



Do You Observe What I Observe? The Predictors and Consequences of Discordance in Teacher and Evaluator Ratings of Teacher Performance

Seth B. Hunter

George Mason University

Matthew P. Steinberg

George Mason University

Districts nationwide have revised their educator evaluation systems, increasing the frequency with which administrators observe and evaluate teacher instruction. Yet, limited insight exists on the role of evaluator feedback for instructional improvement. Relying on unique observation-level data, we examine the alignment between evaluator and teacher assessments of teacher instruction and the potential consequences for teacher productivity and mobility. We show that teachers and evaluators typically rate teacher performance similarly during classroom observations, but with significant variability in teacher-evaluator ratings. While teacher performance improves across multiple classroom observations, evaluator ratings likely overstate productivity improvements among the lowest-performing teachers. Evaluators, but not teachers, systematically rate teacher performance lower in classrooms serving higher concentrations of economically disadvantaged students. And while teacher performance improves when evaluators provide more critical feedback about teacher instruction, teachers receiving critical feedback may seek alternative teaching assignments in schools with less critical evaluation settings. We discuss the implications of these findings for the design, implementation and impact of educator evaluation systems.

VERSION: November 2022

Suggested citation: Hunter, Seth B., and Matthew P. Steinberg. (2022). Do You Observe What I Observe? The Predictors and Consequences of Discordance in Teacher and Evaluator Ratings of Teacher Performance. (EdWorkingPaper: 22-676). Retrieved from Annenberg Institute at Brown University: <https://doi.org/10.26300/97k9-br18>

**Do You Observe What I Observe?
The Predictors and Consequences of Discordance in Teacher and Evaluator
Ratings of Teacher Performance**

Seth B. Hunter
George Mason University

Matthew P. Steinberg
George Mason University

November 10, 2022

The authors thank their partners from the unnamed school district for helpful critical and positive feedback. Seth Hunter is an assistant professor of education at George Mason University (shunte@gmu.edu); Matthew Steinberg is an associate professor of education and public policy at George Mason University (msteinb6@gmu.edu). Authors are listed in alphabetical order.

Abstract

Districts nationwide have revised their educator evaluation systems, increasing the frequency with which administrators observe and evaluate teacher instruction. Yet, limited insight exists on the role of evaluator feedback for instructional improvement. Relying on unique observation-level data, we examine the alignment between evaluator and teacher assessments of teacher instruction and the potential consequences for teacher productivity and mobility. We show that teachers and evaluators typically rate teacher performance similarly during classroom observations, but with significant variability in teacher-evaluator ratings. While teacher performance improves across multiple classroom observations, evaluator ratings likely overstate productivity improvements among the lowest-performing teachers. Evaluators, but not teachers, systematically rate teacher performance lower in classrooms serving higher concentrations of economically disadvantaged students. And while teacher performance improves when evaluators provide more critical feedback about teacher instruction, teachers receiving critical feedback may seek alternative teaching assignments in schools with less critical evaluation settings. We discuss the implications of these findings for the design, implementation and impact of educator evaluation systems.

Keywords: Education policy; educator evaluation; performance feedback; school/teacher effectiveness; supervision; school administrators

Introduction

Since 2010, districts nationwide have revised their approaches to evaluating teacher performance. Recently revised systems of teacher evaluation have not only strengthened the accountability function of evaluation by linking high-stakes consequences (e.g., remediation plans and tenure revocation) to teacher performance ratings, but have also addressed the developmental function of evaluation by increasing the frequency with which teachers' instructional practices are observed and evaluated by their school administrators (Steinberg & Donaldson, 2016). Classroom observations, which account for the overwhelming share of a teacher's summative performance rating in these systems, afford school administrators greater opportunity to formatively assess and provide feedback to teachers about their instructional performance. Prior evidence from Chicago and Cincinnati finds that frequent observation and feedback of teacher practice can improve student performance (Steinberg & Sartain 2015; Taylor & Tyler 2012). Yet, more recent evidence from Tennessee, Missouri, and state-level analyses have found limited effects of teacher evaluation reform on student achievement (Bleiberg et al., 2021; Hunter & Bowser, 2021; Hunter & Springer, 2022). And, while existing evidence suggests that school administrators play a critical role in the implementation and impact of educator evaluation systems, there is limited evidence on the role of evaluator feedback for improving teacher instructional performance in the context of observation-based systems of teacher evaluation.

In this paper, we provide needed empirical insight into the role of evaluator feedback for instructional improvement by examining a critical aspect of teacher evaluation systems – whether expectations about and assessments of teacher instructional performance are aligned between school administrators and teachers (Church et al., 2019; Hunter, 2021). We examine the

extent to which teachers and their evaluators similarly rate teacher instructional performance during classroom observations and the implications of the potential divergence in these ratings for teacher performance and productivity. We address the following questions: (1) Do teachers and evaluators similarly rate teacher instructional performance? (2) What explains the observed differences in teacher and evaluator ratings of teacher performance? (3) What are the consequences of differences in teacher and evaluator ratings on teacher productivity and mobility? Indeed, if teachers and evaluators are more similar, or concordant, in their beliefs about a teacher's instructional performance, then teachers may be more likely to incorporate evaluator feedback into their instructional practices (Atwater et al., 1998; Long, 2019). Yet, if teachers overrate their own performance relative to their evaluators, then teachers may resist evaluator feedback and therefore be less likely to implement feedback recommendations and update their instructional practices (Atwater et al., 1998; Conrad & Hackmann, 2020). In contrast, if teachers underrate their performance relative to their evaluators, then they may be less likely to pursue instructional improvement (Atwater et al., 1998).

To address these questions, we rely on unique administrative data from a large urban school district located in the Southern United States on the performance evaluations of educators collected by the district's central office. Like nearly every school district in the United States, teachers in this district are evaluated by school administrators during at least one formal classroom observation of their instructional performance. Unlike any other evaluation setting (to our knowledge), this district also expects teachers to submit a self-assessment of their instructional performance after each observation. Relying on observation-level ratings of teacher performance from both teachers and their evaluators, we first examine the magnitude, distribution and within-year patterns of classroom observation scores. We pay particular

attention to score discordance – the difference between evaluator and teacher self-assessment scores – and the extent of critical feedback in cases where evaluators rate teacher instructional performance lower than teachers’ own self-assessments. We then examine the classroom-level predictors of classroom observation scores and score discordance. Indeed, given prior evidence that the composition of a teacher’s classroom explains meaningful differences in teacher performance ratings during classroom observations (Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016; Steinberg & Sartain, 2021; Whitehurst et al., 2014), we focus on the extent to which the economic and performance characteristics of a teacher’s students explain differences in teacher and evaluator assessments of teacher performance. We then examine if discordance in classroom observation scores – and the extent to which evaluators rate teacher performance more critically than do teacher self-assessments – is related to teacher productivity and patterns of post-observation teacher mobility. Notably, our study overcomes limitations in prior work where only evaluator scores were available (Cherasaro et al., 2016; Hunter & Springer, 2022; Kraft & Christian, 2021) by incorporating both evaluator and teacher assessments of teacher performance into a policy-relevant measure of performance feedback.

We find that, on average, teachers and their evaluators rate teacher performance similarly during classroom observations; yet, there is significant variability in the discordance of teacher-evaluator ratings. We further find that teacher performance improves across multiple within-year classroom observations, reflected by ratings from both teachers and their 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 evaluators provide increasingly more positive feedback about teacher performance to the lowest-performing teachers (i.e., those

who receive the most annual observations). We also show that evaluators likely inflate the fourth (and final) observation score for the lowest-performing teachers, evidence consistent with interviews with school principals suggesting that school leaders (i.e., evaluators) may intentionally overstate the evaluation ratings of teachers to avoid the time and effort associated with removing and replacing the lowest-performing teachers (Kraft & Gilmour, 2016a).

Notably, we also find that teachers teaching greater concentrations of economically disadvantaged students receive, on average, more critical feedback from their evaluators about their instructional performance than teachers in classrooms with fewer economically disadvantaged students. This finding is consistent with prior evidence that evaluators rate classroom observations lower for teachers whose students are lower-achieving, more nonwhite and more economically disadvantaged (Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016). However, we extend this prior work by showing that teachers themselves, unlike their evaluators, do not differently rate their instructional performance when teaching more academically and economically disadvantaged students. Thus, we show that evaluator bias may be an important source of discordance in teacher and evaluator ratings in the context of the classroom observation process.

Finally, when teachers receive more critical feedback about their instructional performance from their evaluators, their instructional performance – as measured by evaluator ratings during subsequent classroom observations – improves, and these improvements in teacher productivity are also reflected in improvements in a teacher's contribution to student achievement growth. At the same time, teachers who receive more critical feedback from their evaluators may also seek alternative teaching assignments in schools where administrators provide less critical feedback as part of the teacher evaluation process.

Taken together, these findings suggest that teachers and their evaluators typically view (and rate) teacher instructional performance similarly. Yet, when teacher and evaluator scores diverge and evaluators provide more critical assessments of teacher performance, teachers have the opportunity to incorporate this critical feedback into their instructional performance. In these cases, we find that critical feedback from evaluators is related to not only improvements in teacher practice, but also improvements in a teacher's contribution to student learning. And even though many school leaders may wish to avoid complicated conversations with their educators about areas for instructional improvement, our findings indicate that these conversations can have important implications for the productivity of a school's teachers and the academic performance of a school's students. In this way, the provision of critical feedback to educators can support the developmental goals of performance evaluation.

Background and Related Literature

We draw on research within and beyond K-12 education settings to conceptualize the potential discordance in performance evaluation ratings between teachers and their evaluators. Performance management and education scholars have identified three domains which shape how performance feedback might improve employee productivity. First, the policy and organizational context – including organizational culture and climate – may influence how performance feedback affects employee performance (Levy & Williams, 2004; Pichler, 2012). For example, a school espousing a culture of continuous professional learning may implement a very different teacher evaluation system than a school without such a culture (Marsh et al., 2017). Second, the potential for evaluator feedback to improve employee performance likely depends on the evaluator's human capital (Donaldson & Firestone, 2021; Levy & Williams, 2004; London & Smither, 2002). Prior work suggests that the accuracy of evaluator scoring

(Kluger & DeNisi, 1996; Levy & Williams, 2004), the type of performance feedback provided (Hattie & Timperley, 2016; Hunter & Springer, 2022; Kraft & Christian, 2021), and evaluator affectations during evaluation conferences where performance feedback is provided may influence the extent to which feedback improves performance (Donaldson, 2021; Donaldson & Firestone, 2021; Glickman et al., 2018; Levy & Williams, 2004). Third, evaluators might provide detailed performance feedback in organizations that support employee development but employees may still react to the feedback negatively and unproductively (Conrad & Hackmann, 2020; Levy & Williams, 2004; London & Smither, 2002). Although employees' responses to feedback (and, specifically, critical feedback) may depend on several psychological factors, scholars subsume these factors under “feedback orientation” – an employee's receptivity to feedback and the extent to which the employee welcomes performance-enhancing guidance (London & Smither, 2002; Quintelier et al., 2020a, 2020b). Our study focuses on an underexamined yet conceptually critical aspect of teacher performance feedback - the discordance between evaluator and teacher assessments of the teacher's instructional performance.

To what extent might there be discordance in evaluator and teacher ratings, and what might be the direction of this discordance? A substantial body of theoretical and empirical research suggests that teacher self-assessments of their performance will aim to preserve their self-image, suggesting that self-assessments may exceed evaluator scores, on average (Atwater et al., 1998). Indeed, a meta-analysis of more than 100 psychological studies concludes that self-assessments tend to surpass evaluator scores (Heidemeier & Moser, 2009). Yet, prior studies in K-12 school contexts consistently find that evaluators rate teacher performance highly, with little variability in the summative annual ratings that teachers receive (Grissom & Loeb, 2017; Kraft

& Gilmour, 2016b; Weisberg et al., 2009). Further 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, evaluators might also rate teacher performance in ways that they believe will surpass teacher expectations (i.e., self-assessments) to avoid conflicts with teachers that follow the issuance of low performance scores (Donaldson, 2021; Halverson et al., 2004; Kraft & Gilmour, 2016a). To our knowledge, the only empirical research describing the differences in performance ratings between evaluators and teachers suggests that self-assessments and evaluator scores are similar, on average (Hunter, 2021). However, Hunter (2021) compared latent constructs of teacher performance; our study extends the literature by investigating the magnitude and variability of teacher-observation-year discordance in performance ratings.

To the extent that teacher self-assessments and evaluator ratings differ, the discordance in ratings may also reflect differences in the classroom settings in which teachers teach. Indeed, prior evidence shows that teachers teaching classes with higher proportions of historically lower-achieving, economically disadvantaged, or nonwhite students receive lower performance ratings from their evaluators (Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016).¹ When teacher self assessment scores and evaluator scores similarly reflect differences in teachers' classroom settings, there should be little (to no) discordance in these ratings. However, if differences in classroom settings do not influence teacher self-assessments, these characteristics would explain, in part, the discordance in teacher self-assessments and evaluator ratings. To our knowledge, no

¹ Although other work examines whether teacher, evaluator, or system characteristics bias evaluator scores (Campbell & Ronfeldt, 2018; Grissom & Bartanen, 2022; Hunter, 2020; Steinberg & Sartain, 2021), we focus on classroom characteristics.

research has yet examined the relationships between classroom characteristics and teacher self-assessments. Without such evidence, we have little insight into whether evaluator ratings of teachers in different classroom environments introduce bias into the evaluation process, or instead reflect agreed-upon differences in teacher performance ratings between teachers and their evaluators.

In addition to the predictors of discordance in teacher and evaluator ratings, evidence suggests that the consequences of discordance depend on its direction. If evaluator scores exceed self-assessments, which we characterize as *positive feedback from evaluators*, teachers may respond by maintaining or reducing their performance effort while satisfying evaluator performance expectations (Atwater et al., 1998; Carver & Scheier, 1982). Positive feedback from evaluators may also prompt teachers to remain in their school to continue receiving praise, which may elevate their self-image (Alicke & Sedikides, 2009; Church et al., 2019; Kluger & DeNisi, 1996; Vohra & Singh, 2005). While the consequences of positive feedback are relatively straightforward, the effects of *critical feedback from evaluators* – cases where teacher self-assessment scores are higher than evaluator scores – are not. Critical feedback from evaluators might motivate teachers to improve because they learn that they are performing below evaluator expectations (Carver & Scheier, 1982; Locke & Latham, 2002). Additionally, critical feedback may improve teaching directly by offering actionable next steps that teachers can follow to achieve those goals by, for example, recommending instructional strategies and areas for instructional improvements (Hunter & Springer, 2022; Kraft & Christian, 2021; Levy & Williams, 2004). Evaluators might also point teachers to specific professional learning opportunities, such as peer coaching or workshops, to improve their instructional performance (Hunter & Springer, 2022; Kraft & Christian, 2021).

At the same time, critical feedback from evaluators may prompt negative teacher reactions, leading to unintended adverse effects (Locke & Latham, 2002; London & Smither, 2002). Teachers may respond to critical feedback from evaluators with disbelief or distrust – teachers may view the lower evaluator ratings as unfair or untrue – inducing adverse reactions to feedback recommendations that may be counterproductive to instructional improvements (Ford et al., 2017; London & Smither, 2002; Quintelier et al., 2020a, 2020b). Alternatively, teachers may accept relatively lower evaluator ratings as valid and fair performance assessments. In such cases, if evaluator ratings are substantially lower than teacher self-assessments, this may raise questions in the teacher’s mind as to whether they can realistically reach performance expectations, casting doubt on the likelihood that their efforts will result in better performance, which may lead them to abandon improvement efforts (Alicke & Sedikides, 2009; Carver & Scheier, 1982; Church et al., 2019; Hunter, 2022; Kluger & DeNisi, 1996; Vohra & Singh, 2005). And since critical feedback from evaluators may negatively affect teacher self-image (Atwater et al., 1998), it may also cause teachers to seek out schools where evaluators provide more positive (i.e., less critical) feedback.

Study Setting

The setting for this study is a large urban school district located in the Southern United States. In this study, 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 of their instructional performance (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 is an Apprentice who has taught for less than four years or a

Professional who has taught for four or more years). State- and district-specific policies stipulate that a structured post-observation conference between teachers and their evaluators should follow every formal observation. Notably, the district expects teachers to submit a self-assessment of their instructional 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 instructional 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 instructional performance, both 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.

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 planning. This training culminates in a certification exam participants must pass in order 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. 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; prior evidence indicates that administrators assign evaluators nonrandomly and strategically (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 experience conducting observations 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 instructional

improvement, and plans for instructional 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.

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.

Data & Sample

We rely on administrative data from the 2017-18 and 2018-19 school years collected by a large urban school district located in the Southern United States. 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).² 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 (*Comp-Cont*).³ The second sample also includes K-12 teachers and, when relying on prior-year measures, this sample 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 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

² Following prior work on the construction of observation-level teacher performance scores (Garrett & Steinberg, 2015; Ho & Kane, 2013; Hunter, 2021; 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.

³ 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.

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]

In the sections that follow, we first examine the magnitude, distribution and within-year patterns of classroom observation scores from both evaluators and teacher self-assessments. We pay particular attention to score discordance, which is the difference between evaluator and teacher self-assessment scores for each observation occurrence. Next, we examine the classroom-level predictors of classroom observation scores and score discordance. We then examine the potential consequences of score discordance on teacher productivity and mobility.

Discordance in Classroom Observation Scores

We construct a measure of the discordance in classroom observation scores between teacher self-assessments and evaluator assessments. 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 score discordance as follows:

$$(1) \text{Discordance}_{jkt} = Score_{jkt}^{teacher} - Score_{jkt}^{evaluator}$$

When an evaluator rates a teacher's performance as high (or higher) than the teacher's self-assessment, $Discordance_{jkt} \leq 0$, which we characterize as *positive feedback from evaluators* on a teacher's instructional performance. Alternatively, we characterize $Discordance_{jkt} > 0$ as *critical feedback from evaluators* in cases when evaluators rate teacher performance lower than the teacher's self-assessment. In this section, we describe the magnitude, distribution and within-year patterns of score discordance.

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 score discordance 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 discordance scores (SD=0.55).

<Figure 1 about here>

To better understand the underlying patterns of score discordance, Figure 2 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 – across the distribution of total annual observations received, a pattern which is consistent with the fact that lower-performing teachers are annually assigned

more classroom observations. At the same time, average 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.

<Figure 2 about here>

To formally examine the within-year patterns of observation scores presented in Figure 2, 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 $Discordance_{jkt}$. 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, and u_{jkt} is the error term. The coefficient δ represents the magnitude by which observation or discordance scores 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 dyads within each school 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 on 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 discordance between teacher and evaluator scores decreases by 0.06 points (approximately 0.11 standard deviations of discordance scores). This suggests that as teachers receive additional observations their evaluators increasingly provide more positive feedback on their instructional performance. 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, a result consistent with prior evidence on evaluator rating behavior (Kraft & Gilmour, 2016a). Thus, the degree of positive feedback received and the magnitude of score discordance is largest for teachers receiving four annual observations (-0.07 points) – the lowest performing teachers, on average – while we find no discordance in scores among higher-performing teachers receiving two annual observations (see column IV).

<Table 2 about here>

Together, evidence from Figure 2 and Table 2 (Panel A) point to important patterns in the within-year growth in teacher instructional performance. First, ratings by both teachers and their evaluators indicate that teacher performance improves across multiple classroom observations. This result is consistent with research showing not only that subjective performance ratings in the form of classroom observations reflect important information about teacher effectiveness, but also that classroom observation scores capture substantively large within-teacher improvements over time in their instructional practices (Kraft et al., 2020). Second, while the rates of growth in

teacher self-assessments are homogenous across teachers receiving different numbers of annual observations, evaluator scores exhibit higher growth rates for teachers receiving more annual observations. Furthermore, the evidence also suggests that evaluators inflate the fourth (and final) observation score for the lowest-performing teachers – those who receive four annual observations. This is in light of prior evidence from interviews with school principals suggesting that school leaders (i.e., evaluators) may intentionally overstate the evaluation ratings of teachers to avoid, for example, the time and effort associated with removing and replacing the lowest-performing teachers (Kraft & Gilmour, 2016a).

To examine the influence of the fourth (and final) evaluator score (i.e., $Score_{j4t}^{evaluator}$) on the measured growth in instructional 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 2). Then, we create a new variable – $\widehat{Score_{jkt}^{evaluator}}$ – and replace the actual evaluator score from a teacher’s fourth classroom observation ($Score_{j4t}^{evaluator}$) with the predicted fourth-observation evaluator score

$(Score_{j4t}^{evaluator})$. Next, we investigate whether (and the extent to which) the instructional 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 predicated fourth-observation evaluator score (see Table 2, Panel C); all other scores use the observed evaluator and teacher scores.

Table 2 (Panel B) presents results showing the performance growth of all teachers (and by total observations received) across just the first three classroom observations; Table 2 (Panel C) presents results showing the performance growth of all teachers when we replace the actual fourth observation score from evaluators with the predicted fourth observation score. Results reveal that the growth in teacher instructional 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 discordance 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, an empirical

result consistent with findings from interviews with school principals from Kraft and Gilmour (2016a) that document why so few teachers receive low evaluation ratings as well as why teacher evaluation ratings did not reflect evaluators' perceptions of teachers' actual instructional performance. Indeed, this result may further support prior interview evidence that evaluators (typically the school's principal) may wish to avoid giving teachers low ratings because of the intensive amount of time required to document low performance and to implement the professional development and improvement plans that low evaluation ratings typically trigger.

As a robustness check on our primary results on the magnitude of discordance 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). In alternative models, we include either month fixed effects, which control for the within-year timing of each classroom observation or domain fixed effects, which control for potential differences in classroom observation scores by the domain of teacher performance. In all cases, results indicate significant discordance in teacher and evaluator scores occurring during the fourth (and final) classroom observation (see Figure 3).

<Figure 3 about here>

Predictors of Discordance in Classroom Observation Scores

A recent literature has emerged documenting the important influence that the composition of a teacher's classroom has on teacher performance ratings based on classroom observation scores (and, notably, classroom observation scores based only on scores provided by evaluators). Indeed, teachers tend to receive lower classroom observation scores from evaluators in classrooms with lower-achieving and more economically disadvantaged students (Campbell &

Ronfeldt, 2018; Grissom & Bartanen, 2022; Steinberg & Garrett, 2016; Steinberg & Sartain, 2021; Whitehurst et al., 2014). In light of this recent evidence, we next examine the classroom-level predictors of score discordance. In doing so, we also aim to distinguish the source – evaluators, teachers or both – of any bias in classroom observation scores as a function of differences in the composition of a teacher’s classroom.

Figure 4 plots score discordance (at the teacher-year-observation level) and select classroom characteristics, including: prior-year student achievement; proportion of students who are economically disadvantaged; proportion of nonwhite students; and prior-year student disciplinary referrals. We find that teachers who teach in classrooms serving more academically and economically disadvantaged students receive, on average, more critical feedback from evaluators (i.e., score discordance is positively signed) about their instructional performance than teachers in classrooms with more advantaged students. These results are consistent with prior causal evidence showing that evaluators rate classroom observations lower for teachers whose students are lower-achieving, more nonwhite and more economically disadvantaged (Campbell & Ronfeldt, 2018; Steinberg & Garret, 2016).

<Figure 4 about here>

To estimate the conditional associations between classroom characteristics and score discordance, we estimate variants of the following specification:

$$(3) \text{Discordance}_{jkt} = X_{jt}A + W_{jt}B + V_{ekt}C + \lambda_k + \theta_{st} + u_{jkst},$$

where Discordance_{jkt} is the teacher-year-observation discordance score for observation k of teacher j in year t . 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 students

who are economically disadvantaged; proportion of nonwhite students; and prior-year student disciplinary referrals. The vector \mathbf{W} includes controls for the following teacher-level characteristics: gender, race, education level, years of experience, and prior-year observation scores. 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. The variable λ_k is an indicator variable for the count of annual observations a teacher received and controls for within-year heterogeneity in (the level and growth of) classroom observation scores across teachers receiving different numbers of total annual observations (as discussed in the prior section). All observed (and unobserved) differences between school-year cells are controlled for by θ_{st} , and the error term is u_{jkt} . We examine the sensitivity of results from Equation (3) to heterogeneity between specific teacher-by-evaluator dyads, replacing θ_{st} with teacher-by-evaluator fixed effects (η_{je}) and year fixed effects (κ_t) as prior work finds that the assignment of evaluators to teachers is nonrandom (Hunter and Rodriguez, 2021). In alternative versions of equation (3), we replace the discordance score outcome with teacher or evaluator scores to understand whether scores provided by teachers or their evaluators explain differences in the relationship between classroom characteristics and score discordance. Standard errors are clustered at the teacher level. We estimate equation (3) on the analytic sample that is restricted to teachers for whom we can link to the characteristics of the students in their classrooms (i.e., VAM sample).

Table 3 summarizes these results. Score discordance is highly correlated with the proportion of economically disadvantaged students in a teacher’s classroom. Moreover, differences in observation scores across classrooms serving different concentrations of low-income students reflect evaluator (and not teacher) scoring patterns. The proportion of

economically disadvantaged students in a classroom is significantly related to score discordance, on the order of 0.13 points (or 0.24 standard deviations) in models that control for teacher-by-evaluator and year fixed effects (results are qualitatively the same in models that leverage variation in teacher scores within the same school-year cell; i.e., school-year fixed effects). This result suggests that teachers in classrooms with more economically disadvantaged students receive more critical feedback from their evaluators than teachers in classrooms with fewer economically disadvantaged students. Notably, this relationship is driven by the fact that evaluators rate teacher performance significantly lower in classrooms with more economically disadvantaged students (-0.08 points, or 0.13 standard deviations), while teacher self-assessment scores do not reflect variation in the concentration of economically disadvantaged students. Thus, any bias in classroom observation scores that reflect compositional differences in a teacher's classroom are based on differences in how evaluators rate teacher performance, and not how teachers rate their own performance. And, while this result is consistent with prior evidence on the influence of classroom composition on classroom observation scores, our data allow us to distinguish the source of this potential bias in teachers' classroom observation scores.

The Consequences of Discordance on Teacher Productivity and Mobility

To what extent might the type (and degree) of feedback that a teacher receives from evaluators shape instructional performance? Further, does the provision of critical feedback to teachers about their performance influence teacher mobility patterns? In this section, we leverage variation within teacher-by-evaluator-by-year cells to examine how discordance (i.e., critical feedback from evaluators) is associated with within-year teacher productivity. We then leverage variation across years while controlling for differences between teacher-by-evaluator pairs to investigate the association between discordance and annualized measures of teacher productivity

and mobility. In both sets of analyses, we estimate *extensive* and *intensive* margins. Estimates of the extensive margin provide insight into how teacher productivity changes when a teacher receives critical feedback from evaluators instead of positive feedback, independent of the extent of critical feedback. Estimates of the intensive margin estimate how teacher productivity changes as a function of both the type (i.e., positive or critical) and extent of feedback (i.e., whether teacher productivity responds differently the more severe the critical feedback about teacher instruction might be).

To estimate the within-year change in teacher productivity, we first model the extensive margin of critical (versus positive) feedback from evaluators on observation scores. We estimate variants of the following specification:

$$(4) y_{jkt} = \delta I(Discordance_{jt,k-1} > 0) + \phi_{jet} + u_{jkt},$$

where y_{jkt} is $Score_{jkt}^{teacher}$, $Score_{jkt}^{evaluator}$, or $Discordance_{jkt}$. The indicator function $I(Discordance_{jt,k-1} > 0)$ takes a value of one when the j th teacher receives critical feedback from evaluators on their instructional performance during the prior observation occurrence ($k-1$) within year t . The variable ϕ_{jet} captures teacher-by-evaluator-by-year fixed effects, and as with equation (2) we estimate equation (4) using the full sample. At the extensive margin, δ represents the difference in scores between teachers who receive critical instead of positive feedback from evaluators during their prior classroom observation.

To estimate the intensive margin of within-year critical feedback from evaluators, we interact the indicator function $I(Discordance_{jt,k-1} > 0)$ with the degree of score discordance between teachers and evaluators during the prior observation occurrence, which we measure as the absolute value of $Discordance_{jt,k-1}$ (i.e., $|Discordance_{jt,k-1}|$). The inclusion of this interaction term enables insight into whether teacher productivity varies based on both the

direction of instructional feedback from evaluators (positive or critical) and the extent of score discordance between teachers and evaluators during the prior observation (i.e., observation $k-1$).

We specify the model as follows:

$$(5) \ y_{jtk} = \delta I(Discordance_{jt,k-1} > 0) + \beta_1 |Discordance_{jt,k-1}| \\ + \beta_2 [I(Discordance_{jt,k-1} > 0) \cdot |Discordance_{jt,k-1}|] + \phi_{jet} + u_{jtk}$$

In equation (5), β_1 represents the change in teacher productivity associated with a unit increase in score discordance among teachers who received positive feedback from evaluators during their prior classroom observation. β_2 captures any differential change in teacher productivity as a function of the extent of score discordance between teachers and evaluators during the prior observation (i.e., observation $k-1$) among teachers who received critical feedback from evaluators during their prior classroom observation relative to teachers who received positive feedback during their prior observation occurrence. The linear combination of parameters $\beta_1 + \beta_2$ represents the total change in productivity associated with a unit change in score discordance among teachers who received critical feedback from evaluators during the prior observation. The inclusion of teacher-by-evaluator-by-year fixed effects (ϕ_{jet}) restricts all comparisons to the same teacher-by-evaluator dyads within the same year, accounting for all observed and unobserved time-invariant factors among teacher-by-evaluator pairs that may be correlated with both teacher productivity as well as score discordance. Standard errors are clustered at the teacher level.

Table 4 summarizes these results. Teachers who receive critical feedback from evaluators (instead of positive feedback) about their instructional performance during the prior observation are rated, on average, higher by their evaluators but rate their own performance lower in the subsequent observation. At the extensive margin, critical feedback from evaluators in the prior

observation is associated with a subsequent rise in evaluator scores (0.07 units, 0.11 standard deviations, see Panel A, column III), suggesting that critical feedback from evaluators is associated with improvements in within-year teacher productivity. However, teacher self-assessment scores are negatively associated with the receipt of critical feedback from evaluators in the prior observation (-0.09, -0.15 standard deviations, see Panel A, column II). The negative association with teacher self-assessments could reflect a recalibration of teachers' understandings of instructional performance expectations, as teachers who were told that they scored themselves too highly (relative to evaluator scores) in the prior observation lowered their self-assessment during the following observation. Alternatively, the decline in teacher self-assessment scores following the receipt of critical feedback from evaluators may reflect demoralization (Alicke & Sedikides, 2009; Kluger & DeNisi, 1996; Vohra & Singh, 2005).

[Insert Table 4 here]

We also find that teacher instructional performance responds to the intensity of critical feedback from evaluators. A unit increase in positive feedback from evaluators during the prior observation is associated with a 0.07 unit (0.11 standard deviation, see column V) increase in teacher self-assessment and a 0.25 unit (0.39 standard deviation, see column VI) decrease in evaluator scores during the subsequent observation.⁴ In contrast, a unit increase in critical feedback from evaluators is associated with a 0.14 unit decline in teacher self-assessments (0.23 standard deviations, see column V) and 0.07 increase in evaluator scores (0.11 standard deviation, see column VI). Thus, this evidence suggests that teachers respond to more positive feedback from evaluators by reducing their instructional effort, reflected in a decrease in

⁴ Although these associations are large, they are based on substantively large changes in feedback; a unit increase in score discordance is equivalent to 1.82 standard deviations of score discordance.

evaluator scores. Conversely, critical feedback from evaluators appears to result in instructional improvement (captured by higher evaluator scores) while potentially introducing demoralization (captured by lower teacher self-assessment scores). These results are consistent with the empirical and conceptual arguments advanced by Carver and Scheier (1982), Hunter and Springer (2022), and Locke and Latham (2002), among others (see the Background and Related Literature section for more details).

Evidence in Table 4 further suggests that if critical feedback from evaluators (at the extensive and intensive margins) induces teacher demoralization, it is not detrimental to the point of reducing teacher performance, at least as measured by a teacher’s evaluator. However, demoralization may be related to greater mobility among teachers who receive more critical feedback from their evaluators. Furthermore, if critical feedback from evaluators is associated with improvements in teacher performance as measured by evaluator ratings of a teacher’s instructional performance during classroom observations, we also expect a corresponding increase in other measures of teacher performance measures, such as those based on student achievement scores. Notably, results from Table 4 Panel B show that the relationship between critical feedback from evaluators and evaluator scores is not qualitatively different among the subsample of teachers who also receive VAM scores.

Between-Year Productivity and Mobility. Since teacher mobility and VAM scores (based on student achievement) are annualized outcomes, we estimate teacher-year variants of equations 4 and 5. The following equation estimates extensive margins for the annualized outcomes:

$$(6) \ y_{jt} = \delta I(Discordance_{jt} > 0) + X_{jt}A + W_{jt}B + \lambda_k + \theta_{te} + u_{jt}.$$

where y_{jt} is the state-issued VAM score or a binary mobility indicator for teacher j in year t .⁵ 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 discordance scores to the same level: $Discordance_{jt}$ is the j th teacher's average discordance across all observations within year t and $I(Discordance_{jt} > 0)$ takes a value of one when teacher j 's average feedback from all observations in year t is critical feedback. The vectors \mathbf{X} , \mathbf{W} , λ_k , and the term u_{jt} refer to the same quantities as in Equation 3, and we control for teacher-by-evaluator heterogeneity via θ_{te} , which are teacher-by-evaluator fixed effects. The coefficient of interest, δ , represents the average change in an outcome measured at the end of year t for a unit increase in average discordance during school year t . Standard errors are clustered at the teacher level.

We estimate the intensive margins of annualized teacher productivity and mobility with Equation 7:

$$(7) y_{jt} = \delta I(Discordance_{jt} > 0) + \beta_1 |Discordance_{jt}| + \beta_2 [I(Discordance_{jt} > 0) \cdot |Discordance_{jt}|] + X_{jt}A + W_{jt}B + \lambda_k + \theta_{te} + u_{jt},$$

where all variables refer to the same quantities as described in Equation 6 and the interpretation of coefficients is similar to those in equation 4. When VAM is the outcome, equations 6 and 7 rely on the VAM sample; estimates of teacher mobility rely on the full sample. Standard errors are clustered at the teacher level.

⁵ We do not examine summative observation scores as an outcome for econometric reasons. A teacher's summative observation score is their average evaluator-issued observation score across all observations. $Discordance_{jt}$ is a linear function of all evaluator-assigned observation scores. Regressing summative observation scores on $Discordance_{jt}$ effectively amounts to regressing summative observation scores on itself.

Table 5 summarizes these results. Teachers who receive critical feedback on their instructional performance throughout the school year have higher end-of-year VAM scores, on average, than teachers who receive positive feedback on their instructional performance, on the order of 1.30 units (0.20 standard deviations of VAM; see column I, Table 5). These results provide additional support for the association between critical feedback from evaluators on a teacher's instructional performance and improvements in teacher productivity, as shown with classroom observation scores (see Table 4). However, we do not find that changes in VAM depend on the degree of average discordance (see column II, Table 5).

[Insert Table 5 here]

The relationship between annual score discordance and teacher retention suggest that critical feedback may improve teacher productivity at the expense of teacher turnover (see Table 5). Teachers who receive critical feedback from evaluators, 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 (see column III). However, estimates in column IV suggest that a unit increase in positive average feedback from evaluators is associated with a nearly 20-point higher probability that the teacher will remain in their school the following year. Although the change in retention is substantively large, so is the unit change in positive average feedback from evaluators, which is equivalent to 2.08 standard deviations of average discordance scores. Conversely, teachers receiving a unit more critical average feedback from evaluators are 13 percentage points less likely to remain in their school (column IV); although this difference is not statistically significant, it accounts for more than the baseline turnover rate. Results from additional mobility analyses on the probability that teachers exit the district or switch into a new school in the same district are consistent with these retention results

(see Appendix B). Together, the within-year and between-year analyses suggest that while critical feedback from evaluators may improve teacher productivity, it may also push teachers out of their school.

If the teachers who exit their schools for another do so to avoid critical feedback from evaluators, we expect to see these teachers switching to schools that are less likely to provide critical feedback. Indeed, auxiliary analyses support this hypothesis.⁶ We hypothesize that if teachers aimed to enter relatively less critical schools in year $(t + 1)$, they would have sought out less critical school placements. We tested this hypothesis by comparing the year t school-level average discordance of school-switching teachers' sending and receiving schools. Among the sample of teachers who switched schools, those who received critical feedback from evaluators, on average and regardless of degree (i.e., critical average feedback on the extensive margin), switched into receiving schools with 0.05 units (0.09 standard deviations) less discordance than their sending school. Stated differently, teachers who received critical feedback from evaluators

⁶ We examine the school-level discordance scores of the schools school-switching teachers moved into relative to the school-level discordance scores of the school switchers left. We define the difference in school-level discordance scores (y_{sjt}) as $y_{sjt} = \text{Discordance}_{\tilde{s}jt} - \text{Discordance}_{\bar{s}jt}$, where $\text{Discordance}_{\tilde{s}jt}$ is the year t school-level average discordance across all observations in school \tilde{s} , the school the j th teacher switched into for year $(t + 1)$ and $\text{Discordance}_{\bar{s}jt}$ is the year t school-level average discordance across all observations in school \bar{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 \bar{s} if they switch into different schools; y_{sjt} can also vary across switchers entering the same school (\tilde{s}) at the beginning of year $(t + 1)$ if switchers came from different schools. Equation A estimates the relationship between receiving critical feedback, on average, and the school-level discordance of receiving schools relative to sending schools (y_{sjt}), among the teachers who switch schools: (A) $y_{st} = \delta I(\text{Discordance}_{jt} > 0) + \beta_1 \text{VAM}_{\tilde{s}t} + \beta_2 \text{VAM}_{\bar{s}t} + u_{st}$, where $I(\text{Discordance}_{jt} > 0)$ is the same indicator function from equations 6 and 7 and indicates whether teacher j received critical feedback, on average, in year t . $\text{VAM}_{\tilde{s}t}$ is the year t school-level average VAM score across all teachers of tested subjects in school \tilde{s} , the school the j th teacher switched into for year $(t + 1)$ and $\text{VAM}_{\bar{s}t}$ is the year t school-level average VAM score across all teachers of tested subjects in school \bar{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 observation scores or LOE because these are determined by the outcome. The coefficient δ represents the difference in school-level discordance scores in receiving schools relative to sending schools for the school switchers who received critical feedback, on average. Standard errors clustered at the school level. Results from equation A suggest that school switchers who received critical feedback, on average, switch into less discordant schools ($\delta = -0.05$, clustered standard error = 0.02).

and switched schools within the district typically entered schools where teachers were provided with, on average, more positive (i.e., less critical) feedback from their evaluators on their instructional performance.

Discussion

Recent evidence has shown that teacher evaluation systems can positively impact student achievement in specific district settings (Steinberg & Sartain, 2015; Taylor & Tyler, 2012), while larger-scale studies are less sanguine about the potential impact of teacher evaluation (Bleiberg et al., 2021; Hunter & Bowser, 2021). These mixed findings raise questions about the instructional conditions and, more specifically, the developmental function of evaluation systems for supporting improvements in teacher instructional performance (Cohen et al., 2020; Donaldson, 2021; Donaldson & Woulfin, 2018; Hunter & Springer, 2022). In this paper, we rely on unique data from the evaluation system of a large Southern school district to explore one potential mechanism implicit in educator evaluation for teacher development and student improvement – post-observation teacher performance feedback (Donaldson, 2021; Hunter & Springer, 2022). Specifically, we examine the alignment between evaluator ratings and teacher self-assessments of teacher performance (i.e., score discordance), and the potential implications of critical evaluator feedback for improving teacher productivity.

We find that teachers and their evaluators rate teacher performance similarly following the typical classroom observation, but with substantial observation-level variability in teacher and evaluator scores. While these results differ from study contexts outside of K-12 education in which employee self-assessments tend to exceed evaluator ratings (Tetlock & Manstead, 1985; Vansteenkiste et al., 2007), teachers and evaluators in our study context may typically agree about employee performance for three reasons. First, the performance rubric used in this study

delineates teacher performance across multiple proficiency levels and relies on detailed descriptors that link observable evidence about teacher instructional performance with the ratings levels of each performance indicator. Indeed, Danielson's Framework for Teaching and many other standards-based rubrics in K-12 settings are similarly designed (Danielson, 1996; Steinberg & Donaldson, 2016), and the clarity with which these rubrics describe multiple dimensions of teacher performance may increase their reliability when used by teachers and evaluators (Wherry & Bartlett, 1982). Second, the study district encourages schools to hold norming sessions during which evaluators and teachers discuss the design and application of the evaluation rubric as a way to develop a shared understanding of not only how the rubric measures teacher performance but also what evaluators may expect to observe when evaluating and rating teacher practice during classroom observations. Third, prior evidence finds that evaluators wish to avoid conflict with teachers in evaluation contexts (Kraft & Gilmour, 2016a); as such, evaluators may intentionally work with their staff to develop shared understandings of rubric use to avoid differences in how teachers and evaluators observe and rate teacher performance.

We further find that as teachers receive more annual observations of their performance, evaluators provide increasingly positive feedback to teachers. Notably, the within-year evaluator rating gradient increases with the number of observations teachers receive, while teacher self-assessment scores do not vary by the number of annual observations a teacher receives. To the extent that these evaluator rating patterns reflect actual instructional growth in teacher performance, particularly among the lowest-performing teachers, these patterns are consistent with Kraft et al. (2022) who find larger within-teacher returns to experience among less experienced teachers. Because teachers in the study setting who receive more observations are less effective, the heterogeneity in evaluator ratings may also suggest that the higher growth rate

among the least effective teachers narrows the performance gap between the least and most effective teachers. However, this heterogeneity in evaluator ratings may also reflect something other than actual performance improvements. Evaluators may issue increasingly higher ratings (and increasingly more positive feedback) over time to their lowest-performing teachers in order to avoid conflict with teachers and the low evaluation ratings that might trigger teacher dismissal (Kraft & Gilmour, 2016a). Notably, our findings indicate that evaluators systematically inflate the final within-year observation score issued to the least effective teachers. Evaluators might point to inflated end-of-year positive feedback as a reason why their least effective teachers should remain in their schools and build on their end-of-year performance ‘improvement’ next school year. Such evaluator behaviors would be consistent with the evaluator belief that all teachers can improve, rationalizing their hesitance to dismiss the least effective teachers (Rodriguez & Hunter, 2021). Although the data do not allow us to conclude whether the evaluator ratings reflect real or inflated performance growth (survey or qualitative research may be well-suited to examining such mechanisms), the homogeneity in teacher self-assessments across teachers at different performance levels suggests the latter.

Our investigation into the classroom characteristics predicting discordance suggests that student economic disadvantage is the greatest source of disagreement between teachers and evaluators. Teacher-by-evaluator fixed effect models predict that teachers receive a higher degree of critical feedback from evaluators in classrooms serving greater concentrations of economically disadvantaged students. While the concentration of economically disadvantaged students do not shape teacher self-assessments, evaluators issue systematically lower ratings for classrooms with higher proportions of economically disadvantaged students. Consistent with prior evidence (Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016), our findings suggest

that teachers in classrooms with high percentages of economically disadvantaged students receive lower scores, likely reflecting bias in evaluator ratings. Thus, evaluators may require more intensive support to quell this bias; however, recent experimental evidence suggests that workshop-based support may be insufficient (Kraft & Christian, 2021), implying the need for ongoing evaluator coaching. Evaluators might also apply a “consider the opposite” strategy, a scalable, low-cost, and low-intensity intervention in which evaluators describe two reasons why their initial rating is inaccurate, effectively prompting more robust scoring rationales; experimental evidence in public management settings finds that this strategy reduces multiple sources of bias in evaluator ratings (Nagtegaal et al., 2020).

Our findings on the consequences of score discordance suggest that critical feedback from evaluators may improve teacher productivity, which prior work suggests may operate via three mechanisms. First, evaluators providing critical feedback might improve teacher productivity directly by recommending specific teaching practices (Hunter & Springer, 2022). Second, evaluators might improve teacher productivity indirectly by pointing the teacher to suitable professional learning opportunities (e.g., peer mentors, resources) that improve teaching (Hunter & Springer, 2022). Third, when evaluators tell teachers that they performed worse than their self-assessment, it may motivate self-directed teacher improvement, independent of evaluator recommendations (Carver & Scheier, 1982; Kraft & Christian, 2021; Locke & Latham, 2002). In addition to these mechanisms, a novel study examining the written feedback teachers receive suggests that the extent to which feedback improves teacher productivity depends on specific feedback characteristics (e.g., sets teacher performance goals) and teachers’ familiarity with their performance expectations (Hunter & Springer, 2022). At the same time, Hunter and Springer (2022) found no relationships between teacher productivity and several feedback

characteristics that prior survey-based research and studies outside K-12 settings suggested ought to improve teacher performance. Collectively, evidence from the study herein and recent work underscores the need for more research examining the influence of evaluator feedback on teacher performance, its mediators and suppressants, and the environments and conditions in which such feedback may be performance-enhancing for educators.

From a policy perspective, pre- and in-service principal training programs might design and incorporate professional development opportunities to facilitate skill development so that evaluators and school administrators can provide candid and critical feedback to teachers during classroom observations. Indeed, our findings suggest that evaluators who forego the provision of critical feedback may also forego an opportunity to improve teaching and teachers' contributions to student achievement (Hunter & Springer, 2022), underscoring the need to overcome principals' hesitancy around critical feedback (Kraft & Gilmour, 2016a). However, improving the candor and quality of feedback that evaluators may offer to educators may be difficult, as prior work suggests that principals prefer to avoid confrontations in the context of teacher evaluations (Kraft & Gilmour, 2016a) and that evaluators tend to provide simplistic feedback (Hunter & Springer, 2022). Further, policymakers and school leaders might also consider incorporating teacher self-assessments into the design and implementation of educator evaluation systems. Doing so will afford new insights into how teachers rate their own performance relative to their evaluators, and also provide greater insight into the process by which evaluator feedback is aligned with teacher beliefs and may be related to improving teacher instructional performance. Ultimately, policy efforts should recognize the potential that critical feedback can play in driving improvements in teacher instructional practice, and work to develop the capacity of principals to more effectively communicate areas for instructional improvement to their

teachers and develop opportunities for educators to share their assessments of their own instruction. Such efforts will shed new light on the mechanisms for driving improvements to teacher quality and student achievement.

References

- 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.
- Atwater, L. E., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self-Other Agreement: Does It Really Matter? *Personnel Psychology*, 51, 577–598.
- Bleiberg, J., Brunner, E., Harbatkin, E., Kraft, M. A., & Springer, M. G. (2021). *The Effect of Teacher Evaluation on Achievement and Attainment: Evidence from Statewide Reforms* (Working Paper No. 21–496; EdWorkingPaper). Annenberg Institute at Brown University. <https://www.edworkingpapers.com/ai21-496>
- 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.
- Cherasaro, T. L., Brodersen, R. M., Reale, M. L., & Yanoski, D. C. (2016). *Teachers' responses to feedback from evaluators: What feedback characteristics matter?* (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.
- Cohen, J., Loeb, S., Miller, L. C., & Wyckoff, J. H. (2020). Policy Implementation, Principal Agency, and Strategic Action: Improving Teaching Effectiveness in New York City Middle Schools. *Educational Evaluation and Policy Analysis*, 42(1), 134–160. <https://doi.org/10.3102/0162373719893338>

- Conrad, D. L., & Hackmann, D. G. (2020). Implications of Illinois Teacher Evaluation Reforms: Insights from Principals. *Leadership and Policy in Schools*, 1–20.
<https://doi.org/10.1080/15700763.2020.1802761>
- Danielson, C. (n.d.). *The Many Faces of Leadership*. 7.
- Donaldson, M. L. (2021). *Multidisciplinary Perspectives on Teacher Evaluation: Understanding the Research and Theory* (1st ed.). Routledge.
- Donaldson, M. L., & Firestone, W. (2021). Rethinking teacher evaluation using human, social, and material capital. *Journal of Educational Change*, 22(4), 501–534.
<https://doi.org/10.1007/s10833-020-09405-z>
- Donaldson, M. L., & Woulfin, S. (2018). From Tinkering to Going “Rogue”: How Principals Use Agency When Enacting New Teacher Evaluation Systems. *Educational Evaluation and Policy Analysis*, 40(4), 531–556. <https://doi.org/10.3102/0162373718784205>
- Ford, T. G., Sickles, M. E. V., Clark, L. V., Fazio-Brunson, M., & Schween, D. C. (2017). Teacher Self-Efficacy, Professional Commitment, and High-Stakes Teacher Evaluation Policy in Louisiana. *Educational Policy*, 31(2), 202–248.
- 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>
- Glickman, C., Gordon, S., & Ross-Gordon, J. (2018). *Supervision and Instructional Leadership* (10th ed.). Pearson.

- 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., & 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.
- Hattie, J., & Timperley, H. (2016). The Power of Feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- 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. *Bill & Melinda Gates Foundation*.
http://pitt.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwY2BQAB2JnphobmqY1JSSbGSRaJlmZJgErCfMDBKNTc0sUIBWVCKV5m5CDEypeaIMMm6uIc4euqAJi_gC

yJkL8a4uwKaFpam5oRgDC7BfnCrOwJoGjB8gDSwzxYH6xRk4lizDQy0ivf0gXCEYV
68YvH9Jr7BEHFhEg6NX1ljPAACqsie0

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. (2021). Do You Mean What I Mean? Comparing Teacher Performance Self-Scores and Evaluator-Generated Scores. *Journal of Education Human Resources*, e20200026.

<https://doi.org/10.3138/jehr-2020-0026>

Hunter, S. B. (2022). High-leverage teacher evaluation practices for instructional improvement. *Educational Management Administration & Leadership*, 174114322211129.

<https://doi.org/10.1177/17411432221112995>

Hunter, S. B., & Bowser, K. (2021). Identifying the Effects of Next-Generation Teacher Evaluation on Student Achievement in Rural Districts: Evidence from Missouri. *11.03 Teacher Evaluation Systems. Educator Preparation, Professional Development, Performance, and Evaluation*. Association for Education Finance and Policy Annual Conference, Virtual.

https://education.gmu.edu/assets/docs/educational_leadership/HunterBowser_Introducing.pdf

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>
- 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. *Educational Researcher*, 1–31.
- Kraft, M. A., Papay, J. P., & Chi, O. (2020). Teacher Skill Development: Evidenced from Performance Ratings by Principals. *Journal of Policy Analysis and Management*, 39(2), 315–347.
- Levy, P. E., & Williams, J. R. (2004). The Social Context of Performance Appraisal: A Review and Framework for the Future. *Journal of Management*, 30(6), 881–905. <https://doi.org/10.1016/j.jm.2004.06.005>
- Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57(9), 705–717. <https://doi.org/10.1037/0003-066X.57.9.705>

- 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)
- Long, A. (2019). *TEACHERS' PERCEPTIONS AND EXPERIENCES WITH A REFORMED TEACHER EVALUATION SYSTEM: CONDITIONS NECESSARY FOR CHANGING PRACTICE* [Dissertation, The Pennsylvania State University]. https://etda.libraries.psu.edu/files/final_submissions/18672
- Marsh, J. A., Bush-Mecenas, S., Strunk, K. O., Lincove, J. A., & Huguet, A. (2017). Evaluating Teachers in the Big Easy: How Organizational Context Shapes Policy Responses in New Orleans. *Educational Evaluation and Policy Analysis*, 39(4), 539–570. <https://doi.org/10.3102/0162373717698221>
- 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/Jan2013_ACompositeEstimatorofEffectiveTeachingResearchPaper.pdf
- Nagtegaal, R., Tummers, L., Noordegraaf, M., & Bekkers, V. (2020). Designing to Debias: Measuring and Reducing Public Managers' Anchoring Bias. *Public Administration Review*, 80(4), 565–576. <https://doi.org/10.1111/puar.13211>
- Pichler, S. (2012). The social context of performance appraisal and appraisal reactions: A meta-analysis. *Human Resource Management*, 51(5), 709–732. <https://doi.org/10.1002/hrm.21499>
- Quintelier, A., De Maeyer, S., & Vanhoof, J. (2020a). 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>
- Quintelier, A., De Maeyer, S., & Vanhoof, J. (2020b). 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>
- 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>
- 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>

- Tetlock, P. E., & Manstead, A. S. R. (1985). Impression management versus intrapsychic explanations in social psychology: A useful dichotomy? *Psychological Review*, 92, 59–77.
- Vansteenkiste, M., Neyrinck, B., Niemiec, C. P., Soenens, B., Witte, H., & Broeck, A. (2007). On the relations among work value orientations, psychological need satisfaction and job outcomes: A self-determination theory approach. *Journal of Occupational and Organizational Psychology*, 80(2), 251–277. <https://doi.org/10.1348/096317906X111024>
- 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
- Wherry, R. J., & Bartlett, C. J. (1982). THE CONTROL OF BIAS IN RATINGS: A THEORY OF RATING. *Personnel Psychology*, 35(3), 521–551. <https://doi.org/10.1111/j.1744-6570.1982.tb02208.x>
- Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating Teachers with Classroom Observations Lessons Learned in Four Districts* (pp. 1–27).

Tables & Figures

Table 1. Teacher Characteristics

		Annual Classroom Observations			
	All Teachers	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 proportion, except for *Experience*, which reports mean (standard deviation). *Masters+* includes teachers who have more than a master's degree. In Panel B, each cell reports mean (standard deviation) 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, an integer score ranging from -100 – 60. The count of teacher-year observations includes teachers with nonmissing 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
	Discordance	Teacher Scores	Evaluator Scores	Discordance	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(Teacher-Year- 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(Teacher-Year- 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(Teacher-Year- Obs)	20,045	20,045	20,045	20,045	20,045	20,045
N(Teacher-Year)	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. Classroom Predictors of Discordance, Teacher, and Evaluator Scores

	I	II	III	IV	V	VI
	Discordance	Teacher Scores	Evaluator Scores	Discordance	Teacher Scores	Evaluator Scores
Proportion Economically Disadvantaged	0.07* (0.03)	0.01 (0.03)	-0.05* (0.03)	0.13* (0.05)	0.04 (0.06)	-0.082* (0.040)
Proportion Nonwhite	0.00 (0.04)	0.01 (0.04)	0.02 (0.03)	-0.05 (0.04)	0.03 (0.05)	0.084 (0.049)
Prior-Year Achievement	0.01 (0.02)	0.04 (0.02)	0.02 (0.02)	0.01 (0.05)	-0.01 (0.07)	-0.02 (0.06)
Prior-Year Office Referrals	-0.01 (0.02)	-0.02 (0.02)	-0.01 (0.02)	-0.01 (0.02)	0.03 (0.03)	0.04 (0.03)
School-Year FE	X	X	X			
Evaluator-Teacher FE, Year FE				X	X	X
N(Teacher*Year)	1,140	1,140	1,140	1,140	1,140	1,140
N(Teacher*Year*Observation)	2,287	2,287	2,287	2,287	2,287	2,287

Notes: Each column represents a separate regression and coefficients are reported with standard errors (in parentheses) clustered at the teacher-level. The sample includes the VAM sample. All regressions control for teacher and evaluator gender, race/ ethnicity, education level, years of experience (in school-year fixed effect models), and prior-year observation score, in addition to standardized classroom characteristics listed. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 4. Extensive and Intensive Margins of Discordance on Within-Year Change in Teacher Observation Scores

	I	II	III	IV	V	VI
	Discordance	Teacher Scores	Evaluator Scores	Discordance	Teacher Scores	Evaluator Scores
Panel A. Full Sample						
$\delta I(Discordance_{jt,k-1} > 0)$	-0.16*** (0.02)	-0.09*** (0.02)	0.07*** (0.02)	-0.01 (0.02)	-0.02 (0.02)	-0.01 (0.03)
$\beta_1 Discordance_{jt,k-1} $				0.33*** (0.04)	0.07* (0.04)	-0.25*** (0.04)
$\beta_2 I(Discordance_{jt,k-1} > 0)$ $\cdot Discordance_{jt,k-1} $				-0.54*** (0.05)	-0.22*** (0.05)	0.32*** (0.05)
$\beta_1 + \beta_2$				-0.21*** (0.04)	-0.14*** (0.03)	0.07* (0.04)
N(Teacher*Year)	6,661	6,661	6,661	6,661	6,661	6,661
N(Teacher*Year*Observation)	10,837	10,837	10,837	10,837	10,837	10,837
Panel B. VAM Sample						
$\delta I(Discordance_{jt,k-1} > 0)$	-0.19*** (0.04)	-0.13*** (0.04)	0.05 (0.04)	-0.04 (0.05)	-0.05 (0.05)	-0.01 (0.05)
$\beta_1 Discordance_{jt,k-1} $				0.28* (0.11)	0.11 (0.10)	-0.17 (0.09)
$\beta_2 I(Discordance_{jt,k-1} > 0)$ $\cdot Discordance_{jt,k-1} $				-0.53*** (0.13)	-0.30* (0.12)	0.23* (0.12)
$\beta_1 + \beta_2$				-0.25*** (0.07)	-0.19** (0.07)	0.06 (0.07)
N(Teacher*Year)	995	995	995	995	995	995
N(Teacher*Year*Observation)	1,811	1,811	1,811	1,811	1,811	1,811

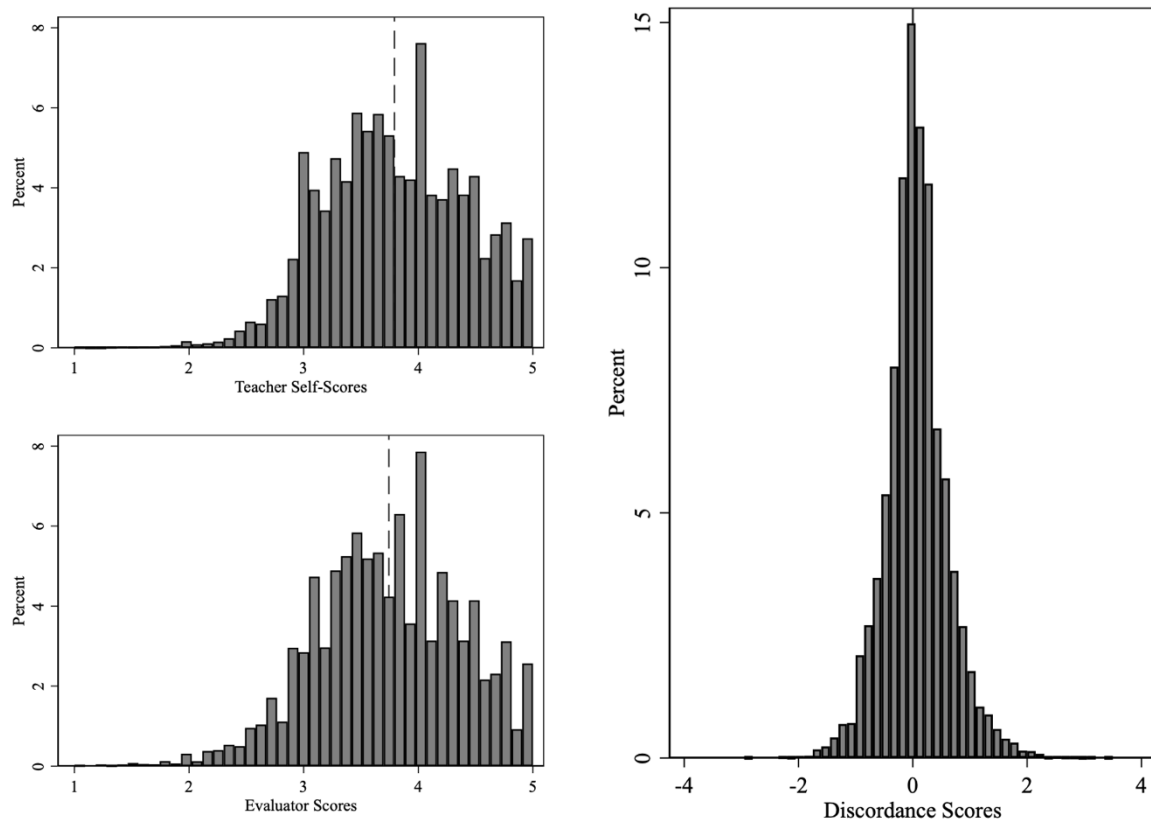
Notes: Each column (within a panel) represents a different regression and coefficients are reported with standard errors (in parentheses) clustered at the teacher-level. Observation scores are regressed on discordance measures and teacher-by-evaluator-by-year FE. Panel A is the subset of teachers from 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 5. Extensive and Intensive Margins of Discordance on Teacher VAM and Teacher Retention

	I	II	III	IV
	VAM		Retained	
$\delta I(Discordance_{jt,k-1} > 0)$	1.30*	1.92	-0.05	0.01
	(0.65)	(1.44)	(0.03)	(0.04)
$\beta_1 Discordance_{jt,k-1} $		0.80		0.19*
		(3.42)		(0.08)
$\beta_2 I(Discordance_{jt,k-1} > 0) \cdot Discordance_{jt,k-1} $		-1.63		-0.32**
		(4.07)		(0.11)
$\beta_1 + \beta_2$		-0.84		-0.13
		(2.16)		(0.07)
N(Teacher*Year)	537	537	3,980	3,980

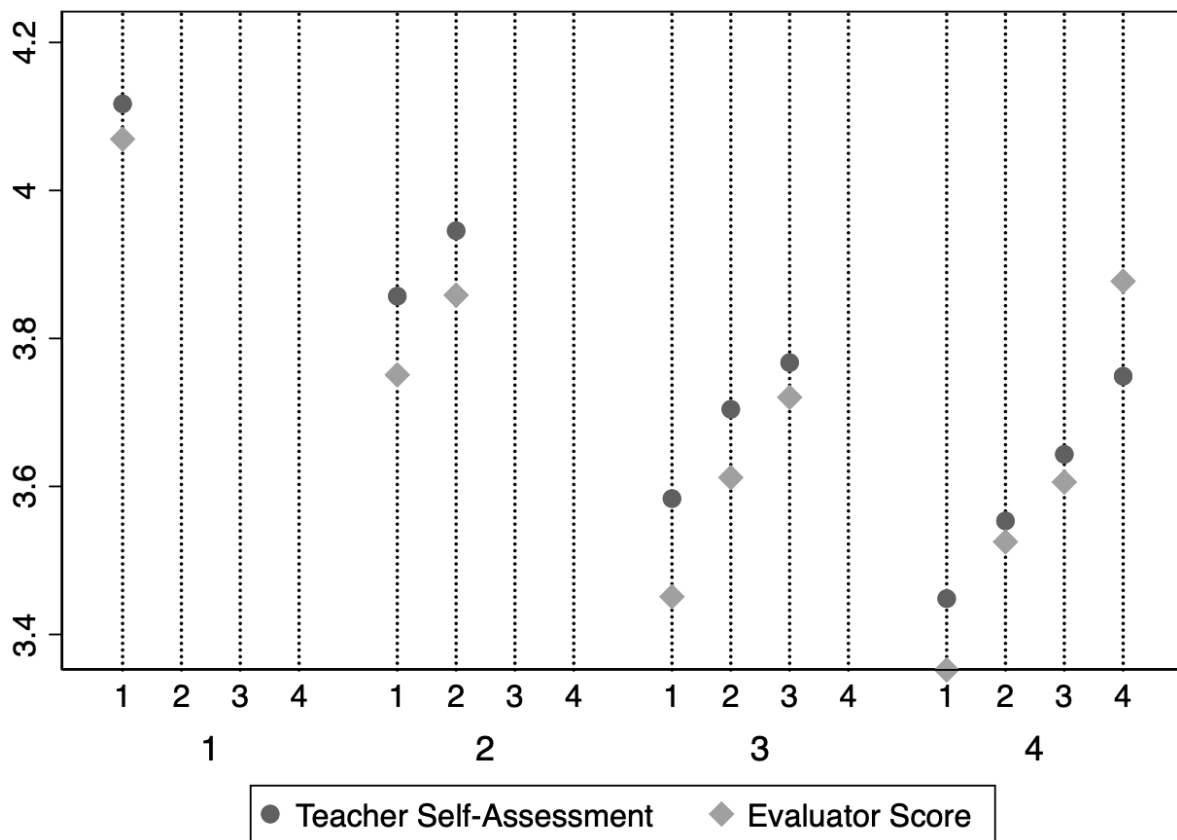
Notes: Each column represents a separate regression and coefficients are reported with standard errors (in parentheses) clustered at the teacher-level. Teacher-years are the unit of analysis. Outcomes are regressed on a discordance measure and teacher-by-evaluator fixed effects and year fixed effects. Columns I and II are limited to teachers of tested subjects with VAM scores. Columns III and IV include all teachers and are 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



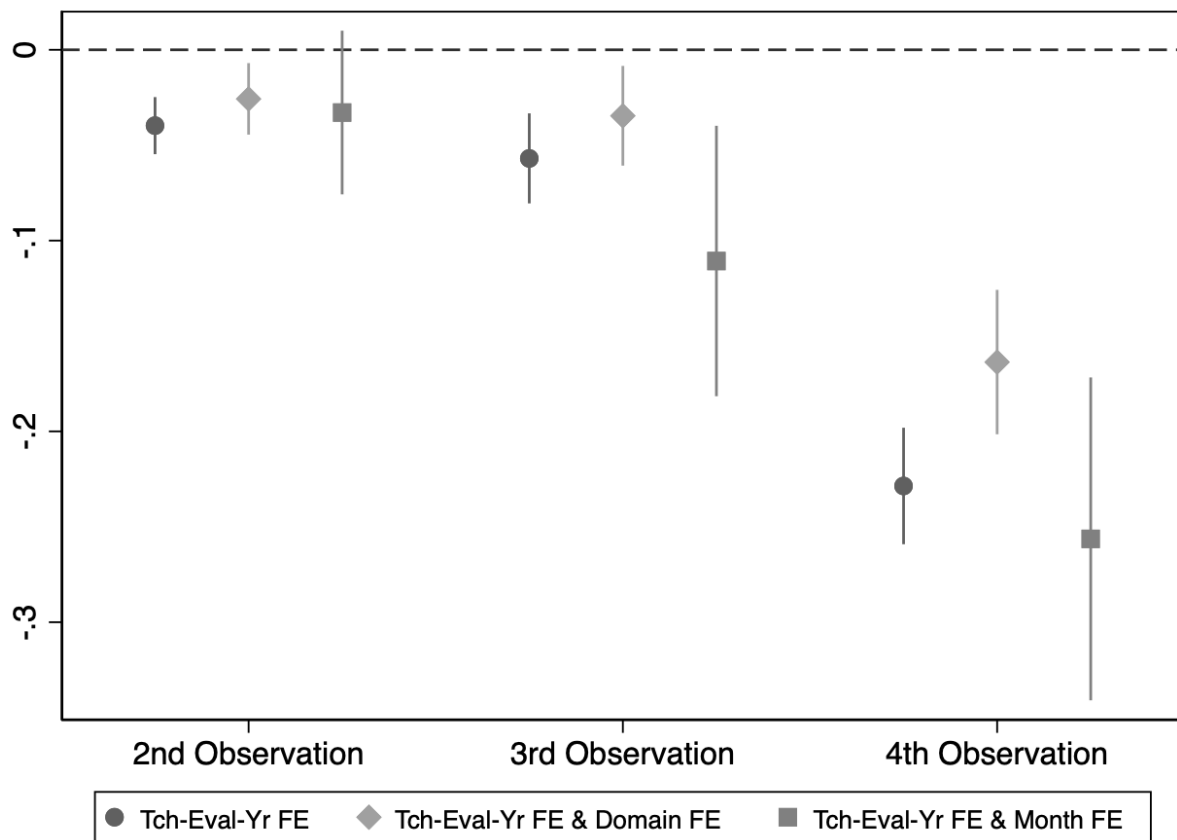
Notes. Observation occurrences are the unit of analysis. Each figure shows the distribution of classroom observation scores. The mean (standard deviation) of teacher self-scores is 3.79 (0.62); the mean (standard deviation) of evaluator scores is 3.73 (0.64); and the mean (standard deviation) of discordance scores is 0.06 (0.55). The count of observations at the teacher-year-observation level is 20,045.

Figure 2. Within-Year Distribution of Observation Scores, by Total Observations Received



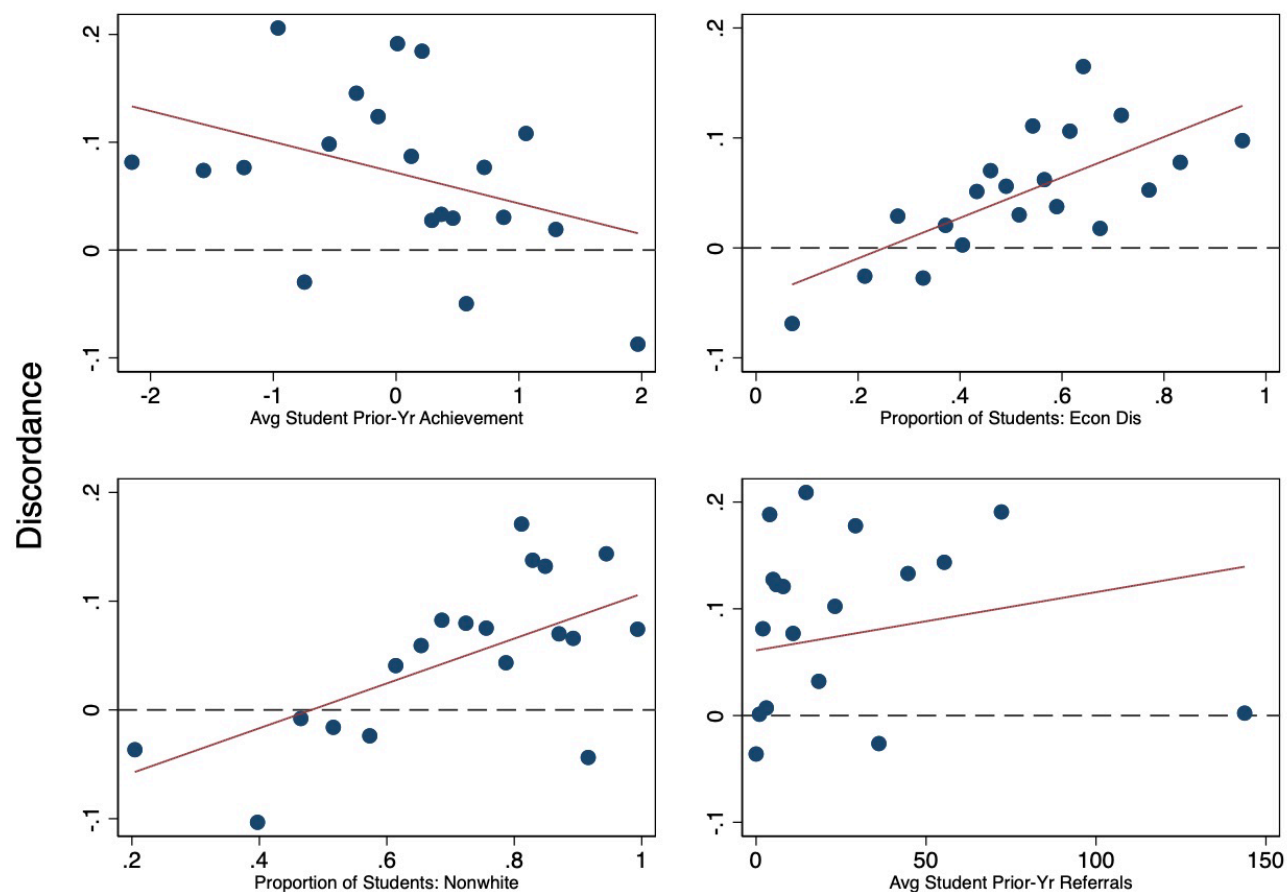
Notes: $N(\text{Teacher-Year-Observation}) = 20,045$ and $N(\text{Teacher-Year}) = 9,070$. The x -axis represents the count (k) of observations received and these are grouped by total annual observations received. Circles represent mean teacher self-assessment scores and diamonds represent mean evaluator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

Figure 3. Nonparametric Estimates of Discordance, by Total Observations Received






Notes: $N(\text{Teacher-Year-Observation}) = 20,045$ and $N(\text{Teacher-Year}) = 9,070$ in Teacher-by-Evaluator-by-Year FE model and Teacher-by-Evaluator-by-Year FE and Domain FE model. $N(\text{Teacher-Year-Observation}) = 9,944$ and $N(\text{Teacher-Year}) = 4,533$; samples differ due to missing month data. Teacher-level clustered standard errors; 95% confidence intervals.



Figure 4. Bivariate Relationship between Discordance and Classroom Characteristics








Notes. Points represent means within bins. $N(\text{Teacher-Year-Observation}) = 6,047$ for Average Student Prior-Year Achievement panel, $N(\text{Teacher-Year-Observation}) = 8,561$ for Proportion of Students: Economically Disadvantaged panel, $N(\text{Teacher-Year-Observation}) = 8,561$ for Proportion of Students: Nonwhite panels, $N(\text{Teacher-Year-Observation}) = 6,537$ for Average Student Prior-Year Referrals. Fit line produced by regressing binned discordance scores on binned baseline characteristic, the sole right-hand side variable in each model. The coefficient (standard error) for prior-year student achievement -0.03 (0.01), student economic disadvantage $= 0.20$ (0.03), nonwhite student $= 0.21$ (0.03), and student prior-year office referrals 0.00 (0.0002).



Appendix A. Classroom Observation Rubric




	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Standards and Objectives 	<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 	<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 	<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 	<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 	<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. Extensive and Intensive Margins of Discordance on Teacher Exit and School Switching

In the main text, Equation (6) estimates extensive margins for teacher 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 school year t (*Exit*) and the second 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.

Table B1 summarizes these results. Evidence in Table 5 suggests that receiving critical feedback, on average, may push teachers out of their school. 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 critical 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 into 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. Results in columns II and IV suggest that teacher exits and school switching is not sensitive to the degree of critical average feedback received.

Table B1. Extensive and Intensive Margins of Discordance on Between-Year Teacher Mobility

	I	II	III	IV
	Exit		Switching	
$\delta I(Discordance_{jt,k-1} > 0)$	0.02 (0.02)	-0.01 (0.02)	0.04 (0.03)	0.01 (0.03)
$\beta_1 Discordance_{jt,k-1} $		-0.07 (0.05)		-0.13 (0.08)
$\beta_2 I(Discordance_{jt,k-1} > 0) \cdot Discordance_{jt,k-1} $		0.15* (0.07)		0.20* (0.10)
$\beta_1 + \beta_2$		0.07 (0.06)		0.07 (0.06)
N(Teacher*Year)	3,980	3,980	3,799	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 discordance measure and teacher-by-evaluator fixed effects and year fixed effects. The sample includes all teachers and estimates are generated by linear probability models. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$