on Teacher Quality. (2019). *NCTQ: Yearbook: State Teacher Policy Database*. National Council on Teacher Quality (NCTQ). <https://www.nctq.org/yearbook/home>
- Pham, L. D., Nguyen, T. D., & Springer, M. G. (2021). Teacher Merit Pay: A Meta-Analysis. *American Educational Research Journal*, 58(3), 527–566.
<https://doi.org/10.3102/0002831220905580>

- Quintelier, A., De Maeyer, S., & Vanhoof, J. (2020a). Determinants of teachers' feedback acceptance during a school inspection visit. *School Effectiveness and School Improvement, 31*(4), 529–547. <https://doi.org/10.1080/09243453.2020.1750432>
- Quintelier, A., De Maeyer, S., & Vanhoof, J. (2020b). The role of feedback acceptance and gaining awareness on teachers' willingness to use inspection feedback. *Educational Assessment, Evaluation and Accountability, 32*(3), 311–333. <https://doi.org/10.1007/s11092-020-09325-9>
- Raaijmakers, S. F., Baars, M., Schaap, L., Paas, F., & Van Gog, T. (2017). Effects of performance feedback valence on perceptions of invested mental effort. *Learning and Instruction, 51*, 36–46. <https://doi.org/10.1016/j.learninstruc.2016.12.002>
- Rodriguez, L. A., & Hunter, S. B. (2021). Making Do: Why Do Administrators Retain Low-Performing Teachers? *Educational Researcher, 50*(9), 673–676. <https://doi.org/10.3102/0013189X211039450>
- Song, M., Wayne, A. J., Garet, M. S., Brown, S., & Rickles, J. (2021). Impact of Providing Teachers and Principals with Performance Feedback on Their Practice and Student Achievement: Evidence from a Large-Scale Randomized Experiment. *Journal of Research on Educational Effectiveness, 1*–26. <https://doi.org/10.1080/19345747.2020.1868030>
- Stecher, B., Holtzman, D., Garet, M., Hamilton, L., Engberg, J., Steiner, E., Robyn, A., Baird, M., Gutierrez, I., Peet, E., Brodziak de los Reyes, I., Fronberg, K., Weinberger, G., Hunter, G., & Chambers, J. (2018). *Improving Teaching Effectiveness: Final Report: The Intensive Partnerships for Effective Teaching Through 2015–2016*. RAND Corporation. <https://doi.org/10.7249/RR2242>

- Steinberg, M. P., & Donaldson, M. L. (2016). The New Educational Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era. *Education Finance and Policy, 11*(3). https://doi.org/10.1162/EDFP_a_00186
- Steinberg, M. P., & Garrett, R. (2016). Classroom Composition and Measured Teacher Performance: What Do Teacher Observation Scores Really Measure? *Educational Evaluation and Policy Analysis, 38*(2), 293–317. <https://doi.org/10.3102/0162373715616249>
- Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago’s Excellence in Teaching Project. *Education Finance and Policy, 10*(4), 535–572. https://doi.org/10.1162/EDFP_a_00173
- Steinberg, M. P., & Sartain, L. (2021). What Explains the Race Gap in Teacher Performance Ratings? Evidence From Chicago Public Schools. *Educational Evaluation and Policy Analysis, 43*(1), 60–82. <https://doi.org/10.3102/0162373720970204>
- Taylor, E. S., & Tyler, J. H. (2012). The Effect of Evaluation on Teacher Performance. *American Economic Review, 102*(7), 3628–3651. <https://doi.org/10.1257/aer.102.7.3628>
- Van Dijk, D., & Kluger, A. N. (2011). Task type as a moderator of positive/negative feedback effects on motivation and performance: A regulatory focus perspective. *Journal of Organizational Behavior, 32*(8), 1084–1105. <https://doi.org/10.1002/job.725>
- Vohra, N., & Singh, M. (2005). Mental traps to avoid while interpreting feedback: Insights from administering feedback to school principals. *Human Resource Development Quarterly, 16*(1), 139–147. <https://doi.org/10.1002/hrdq.1128>
- Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget Effect* (pp. 48–48). http://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf

Wiswede, D., Münte, T. F., & Rüsseler, J. (2009). Negative affect induced by derogatory verbal feedback modulates the neural signature of error detection. *Social Cognitive and Affective Neuroscience*, 4(3), 227–237. <https://doi.org/10.1093/scan/nsp015>

Youngs, P., Kim, J., & Mavrogordato, M. (Eds.). (2020). *Exploring Principal Development and Teacher Outcomes: How Principals Can Strengthen Instruction, Teacher Retention, and Student Achievement* (1st ed.). Routledge.

ⁱ Rigorous nationwide research examining effects on office referrals detects no effects (Liebowitz et al., 2022), though emerging work finds that assigning schools more observations plausibly reduces the number of students receiving exclusionary discipline outcomes by improving classroom management (Hunter et al., 2023).

ⁱⁱ See Appendix B for details about the composite effectiveness score.

ⁱⁱⁱ Following prior work on the construction of observation-level teacher performance scores (Garrett & Steinberg, 2015; Ho & Kane, 2013; Hunter, 2023a; Mihaly et al., 2013), we average across all 19 items at the observation (k) level to construct a teacher performance score at the observation occurrence level.

^{iv} School administrators also receive a *de facto* summative observation score based on a portfolio of evidence and two observations per year. We link these summative scores to evaluators who are school administrators.

^v If evaluator and teacher scores are measured with substantial error, *Valence* is more error prone. To the extent these conditions are true, a *Valence* predictor would attenuate coefficients, suggesting that the subsequent findings are conservative.

^{vi} We explored whether the relationship between valence and teacher performance exhibited diminishing marginal returns by including second-order polynomials of $|Valence_{jt}|$. The second-order term was never statistically significant at conventional levels. Results are available upon request.

^{vii} We do not examine summative evaluator scores as an outcome for econometric reasons. A teacher's summative observation score is their average evaluator score across all observations. $Valence_{jt}$ is a function of all evaluator-assigned observation scores, like the summative evaluator score. Regressing summative evaluator scores on $Valence_{jt}$ effectively regresses summative observation scores on itself.

^{viii} Although these associations are large, they are based on substantively large changes in feedback valence; a unit increase in intensive valence is equivalent to 1.82 standard deviations of intensive valence.

Tables & Figures

Table 1. Teacher Characteristics

	All Teachers	Annual Classroom Observations			
		One	Two	Three	Four
<u>Panel A: Teacher Characteristics</u>					
Female	0.80	0.82	0.79	0.78	0.79
Nonwhite	0.26	0.25	0.28	0.24	0.23
Experience	9.70 (9.31)	12.65 (9.16)	12.53 (9.12)	3.37 (5.70)	1.87 (3.66)
Masters+	0.07	0.10	0.08	0.06	0.03
<u>Panel B: Teacher Performance Measures</u>					
Summative Observation Score	3.57 (0.45)	3.57 (0.44)	3.56 (0.44)	3.51 (0.47)	3.36 (0.60)
Comp-Cont	334.80 (78.69)	342.53 (79.49)	335.13 (78.29)	315.53 (78.35)	330.45 (77.75)
VAM	-1.15 (6.62)	0.18 (6.99)	-1.49 (5.69)	-3.04 (7.33)	-1.92 (5.36)
N(Teacher-Year)	9,070	5,049	2,828	737	456

Notes. In Panel A, each cell reports proportions, except for *Experience*, which reports means (standard deviations). *Masters+* includes teachers who have more than a master’s degree. In Panel B, each cell reports means (standard deviations) of the teacher performance measure from year *t*. *Observation Score* represents teacher performance ratings from formal classroom observations and range from 1 – 5. *Comp-Cont* is the composite teacher effectiveness score which is a continuous measure from 100 – 500. *VAM* is a state-issued value-added measure ranging from -100 to 60. The count of teacher-year observations includes teachers with non-missing teacher self-assessment and evaluator scores (some teachers are missing values for some characteristics in the table); there are 5,251 unique teachers in the sample.

Table 2. Within-Year Changes in Classroom Observation Scores

	I	II	III	IV	V	VI
	Intensive Valence	Teacher Scores	Evaluator Scores	Intensive Valence	Teacher Scores	Evaluator Scores
Panel A. All Scores						
Observations	0.06*** (0.00)	0.10*** (0.00)	0.16*** (0.00)			
2 Annual Obs: Observations				0.02 (0.01)	0.10*** (0.01)	0.11*** (0.01)
3 Annual Obs: Observations				0.04*** (0.01)	0.09*** (0.01)	0.14*** (0.01)
4 Annual Obs: Observations				0.07*** (0.01)	0.10*** (0.01)	0.18*** (0.01)
N(Tch-Yr-Obs)	20,045	20,045	20,045	20,045	20,045	20,045
Panel B. Excluding Fourth Score						
Observations	0.03*** (0.01)	0.10*** (0.01)	0.13*** (0.00)			
2 Annual Obs: Observations				0.02 (0.01)	0.10*** (0.01)	0.11*** (0.01)
3 Annual Obs: Observations				0.04*** (0.01)	0.09*** (0.01)	0.14*** (0.01)
4 Annual Obs: Observations				0.03*** (0.01)	0.10*** (0.01)	0.13*** (0.01)
N(Tch -Yr-Obs)	18,257	18,257	18,257	18,257	18,257	18,257
Panel C. First Through Third Scores and Predicted Fourth Score						
Observations	0.03*** (0.00)	0.10*** (0.00)	0.13*** (0.00)			
2 Annual Obs: Observations				0.02 (0.01)	0.10*** (0.01)	0.11*** (0.01)
3 Annual Obs: Observations				0.04*** (0.01)	0.09*** (0.01)	0.14*** (0.01)
4 Annual Obs: Observations				0.03*** (0.00)	0.10*** (0.00)	0.13*** (0.00)
N(Tch-Yr-Obs)	20,045	20,045	20,045	20,045	20,045	20,045
N(Tch-Yr)	9,070	9,070	9,070	9,070	9,070	9,070

Notes. Each column (within a panel) is a separate regression. Coefficients reported with standard errors (in parentheses) clustered at the teacher-level. Outcomes are regressed on a nonparametric operationalization of the k th observation and teacher-by-evaluator-by-year fixed effects. Panel A uses the full sample; Panel B excludes the fourth observation score; Panel C uses the full sample with predicted fourth score. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 3. Nonparametric Estimates of Intensive Valence, by Total Observations Received

	I	II	III
2 nd Observation	0.05* (0.02)	0.03 (0.02)	0.04 (0.03)
3 rd Observation	0.06** (0.02)	0.04* (0.02)	0.11** (0.04)
4 th Observation	0.22*** (0.03)	0.16*** (0.03)	0.25*** (0.05)
N(Tch*Yr*Obs)	20045	20045	9944
Eval*Tch*Yr FE	X	X	X
Domain FE		X	
Month FE			X

Notes. N(Teacher-Year-Observation) = 20,045 and N(Teacher-Year) = 9,070 in Teacher-by-Evaluator-by-Year FE model and Teacher-by-Evaluator-by-Year FE and Domain FE model. N(Teacher-Year-Observation) = 9,944 and N(Teacher-Year) = 4,533 in Teacher-by-Evaluator-by-Year FE and Month FE model; samples differ due to missing month data. Each column represents a different regression and coefficients are reported with standard errors (in parentheses) clustered at the teacher-level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 4. Extensive Negative Valence on Within-Year Change in Evaluator Scores

	I	II	III	IV	V
Extensive Negative Valence	0.07* (0.03)	0.07** (0.02)	0.08*** (0.01)	0.08*** (0.01)	0.08*** (0.01)
Prior-Obs Evaluator Score			0.73*** (0.01)	0.66*** (0.01)	0.65*** (0.01)
N(Tch*Yr*Obs)	10840	10840	10840	10840	10840
Year FE			X	X	
Eval FE				X	
Eval*Yr FE					X
Tch*Yr FE	X				
Eval*Tch*Yr FE		X			

Notes. Each column represents a different regression and coefficients are reported with standard errors (in parentheses) clustered at the teacher-level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5. Intensive Negative Valence on Within-Year Change in Evaluator Scores

Extensive Negative Valence	-0.01 (0.03)
Intensive Valence	-0.25*** (0.04)
Extensive Negative Valence * Intensive Valence	0.32*** (0.05)
Ext Negative Valence * Int Valence + Int Valence	0.07* (0.04)
N(Teacher*Year)	6,661
N(Teacher*Year*Observation)	10,837

Notes. Each column represents a different regression and coefficients are reported with standard errors (in parentheses) clustered at the teacher-level. Observation scores are regressed on valence measures and teacher-by-evaluator-by-year FE. Panel A is the subset of teachers from the full sample with at least two annual observations. Panel B is the subset of teachers in Panel A with VAM scores. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 6. Sensitivity Tests for Teacher-by-Evaluator-Year FE Estimates

		I	II	III
		$R^2_{Valence \sim OV X}$	$R^2_{Eval\ Scores \sim OV X, Valence}$	Coef (95% CI)
I	1.00x Prior-Obs Evaluator Score	0.010	0.013	0.04** (0.02, 0.06)
II	2.00x Prior-Obs Evaluator Score	0.020	0.026	0.03* (0.01, 0.05)
III	3.00x Prior-Obs Evaluator Score	0.030	0.039	0.01 (-0.01, 0.03)
IV	1.00x Prior-Obs Teacher Score	0.125	0.0001	0.06*** (0.03, 0.11)
V	2.00x Prior-Obs Teacher Score	0.249	0.0002	0.05*** (0.02, 0.1)
VI	3.00x Prior-Obs Teacher Score	0.374	0.0003	0.05*** (0.01, 0.09)
N(Teacher*Year*Observation)				10,837

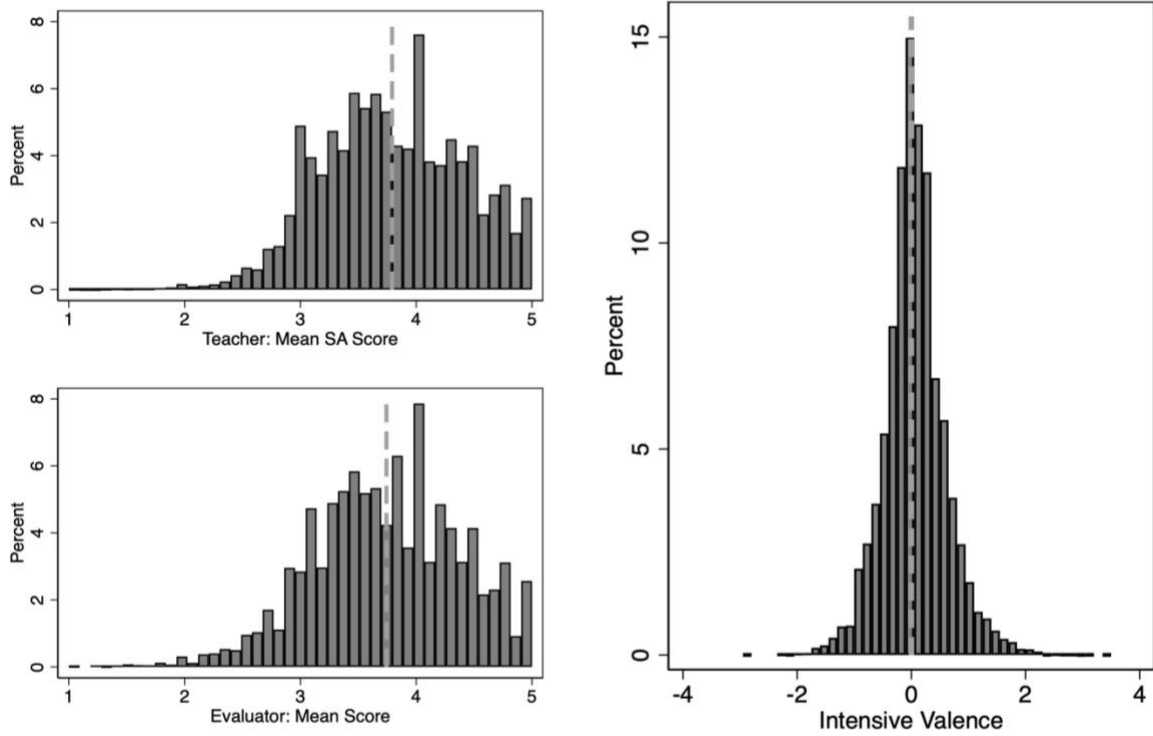
Notes. Models apply Equation 4 (teacher-by-evaluator-by-year fixed effects). *Valence* is the binary measure of negative balance, *OV* the hypothetical omitted variables, and *X* fixed effects. $R^2_{Valence \sim OV|X}$ represents the proportion of explained residual variation in *Valence*. $R^2_{Eval\ Scores \sim OV|X, Valence}$ represents the proportion of explained residual variation in the *Evaluator Score* outcome. “Coef” is the estimated treatment effect if Equation 4 controls for *OV*. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7. Annualized Extensive Negative Valence on Teacher VAM and Retention

	I	II
	VAM	Retained
Summative Extensive Negative Valence	1.30* (0.65)	-0.05 (0.03)
N(Teacher*Year)	537	3,980

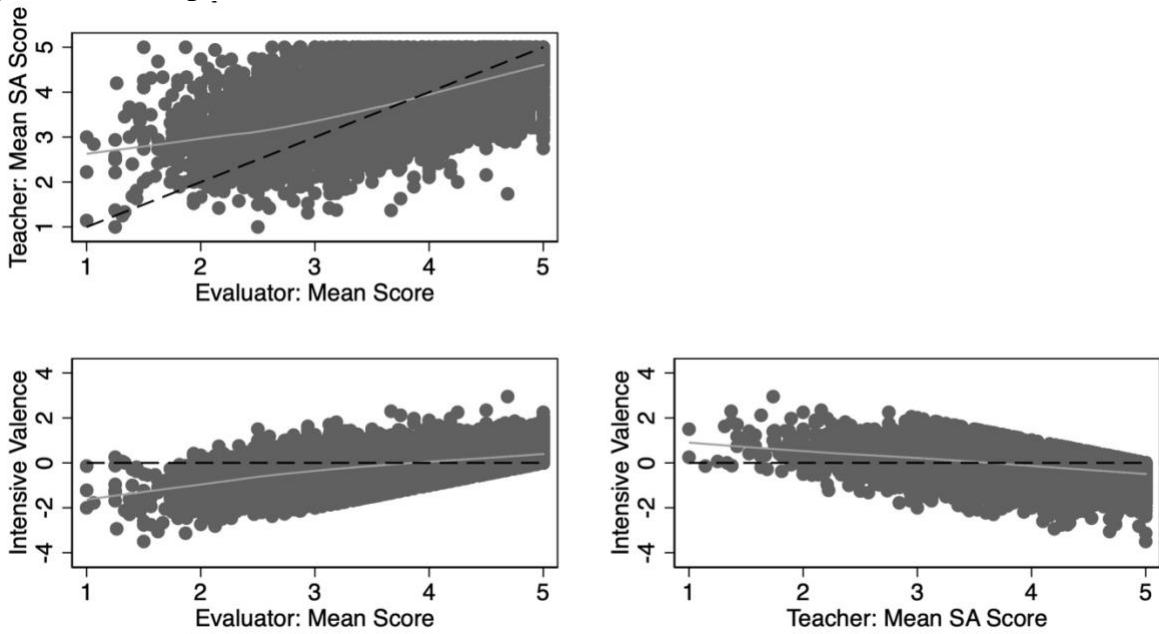
Notes. Each column represents a separate regression and coefficients are reported with standard errors (in parentheses) clustered at the teacher-level. Teacher-years are the unit of analysis. Outcomes are regressed on a binary measure of negative valence and teacher-by-evaluator fixed effects and year fixed effects. Column I is limited to teachers of tested subjects with VAM scores. Column II includes all teachers and is estimated by linear probability models. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1. Distribution of Classroom Observation Scores, by Evaluator and Teacher



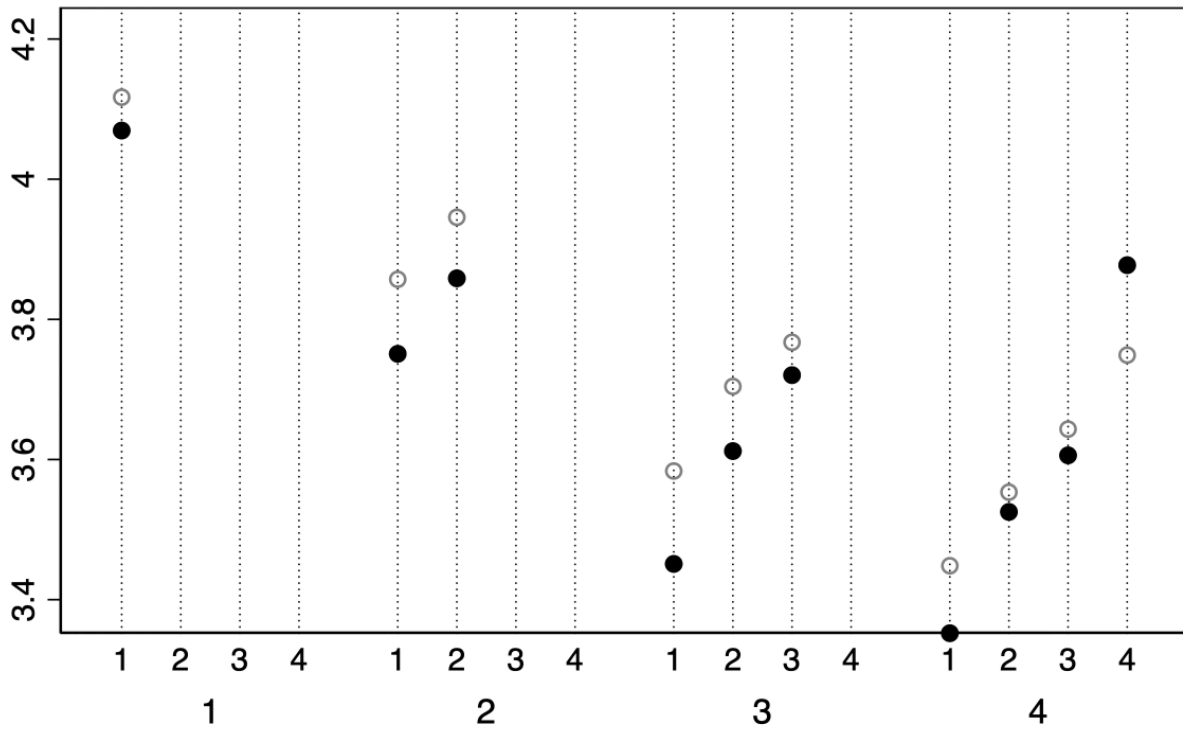
Notes. Observation occurrences are the unit of analysis. Each figure shows the distribution of classroom observation scores. The mean (standard deviation) of teacher self-scores is 3.79 (0.62); the mean (standard deviation) of evaluator scores is 3.74 (0.64); and the mean (standard deviation) of discordance scores is -0.06 (0.54). The count of observations at the teacher-year-observation level is 20,045.

Figure 2. Scatterplots and Lowess Curves: Teacher-Observation-Year Means



Notes. Observation occurrences are the unit of analysis. Each figure plots teacher-observation-year classroom observation scores. Solid gray lines represent lowess curves and the dashed black line represents intensive valence of zero. The count of observations at the teacher-year-observation level is 20,045.

Figure 3. Within-Year Distribution of Observation Scores, by Total Observations Received








Notes. $N(\text{Teacher-Year-Observation}) = 20,045$ and $N(\text{Teacher-Year}) = 9,070$. The x-axis represents the count (k) of observations received and these are grouped by total annual observations received. Circles represent mean teacher self-assessment scores and diamonds represent mean evaluator scores for count k in each annual observation group.



Appendix A. Classroom Observation Rubric

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Standards and Objectives <div style="border: 1px solid black; width: 40px; height: 40px; margin: 10px auto;"></div>	<ul style="list-style-type: none"> All learning objectives are clearly and explicitly communicated, connected to state standards, and referenced throughout lesson. Sub-objectives are aligned and logically sequenced to the lesson's major objective. Learning objectives are: (a) consistently connected to what students have previously learned, (b) known from life experiences, and (c) integrated with other disciplines. Expectations for student performance are clear, demanding, and high. There is evidence that most students demonstrate mastery of the daily objective that supports significant progress towards mastery of a standard. 	<ul style="list-style-type: none"> Most learning objectives are communicated, connected to state standards, and referenced throughout lesson. Sub-objectives are mostly aligned to the lesson's major objective. Learning objectives are connected to what students have previously learned. Expectations for student performance are clear. There is evidence that most students demonstrate mastery of the daily objective that supports significant progress towards mastery of a standard. 	<ul style="list-style-type: none"> Few learning objectives are communicated, connected to state standards, and referenced throughout lesson. Sub-objectives are inconsistently aligned to the lesson's major objective. Learning objectives are rarely connected to what students have previously learned. Expectations for student performance are vague. There is evidence that few students demonstrate mastery of the daily objective that supports significant progress towards mastery of a standard.
Motivating Students <div style="border: 1px solid black; width: 40px; height: 40px; margin: 10px auto;"></div>	<ul style="list-style-type: none"> The teacher consistently organizes the content so that it is personally meaningful and relevant to students. The teacher consistently develops learning experiences where inquiry, curiosity, and exploration are valued. The teacher regularly reinforces and rewards effort. 	<ul style="list-style-type: none"> The teacher sometimes organizes the content so that it is personally meaningful and relevant to students. The teacher sometimes develops learning experiences where inquiry, curiosity, and exploration are valued. The teacher sometimes reinforces and rewards effort. 	<ul style="list-style-type: none"> The teacher rarely organizes the content so that it is personally meaningful and relevant to students. The teacher rarely develops learning experiences where inquiry, curiosity, and exploration are valued. The teacher rarely reinforces and rewards effort.
Presenting Instructional Content <div style="border: 1px solid black; width: 40px; height: 40px; margin: 10px auto;"></div>	<p>Presentation of content always includes:</p> <ul style="list-style-type: none"> visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; examples, illustrations, analogies, and labels for new concepts and ideas; effective modeling of thinking process by the teacher and/or students guided by the teacher to demonstrate performance expectations; concise communication; logical sequencing and segmenting; all essential information; and no irrelevant, confusing, or non-essential information. 	<p>Presentation of content most of the time includes:</p> <ul style="list-style-type: none"> visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; examples, illustrations, analogies, and labels for new concepts and ideas; modeling by the teacher to demonstrate performance expectations; concise communication; logical sequencing and segmenting; all essential information; and no irrelevant, confusing, or non-essential information. 	<p>Presentation of content rarely includes:</p> <ul style="list-style-type: none"> visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; examples, illustrations, analogies, and labels for new concepts and ideas; modeling by the teacher to demonstrate performance expectations; concise communication; logical sequencing and segmenting; all essential information; and relevant, coherent, or essential information.

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Lesson Structure and Pacing <div style="border: 1px solid black; width: 40px; height: 40px; margin: 10px auto;"></div>	<ul style="list-style-type: none"> The lesson starts promptly. The lesson's structure is coherent, with a beginning, middle, and end. The lesson includes time for reflection. Pacing is brisk and provides many opportunities for individual students who progress at different learning rates. Routines for distributing materials are seamless. No instructional time is lost during transitions. 	<ul style="list-style-type: none"> The lesson starts promptly. The lesson's structure is coherent, with a beginning, middle, and end. Pacing is appropriate and sometimes provides opportunities for students who progress at different learning rates. Routines for distributing materials are efficient. Little instructional time is lost during transitions. 	<ul style="list-style-type: none"> The lesson does not start promptly. The lesson has a structure, but it may be missing closure or introductory elements. Pacing is appropriate for less than half of the students and rarely provides opportunities for students who progress at different learning rates. Routines for distributing materials are inefficient. Considerable time is lost during transitions.
Activities and Materials <div style="border: 1px solid black; width: 40px; height: 40px; margin: 10px auto;"></div>	<ul style="list-style-type: none"> Activities and materials include all of the following: <ul style="list-style-type: none"> support the lesson objectives, are challenging, sustain students' attention, elicit a variety of thinking, provide time for reflection, are relevant to students' lives, provide opportunities for student-to-student interaction, induce student curiosity and suspense, provide students with choices, incorporate multimedia and technology, and incorporate resources beyond the school curriculum texts (e.g., teacher-made materials, manipulatives, resources from museums, cultural centers, etc.). In addition, sometimes activities are game-like, involve simulations, require creating products, and demand self-direction and self-monitoring. The preponderance of activities demand complex thinking and analysis. Texts and tasks are appropriately complex. 	<ul style="list-style-type: none"> Activities and materials include most of the following: <ul style="list-style-type: none"> support the lesson objectives, are challenging, sustain students' attention, elicit a variety of thinking, provide time for reflection, are relevant to students' lives, provide opportunities for student-to-student interaction, induce student curiosity and suspense; provide students with choices, incorporate multimedia and technology, and incorporate resources beyond the school curriculum texts (e.g., teacher-made materials, manipulatives, resources from museums, cultural centers, etc.). Texts and tasks are appropriately complex. 	<ul style="list-style-type: none"> Activities and materials include few of the following: <ul style="list-style-type: none"> support the lesson objectives, are challenging, sustain students' attention, elicit a variety of thinking, provide time for reflection, are relevant to students' lives, provide opportunities for student to student interaction, induce student curiosity and suspense, provide students with choices, incorporate multimedia and technology, and incorporate resources beyond the school curriculum texts (e.g., teacher made materials, manipulatives, resources from museums, etc.).

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Questioning 	<ul style="list-style-type: none"> Teacher questions are varied and high quality, providing a balanced mix of question types: <ul style="list-style-type: none"> knowledge and comprehension, application and analysis, and creation and evaluation. Questions require students to regularly cite evidence throughout lesson. Questions are consistently purposeful and coherent. A high frequency of questions is asked. Questions are consistently sequenced with attention to the instructional goals. Questions regularly require active responses (e.g., whole class signaling, choral responses, written and shared responses, or group and individual answers). Wait time (3-5 seconds) is consistently provided. The teacher calls on volunteers and non-volunteers, and a balance of students based on ability and sex. Students generate questions that lead to further inquiry and self-directed learning. Questions regularly assess and advance student understanding. When text is involved, majority of questions are text-based. 	<ul style="list-style-type: none"> Teacher questions are varied and high quality providing for some, but not all, question types: <ul style="list-style-type: none"> knowledge and comprehension, application and analysis, and creation and evaluation. Questions usually require students to cite evidence. Questions are usually purposeful and coherent. A moderate frequency of questions asked. Questions are sometimes sequenced with attention to the instructional goals. Questions sometimes require active responses (e.g., whole class signaling, choral responses, or group and individual answers). Wait time is sometimes provided. The teacher calls on volunteers and non-volunteers, and a balance of students based on ability and sex. When text is involved, majority of questions are text-based. 	<ul style="list-style-type: none"> Teacher questions are inconsistent in quality and include few question types: <ul style="list-style-type: none"> knowledge and comprehension, application and analysis, and creation and evaluation. Questions are random and lack coherence. A low frequency of questions is asked. Questions are rarely sequenced with attention to the instructional goals. Questions rarely require active responses (e.g., whole class signaling, choral responses, or group and individual answers). Wait time is inconsistently provided. The teacher mostly calls on volunteers and high-ability students.
Academic Feedback 	<ul style="list-style-type: none"> Oral and written feedback is consistently academically focused, frequent, high quality and references expectations. Feedback is frequently given during guided practice and homework review. The teacher circulates to prompt student thinking, assess each student's progress, and provide individual feedback. Feedback from students is regularly used to monitor and adjust instruction. Teacher engages students in giving specific and high-quality feedback to one another. 	<ul style="list-style-type: none"> Oral and written feedback is mostly academically focused, frequent, and mostly high quality. Feedback is sometimes given during guided practice and homework review. The teacher circulates during instructional activities to support engagement, and monitor student work. Feedback from students is sometimes used to monitor and adjust instruction. 	<ul style="list-style-type: none"> The quality and timeliness of feedback is inconsistent. Feedback is rarely given during guided practice and homework review. The teacher circulates during instructional activities but monitors mostly behavior. Feedback from students is rarely used to monitor or adjust instruction.

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Grouping Students 	<ul style="list-style-type: none"> The instructional grouping arrangements (either whole-class, small groups, pairs, individual; heterogeneous or homogenous ability) consistently maximize student understanding and learning efficiency. All students in groups know their roles, responsibilities, and group work expectations. All students participating in groups are held accountable for group work and individual work. Instructional group composition is varied (e.g., race, gender, ability, and age) to best accomplish the goals of the lesson. Instructional groups facilitate opportunities for students to set goals, reflect on, and evaluate their learning. 	<ul style="list-style-type: none"> The instructional grouping arrangements (either whole class, small groups, pairs, individual; heterogeneous or homogenous ability) adequately enhance student understanding and learning efficiency. Most students in groups know their roles, responsibilities, and group work expectations. Most students participating in groups are held accountable for group work and individual work. Instructional group composition is varied (e.g., race, gender, ability, and age) most of the time to best accomplish the goals of the lesson. 	<ul style="list-style-type: none"> The instructional grouping arrangements (either whole-class, small groups, pairs, individual; heterogeneous or homogenous ability) inhibit student understanding and learning efficiency. Few students in groups know their roles, responsibilities, and group work expectations. Few students participating in groups are held accountable for group work and individual work. Instructional group composition remains unchanged irrespective of the learning and instructional goals of a lesson.
Teacher Content Knowledge 	<ul style="list-style-type: none"> Teacher displays extensive content knowledge of all the subjects she or he teaches. Teacher regularly implements a variety of subject-specific instructional strategies to enhance student content knowledge. The teacher regularly highlights key concepts and ideas and uses them as bases to connect other powerful ideas. Limited content is taught in sufficient depth to allow for the development of understanding. 	<ul style="list-style-type: none"> Teacher displays accurate content knowledge of all the subjects he or she teaches. Teacher sometimes implements subject-specific instructional strategies to enhance student content knowledge. The teacher sometimes highlights key concepts and ideas and uses them as bases to connect other powerful ideas. 	<ul style="list-style-type: none"> Teacher displays under-developed content knowledge in several subject areas. Teacher rarely implements subject-specific instructional strategies to enhance student content knowledge. Teacher does not understand key concepts and ideas in the discipline and therefore presents content in a disconnected manner.
Teacher Knowledge of Students 	<ul style="list-style-type: none"> Teacher practices display understanding of each student's anticipated learning difficulties. Teacher practices regularly incorporate student interests and cultural heritage. Teacher regularly provides differentiated instructional methods and content to ensure children have the opportunity to master what is being taught. 	<ul style="list-style-type: none"> Teacher practices display understanding of some student anticipated learning difficulties. Teacher practices sometimes incorporate student interests and cultural heritage. Teacher sometimes provides differentiated instructional methods and content to ensure children have the opportunity to master what is being taught. 	<ul style="list-style-type: none"> Teacher practices demonstrate minimal knowledge of students anticipated learning difficulties. Teacher practices rarely incorporate student interests or cultural heritage. Teacher practices demonstrate little differentiation of instructional methods or content.

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Thinking 	<ul style="list-style-type: none"> • The teacher thoroughly teaches two or more types of thinking: <ul style="list-style-type: none"> ○ analytical thinking, where students analyze, compare and contrast, and evaluate and explain information; ○ practical thinking, where students use, apply, and implement what they learn in real-life scenarios; ○ creative thinking, where students create, design, imagine, and suppose; and ○ research-based thinking, where students explore and review a variety of ideas, models, and solutions to problems. • The teacher provides opportunities where students: <ul style="list-style-type: none"> ○ generate a variety of ideas and alternatives, ○ analyze problems from multiple perspectives and viewpoints, and ○ monitor their thinking to insure that they understand what they are learning, are attending to critical information, and are aware of the learning strategies that they are using and why. 	<ul style="list-style-type: none"> • The teacher thoroughly teaches one or more types of thinking: <ul style="list-style-type: none"> ○ analytical thinking, where students analyze, compare and contrast, and evaluate and explain information; ○ practical thinking, where students use, apply, and implement what they learn in real-life scenarios; ○ creative thinking, where students create, design, imagine, and suppose; and ○ research-based thinking, where students explore and review a variety of ideas, models, and solutions to problems. • The teacher provides opportunities where students: <ul style="list-style-type: none"> ○ generate a variety of ideas and alternatives, and ○ analyze problems from multiple perspectives and viewpoints. 	<ul style="list-style-type: none"> • The teacher implements no learning experiences that thoroughly teach any type of thinking. • The teacher provides no opportunities where students: <ul style="list-style-type: none"> ○ generate a variety of ideas and alternatives, or ○ analyze problems from multiple perspectives and viewpoints.
Problem-Solving 	<p>The teacher implements activities that teach and reinforce three or more of the following problem-solving types:</p> <ul style="list-style-type: none"> • Abstraction • Categorization • Drawing Conclusions/Justifying Solutions • Predicting Outcomes • Observing and Experimenting • Improving Solutions • Identifying Relevant/Irrelevant Information • Generating Ideas • Creating and Designing 	<p>The teacher implements activities that teach two of the following problem-solving types:</p> <ul style="list-style-type: none"> • Abstraction • Categorization • Drawing Conclusions/Justifying Solution • Predicting Outcomes • Observing and Experimenting • Improving Solutions • Identifying Relevant/Irrelevant Information • Generating Ideas • Creating and Designing 	<p>The teacher implements no activities that teach the following problem-solving types:</p> <ul style="list-style-type: none"> • Abstraction • Categorization • Drawing Conclusions/Justifying Solution • Predicting Outcomes • Observing and Experimenting • Improving Solutions • Identifying Relevant/Irrelevant Information • Generating Ideas • Creating and Designing

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Instructional Plans	Instructional plans include: <ul style="list-style-type: none"> measurable and explicit goals aligned to state content standards; activities, materials, and assessments that: <ul style="list-style-type: none"> are aligned to state standards, are sequenced from basic to complex, build on prior student knowledge, are relevant to students' lives, and integrate other disciplines, and provide appropriate time for student work, student reflection, and lesson unit and closure; evidence that plan is appropriate for the age, knowledge, and interests of all learners; and evidence that the plan provides regular opportunities to accommodate individual student needs. 	Instructional plans include: <ul style="list-style-type: none"> goals aligned to state content standards, activities, materials, and assessments that: <ul style="list-style-type: none"> are aligned to state standards, are sequenced from basic to complex, build on prior student knowledge, and provide appropriate time for student work, and lesson and unit closure; evidence that plan is appropriate for the age, knowledge, and interests of most learners; and evidence that the plan provides some opportunities to accommodate individual student needs. 	Instructional plans include: <ul style="list-style-type: none"> few goals aligned to state content standards, activities, materials, and assessments that: <ul style="list-style-type: none"> are rarely aligned to state standards, are rarely logically sequenced, rarely build on prior student knowledge, and inconsistently provide time for student work, and lesson and unit closure; and little evidence that the plan provides some opportunities to accommodate individual student needs.
Student Work	Assignments require students to: <ul style="list-style-type: none"> organize, interpret, analyze, synthesize, and evaluate information rather than reproduce it, draw conclusions, make generalizations, and produce arguments that are supported through extended writing, and connect what they are learning to experiences, observations, feelings, or situations significant in their daily lives both inside and outside of school. 	Assignments require students to: <ul style="list-style-type: none"> interpret information rather than reproduce it, draw conclusions and support them through writing, and connect what they are learning to prior learning and some life experiences. 	Assignments require students to: <ul style="list-style-type: none"> mostly reproduce information, rarely draw conclusions and support them through writing, and rarely connect what they are learning to prior learning or life experiences.
Assessment	Assessment plans: <ul style="list-style-type: none"> are aligned with state content standards; have clear measurement criteria; measure student performance in more than three ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); require extended written tasks; are portfolio based with clear illustrations of student progress toward state content standards; and include descriptions of how assessment results will be used to inform future instruction. 	Assessment plans: <ul style="list-style-type: none"> are aligned with state content standards; have measurement criteria; measure student performance in more than two ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); require written tasks; and include performance checks throughout the school year. 	Assessment plans: <ul style="list-style-type: none"> are rarely aligned with state content standards; have ambiguous measurement criteria; measure student performance in less than two ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); and include performance checks, although the purpose of these checks is not clear.

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Expectations	<ul style="list-style-type: none"> Teacher sets high and demanding academic expectations for every student. Teacher encourages students to learn from mistakes. Teacher creates learning opportunities where all students can experience success. Students take initiative and follow through with their own work. Teacher optimizes instructional time, teaches more material, and demands better performance from every student. 	<ul style="list-style-type: none"> Teacher sets high and demanding academic expectations for every student. Teacher encourages students to learn from mistakes. Teacher creates learning opportunities where most students can experience success. Students complete their work according to teacher expectations. 	<ul style="list-style-type: none"> Teacher expectations are not sufficiently high for every student. Teacher creates an environment where mistakes and failure are not viewed as learning experiences. Students demonstrate little or no pride in the quality of their work.
Managing Student Behavior	<ul style="list-style-type: none"> Students are consistently well behaved and on task. Teacher and students establish clear rules for learning and behavior. The teacher overlooks inconsequential behavior. The teacher deals with students who have caused disruptions rather than the entire class. The teacher attends to disruptions quickly and firmly. 	<ul style="list-style-type: none"> Students are mostly well behaved and on task, some minor learning disruptions may occur. Teacher establishes rules for learning and behavior. The teacher uses some techniques, such as social approval, contingent activities, and consequences, to maintain appropriate student behavior. The teacher overlooks some inconsequential behavior, but at other times, stops the lesson to address it. The teacher deals with students who have caused disruptions, yet sometimes he or she addresses the entire class. 	<ul style="list-style-type: none"> Students are not well behaved and are often off task. Teacher establishes few rules for learning and behavior. The teacher uses few techniques to maintain appropriate student behavior. The teacher cannot distinguish between inconsequential behavior and inappropriate behavior. Disruptions frequently interrupt instruction.
Environment	The classroom: <ul style="list-style-type: none"> welcomes all members and guests, is organized and understandable to all students, supplies, equipment, and resources are all easily and readily accessible, displays student work that frequently changes, and is arranged to promote individual and group learning. 	The classroom: <ul style="list-style-type: none"> welcomes most members and guests, is organized and understandable to most students, supplies, equipment, and resources are accessible, displays student work, and is arranged to promote individual and group learning. 	The classroom: <ul style="list-style-type: none"> is somewhat cold and uninviting, is not well organized and understandable to students, supplies, equipment, and resources are difficult to access, does not display student work, and is not arranged to promote group learning.
Respectful Culture	<ul style="list-style-type: none"> Teacher-student interactions demonstrate caring and respect for one another. Students exhibit caring and respect for one another. Positive relationships and interdependence characterize the classroom. 	<ul style="list-style-type: none"> Teacher-student interactions are generally friendly, but may reflect occasional inconsistencies, favoritism, or disregard for students' cultures. Students exhibit respect for the teacher and are generally polite to each other. Teacher is sometimes receptive to the interests and opinions of students. 	<ul style="list-style-type: none"> Teacher-student interactions are sometimes authoritarian, negative, or inappropriate. Students exhibit disrespect for the teacher. Student interaction is characterized by conflict, sarcasm, or put-downs. Teacher is not receptive to interests and opinions of students.

Appendix B. Composite Teacher Effectiveness Scores

A teacher's summative annual effectiveness score is based on multiple teacher performance measures, including classroom observation, growth, and achievement scores. The growth component for teachers of tested grades/subjects is a state-issued teacher value-added measure (VAM). Growth scores for teachers of untested grades/subjects are based on school- or district-wide student outcomes (e.g., accountability test scores, school-wide VAM scores). Achievement measures are grade-, school-, or district-wide student achievement outcomes (e.g., ACT scores and high school graduation rates). Teachers receive their summative observation, growth, and achievement scores, along with their summative annual effectiveness score (*Comp-Cont*), prior to the start of the next school year since the summative effectiveness score largely determines the number of state-assigned annual observations a teacher is to receive. During the study period, growth scores comprised 35% of teachers' composite effectiveness score (*Comp-Cont*); achievement scores comprised 15%; and summative observation scores comprised 50%.

Appendix C. Extensive Margins of Negative Feedback on Teacher Exit and School

Switching

In the main text, Equation (6) estimates extensive margins for teacher mobility. In this appendix, we estimate margins for exit and school switching, where y_{jt} is one of two binary mobility indicators for teacher j in year t . The first measure indicates whether (or not) teacher j exits the state public educator market at the end of the 2018-19 school year (*Exit*); teachers who exit the district at the end of the 2017-18 school year do not contribute to the estimates directly because they appear in the data for one year only. The second measure indicates whether (or not) teacher j switches into a new school (in the same district) after the end of school year t (*Switches*). All other quantities are as described in Equation 6 in the main text. We estimate the intensive margins on *Exit* and *Switches* using Equation 7 from the main text.

The findings in Table B1 suggest that the lower probability of retention may be due to teachers seeking out new schools instead of exiting the profession. Teachers who receive negative feedback from evaluators, on average, are two percentage points more likely to exit the profession (column I) and four percentage points more likely to switch to a new school (column III) the following year, though neither estimate is statistically significant. Nonetheless, the four-percentage point estimate is approximately one-third the baseline teacher turnover rate (12%), making it a substantively large change.

We examine the school-level valence of the schools that school-switching teachers moved into relative to the school-level valence of the school switchers left. We define the difference in school-level valence scores (y_{sjt}) as $y_{sjt} = Valence_{\tilde{s}jt} - Valence_{s_{jt}}$, where $Valence_{\tilde{s}jt}$ is the year t school-level average valence intensity across all observations in school \tilde{s} , the school the j th teacher switched into for year $(t + 1)$ and $Valence_{s_{jt}}$ is the year t school-

level average valence intensity across all observations in school \vec{s} , the school the j th teacher left at the end of year t . Thus, y_{sjt} may vary across school-switching teachers leaving school \vec{s} if they switch into different schools; y_{sjt} can also vary across switchers entering the same school ($\vec{\tilde{s}}$) at the beginning of year $(t + 1)$ if switchers came from different schools. Equation A estimates the relationship between receiving negative feedback, on average, and the school-level valence intensity of receiving schools relative to sending schools (y_{sjt}), among the teachers who switch schools: (A) $y_{st} = \delta I(\text{Valence}_{jt} < 0) + \beta_1 VAM_{\vec{\tilde{s}}t} + \beta_2 VAM_{\vec{s}t} + u_{st}$, where $I(\text{Valence}_{jt} < 0)$ is the same indicator function from prior equations and indicates whether teacher j received negative feedback, on average, in year t . $VAM_{\vec{\tilde{s}}t}$ is the year t school-level average VAM score across all teachers of tested subjects in school $\vec{\tilde{s}}$, the school the j th teacher switched into for year $(t + 1)$ and $VAM_{\vec{s}t}$ is the year t school-level average VAM score across all teachers of tested subjects in school \vec{s} , the school the j th teacher left at the end of year t . We apply equation A to the sample of teachers who switched schools from year t to $(t + 1)$ only. We do not control for school-level average teacher evaluator scores or LOE because these are determined by the outcome. The coefficient δ represents the difference in school-level valence intensity scores in receiving schools relative to sending schools for the school switchers who received negative feedback, on average. Standard errors clustered at the school level. Results from equation A suggest that school switchers who received negative feedback, on average, switch into schools with relatively more positive valence intensity ($\delta = 0.05$, clustered standard error = 0.02).

Table C1. Extensive Margins of Negative Feedback on Between-Year Teacher Mobility

	I	III
	Exit	Switching
Extensive Negative Feedback	0.02 (0.02)	0.04 (0.03)
N(Teacher*Year)	3,980	3,799

Notes: Teacher-years are the unit of analysis. Each column represents a separate regression and coefficients are reported with standard errors (in parentheses) clustered at the teacher-level. Outcomes are regressed on a measure of extensive feedback and teacher-by-evaluator fixed effects and year fixed effects. The sample includes all teachers and estimates are generated by linear probability models. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$