

# Critical Feedback Characteristics, Teacher Human Capital, and Early-Career Teacher Performance: A Mixed-Methods Analysis

Seth B. Hunter 

George Mason University

Matthew G. Springer 

The University of North Carolina at Chapel Hill

*Most education agencies have implemented new teacher evaluation systems that promise to improve teacher performance. Post-observation performance feedback is a theoretically important driver of this promise as it should ultimately develop teacher-specific weaknesses. This is the first large-scale study to use the written feedback provided to early-career teachers during formal post-observation conferences and quantitatively link critical feedback characteristics (CFCs) to measures of teacher human capital. We find that most conferences do not include CFCs, that feedback is typically unidimensional, and that less effective early-career teachers receive higher shares of CFCs. However, goal-setting is the only CFC associated with subsequent teacher performance. Beginning and less-educated teachers, for whom goal-setting may clarify performance expectations, drive this relationship.*

**Keywords:** *evaluation, school/teacher effectiveness, supervision, educational policy, mixed methods, qualitative research, regression analyses*

## Introduction

OVER the last 20 years, policymakers in almost every state have substantially transformed teacher evaluation practices (National Council on Teacher Quality, 2019; Bleiberg & Harbatkin, 2020; Hunter, 2021; Steinberg & Donaldson, 2016).<sup>1</sup> A growing body of work suggests that the success of these reforms rests partly on teacher observation processes and the policies that shape them (Donaldson, 2021). Specifically, teaching improvements seem to depend on the quality of observers' training (Steinberg & Sartain, 2015), observers' discretion in the implementation of teacher observations, structured observation conferences, or post-observation improvement plans (Donaldson & Woulfin, 2018; Hunter & Ege, 2021; Marsh et al., 2017), and teachers' receipt of

observations (Phipps, 2018; Phipps & Wiseman, 2021). Recent quasi-experimental work also indirectly implies that observations improve teachers' performance by increasing their motivation via accountability mechanisms and performance feedback (Phipps & Wiseman, 2021).<sup>2</sup>

Although feedback generally improves teacher performance, theoretical and empirical research suggests that certain feedback characteristics are more potent than others. Specifically, research both within and beyond K–12 settings suggests a positive association between improvements in performance and feedback that (a) is aligned with an improvement area, (b) discusses the feedback's evidential basis, (c) sets specific improvement goals, and (d) includes actionable next steps (Cherasaro et al., 2016; DeNisi &

Murphy, 2017; Hattie & Timperley, 2016; Ilgen et al., 1979).<sup>3</sup> We call these four qualities *critical feedback characteristics* (CFCs). Because students taught by more effective teachers experience better short- and long-term academic and nonacademic outcomes (e.g., Chetty et al., 2014; Doan, 2019; Jackson, 2018), we suspect that policymakers want observation conferences to include performance-enhancing CFCs, especially if their provision is low-cost.<sup>4</sup>

Prior studies explore the prevalence of effective feedback, broadly defined, in current or recently implemented teacher evaluation settings. We extend this work in several ways. Survey and small-scale qualitative research studies report that teachers receive effective feedback (Cherasaro et al., 2016; Donaldson et al., 2014; Finster & Milanowski, 2018; Long, 2019; Sun et al., 2016), but the data collected by these studies are self-reports. Furthermore, survey studies predominantly collect data near the end of a school year while asking participants to recall feedback received over the year, potentially introducing recall bias. Therefore, it is unclear how prior data collection methods have shaped what we know about feedback provision.

Our study is radically different from prior work. We qualitatively code the micro-textual feedback observers provided to approximately 1,200 early-career Tennessee teachers throughout an academic year. We then transform these qualitative codes into quantitative measures of CFCs. As detailed in foundational work by Fesler et al. (2019), access to text data sets, and ours, specifically, now affords researchers the ability to ask questions that were not previously possible. Our data collection methods differ from prior quantitative work in another meaningful manner. Responses are missing from the sampling frames of previous survey studies due to nonresponse, potentially introducing nonresponse bias. As we randomly select teachers from administrative data that includes no missing data from the entire sampling frame, our study avoids nonresponse bias.

We also extend prior work on feedback presenting CFCs. First, we explore whether teachers receiving high concentrations of one feedback characteristic across their post-observation conferences also receive high concentrations of other characteristics. Second, we assess who is receiving

CFCs by drawing on strategic management of human capital theory, which examines how policymakers and school leaders use measures of educator human capital for educator development and other purposes (Odden & Kelly, 2008). Although prior research reports that decision-makers use such measures to strategically manage educator human capital (Cohen-Vogel, 2011; Goldring et al., 2015; Grissom & Bartanen, 2018; Grissom et al., 2017), we are unaware of any studies that examine whether observers issue feedback based on these measures. We hypothesize that teachers with lower human capital (specifically, observation scores, composite effectiveness scores, education level, and experience) receive more CFCs, given that these teachers need more development than teachers with higher levels of human capital.

Finally, we examine the relationships between feedback characteristics and subsequent performance. Prior studies have tried to clarify these connections using participant self-reports. We use a unique data set of externally coded feedback by estimating associations between each of the CFCs and subsequent value-added and observation scores. We also explore whether these associations depend on teacher human capital measures.

The purpose of this study is to extend our understanding of teacher evaluation processes as recently applied in field settings, which is critically important to education policy and practice considering the significant financial investments in teacher evaluation and its use as a focal mechanism to enhance teacher practice and effectiveness. Specifically, we ask the following questions:

- To what extent do teachers' early-career post-observation conferences include the CFCs, that is, feedback which (a) is aligned to improvement targets, (b) refers to evidence, (c) sets goals, and (d) is actionable?
- To what extent is the receipt of one CFC correlated with the receipt of other CFCs?
- To what extent are teacher post-observation conferences exhibiting the examined CFCs associated with teacher baseline observation scores, composite effectiveness scores, education level, and experience?

- To what extent are teacher performance outcomes associated with the proportion of teacher post-observation conferences exhibiting the examined CFCs? Which of the four examined teacher measures of human capital moderate these associations?

The article proceeds as follows. First, we describe literature concerning feedback characteristics and employee improvement. We then discuss strategic human management as a conceptual frame for our work, study context, and data and methods. Finally, the article ends with a presentation and interpretation of our findings, followed by a discussion of implications for future research and policy and practice.

### Feedback Characteristics and Performance Improvement

We consult educational, psychological, and management literature, as the latter fields include decades of established empirical work on the relationships, mediators, and moderators of feedback provided to employees. Research within and beyond the K–12 sector finds positive and negative net associations between changes in employee (teacher) performance and feedback characteristics (e.g., DeNisi & Murphy, 2017; Hattie & Timperley, 2016). Despite this mixed evidence, scholars consistently identify specific feedback characteristics as performance-enhancing. We review prior research to identify feedback characteristics associated with performance improvement, focusing specifically on those characteristics communicated via written feedback. Ultimately, we use these four CFCs to qualitatively code the textual feedback before transforming it into quantitative measures. As displayed and defined in Table 1, the four CFCs are as follows: (a) *Alignment with an area of improvement*, (b) *Evidence*, (c) *Goal-Setting*, and (d) *Action*.

#### *Alignment With an Area of Improvement*

Scholars argue that feedback is more effective when aligned to one job skill because it allows for in-depth analysis and greater focus during post-observation conferences (Cannon & Witherspoon, 2005; Cornelius & Nagro, 2014;

Tuytens & Devos, 2011). For instance, feedback about “questioning” may be too broad; instead, effective feedback might focus on particular questioning techniques such as “the degree to which all students in a class participate in answering questions” or “use of cognitively demanding questions that push student thinking” (Archer et al., 2016). In addition, research has shown that feedback that discusses several areas where improvement is warranted may overwhelm recipients, thus limiting teachers’ responsiveness to observer recommendations (Brinko, 1993; Hattie & Timperley, 2016; Scheeler et al., 2004). We apply these results in two ways. First, we analyze written feedback meant to focus on a single area for teacher improvement (see the Study Context section for details about feedback design). Second, we examine the feedback to determine whether it aligned with the observer-identified area of improvement or not.

#### *Evidence*

Feedback referencing evidence from an observation is one of the most consistently identified characteristics of performance-enhancing feedback (Feeney, 2007; Hattie & Timperley, 2016; Hemmeter et al., 2011; Tuytens & Devos, 2011). Previous research and practitioner resources suggest that feedback effectiveness depends on the specificity and objectivity of the evidence referenced in the feedback (Glickman et al., 2018; Hill & Grossman, 2013; Ilgen et al., 1979). Prior work also suggests that referencing evidence promotes teachers’ perceptions of observer credibility and is positively associated with teachers’ acting on the recommendations from the feedback (Archer et al., 2016; Ilgen et al., 1979; Thurlings et al., 2013; Tuytens & Devos, 2013). Evidence obtained from a teacher observation might be a direct quote of something the teacher said to students, something the teacher wrote on the board, or something students wrote in their papers. We define feedback as including evidence if it describes specific teacher or student behaviors or work products.

#### *Goal-Setting*

Several feedback studies consider goal-setting a characteristic of performance-enhancing feedback

TABLE 1  
*Qualitative Codes: Feedback Characteristics*

Characteristic	Definition	Examples	Interrater percent agreement
Alignment	Feedback references language from the weakest indicator	Oral feedback during the lesson was academically focused. For future lessons, students will engage in effective feedback to each other	89.0
Evidence	Feedback mentions behavior of teacher/ students, although description may be vague	The teacher’s questions were varied and of high quality, mostly relied on volunteers	86.4
Goal-Setting	Feedback directly tells the teacher to change behavior or instruction	By the end of the third 9 weeks, the teacher will include clear measurement criteria for all assessments	84.7
Actionable	Feedback suggests a mechanism by which teacher can change behavior or instruction.	Continue to use strategies to increase student interaction and curiosity like, “What can we add to the answer to help it be even more accurate?” Allow 9 or so seconds of wait time before offering help or hints. This should greatly increase participation and unsolicited responses	91.9

(Cherasaro et al., 2016; DeNisi & Murphy, 2017; Hattie & Timperley, 2016; Hemmeter et al., 2011). In one study, Ivancevich (1982) found that supervisors trained in goal-setting reported that their employees were more likely to take up recommended changes to their work. Goal-setting also signals what supervisors may look for during future observations, clarifying the supervisor’s performance expectations (Ilgen et al., 1979). We characterize feedback as goal-setting if it describes at least one specific behavior for the teacher to change.<sup>5</sup>

*Actionable*

Actionable feedback specifies methods for achieving evaluator-identified goals. Scholars argue that feedback should offer strategies for goal attainment or approaches for closing performance gaps (Carver & Scheier, 1982; Cherasaro et al., 2016; Kimball & Milanowski, 2009; Sun et al., 2016). Observers might recommend strategies based on their knowledge of the individual and the area for improvement or may refer the teacher to other professional learning sources,

such as workshops or peer mentors. We count feedback as actionable if it suggests a mechanism through which the teacher could improve some aspect of her teaching.

**Strategic Management of Human Capital**

The strategic management of human capital in K–12 education aims to improve student outcomes via educator development and the recruitment and retention of talented educators (Hitt & Tucker, 2016; Odden & Kelly, 2008). Research in this area has primarily focused on how school administrators use management information systems to make hiring and classroom assignment decisions (Cannata et al., 2017; Cohen-Vogel et al., 2015; Goldring et al., 2015; Grissom & Bartanen, 2018; Grissom et al., 2017).

The strategic management of human capital may be applicable when providing teachers with critical feedback. Although all employees deserve high-quality feedback, we hypothesize that school administrators may consider teacher human capital when providing critical feedback. Specifically, we hypothesize that teachers in

need of more improvement, that is, teachers with less human capital, will receive more critical feedback. Importantly, we do not assume that school administrators rely exclusively on observable measures to manage teachers. Indeed, prior work indicates that school administrators also consider unmeasured qualities such as commitment and collaborative skills (Grissom & Loeb, 2017; Harris & Sass, 2014; Neumerski et al., 2018).

### Measures of Human Capital

Commonly available measures of educator human capital (e.g., experience, educational attainment, test score value-added, and observation scores) vary in their ability to predict student achievement. Recent research on teacher experience and human capital development has demonstrated that teachers see large gains in effectiveness early in their career (Gershenson, 2016; Harris & Sass, 2011; Ladd & Sorensen, 2017; Papay & Kraft, 2015). Educational attainment is a less reliable measure. Advanced degree attainment on its own is not a reliable predictor of future student achievement (e.g., Harris & Sass, 2011; Wayne & Youngs, 2003). However, teachers with advanced degrees in their assigned subject demonstrate higher effectiveness (e.g., Dee & Cohodes, 2008; Wayne & Youngs, 2003).

Alternative methods for assessing educator human capital include measuring teacher effectiveness via (a) value-added modeling (VAM) or (b) subjective performance ratings, such as scores from classroom observations. Research shows that three consecutive years in a highly effective teacher's classroom (e.g., one with high VAM) can elevate a student's state standardized test-score ranking from the 25th to the 75th percentile (Sanders & Rivers, 1996). More recent work shows that exposure to high value-added teachers increases students' chances of attending college, matriculating at higher ranked colleges, and earning higher salaries (Chetty et al., 2014). We use a teacher's score from the Tennessee Value-Added Assessment System (TVAAS) as a measure of interest, despite some concerns over its operationalization in staffing decisions (Ballou & Springer, 2015; Goldring et al., 2015).

Another widely used approach to measuring teacher quality is classroom observation. Although classroom observations have been used to assess

teachers for far longer than VAM (Brophy & Good, 1986), researchers have only recently begun to evaluate the properties of observation scores (Campbell & Ronfeldt, 2018; Hunter, 2020; Steinberg & Garrett, 2016). The depth of research examining the connections between teacher observation scores and students' outcomes is much more limited. Several recent studies report that classroom observations and student achievement are correlated (Bacher-Hicks et al., 2017; Garrett & Steinberg, 2015; Kane et al., 2011). A recent study from Tennessee found that classroom observation scores not only track teachers' impacts on students' K–12, postsecondary, and labor market outcomes but also that the effects of observation scores are at least comparable to, if not greater than, the effects of teacher value-added scores on various student outcomes including reduced student absences and suspensions (Doan, 2019).

### Study Context: Background on Teacher Evaluation in Tennessee

This study occurs in Tennessee, an ideal setting. Tennessee's education policy requires each teacher to receive at least one observation each school year, with early-career teachers often receiving more. There are clear rules regarding the assignment of teacher observations and structured post-observation feedback (Teacher and Principal Evaluation Policy, 2013). Critically, the Tennessee Department of Education (TNDOE) also expects teacher evaluators to input post-observation feedback into a central data management system that links observer feedback to individual teachers, a key feature we leverage in the current study. We describe the observation system in which our study occurs using the framework developed by Liu et al. (2019).

#### *TEAM Observation System*

Tennessee policymakers adopted the Tennessee Educator Acceleration Model (TEAM) observation and evaluation system in the early 2010s. Teachers are observed using the TEAM rubric (see Supplemental Appendix A, available in the online version of this article) based on Charlotte Danielson's widely used Framework for Teaching. Although the TEAM rubric includes four domains, only three are used for classroom observations: instruction, environment, and planning. These



three domains have 12, four, and three indicators, respectively. The indicators describe specific aspects of standards-based teaching mapped onto three levels of proficiency: *below expectations* (=1), *at expectations* (=3), and *above expectations* (=5).

*Training, Certification, and Accountability.* TNDOE annually provides 2 days of training on using the TEAM rubric, facilitating pre- and post-observation conferences, basic knowledge of Tennessee's evaluation policy, and the characteristics of performance-enhancing feedback (Alexander, 2016). Attendees must pass a certification exam before officially conducting teacher observations (Teacher and Principal Evaluation Policy, 2013). Certified observers need not be school administrators; however, less than 20% of observers are district personnel or peer teachers.

Tennessee policy holds certified observers accountable in three ways. First, observers are expected to generate observation scores that are somewhat aligned with the value-added score of teachers of tested subjects (Teacher and Principal Evaluation Policy, 2013). Observers who consistently generate observation scores that are too far above or too far below a teacher's value-added scores can lose their certification. Second, teachers can file formal grievances if observers do not adhere to policy expectations (Teacher and Principal Evaluation Policy, 2013). For instance, teachers may file a grievance if they do not receive a copy of their observation scores. Third, observers who are school administrators receive their own performance rating (see details below) concerning skills in teacher evaluation and support of teacher professional learning.<sup>6</sup>

*Rating Processes.* Per year, TEAM policy assigns a minimum of four observations to teachers receiving Tennessee's lowest teacher effectiveness score, four or two observations to teachers in the middle categories of effectiveness depending on their certification, and one observation to teachers receiving the highest effectiveness score (Teacher and Principal Evaluation Policy, 2013). Although state policy expects the typical observation to last approximately 15 minutes (Teacher and Principal Evaluation Policy, 2013), teachers report that observations tend to last about 30 minutes (Hunter, 2020). School

administrators decide which observers will conduct observations of which teachers.<sup>7</sup>

Each observation is followed by a timely, structured face-to-face conference, whereas a conference precedes only some observations. Each observation is either announced to the teacher in advance or not. Conferences do not precede unannounced observations but precede announced observations (Teacher and Principal Evaluation Policy, 2013). Tennessee policy states that teachers should receive their post-observation conference within 1 week of each observation (Teacher and Principal Evaluation Policy, 2013). During post-observation conferences, observers discuss an *area of refinement* and an *area of reinforcement*. Each area refers to a single indicator from the TEAM rubric. The area of reinforcement identifies the best aspect of teaching seen during the observation. In contrast, the area of refinement represents the aspect most in need of improvement (Tennessee Department of Education, 2016). Observers are trained to offer suggestions on how teachers might improve their refinement area and to point teachers toward resources that might aid improvement (alignment and actionable feedback). Observers are required to discuss performance ratings across all indicators, not just the areas of refinement and reinforcement. TNDOE expects observers to set improvement timelines with teachers, identifying a timeframe over which the refinement area should improve (goal-setting feedback). Finally, observers discuss scores for each indicator and the basis for each score (evidence-referencing feedback; Alexander, 2016).

### *Level of Effectiveness*

Observation scores and other measures of teacher performance determine each teacher's level of effectiveness (LOE-cont), a continuous composite<sup>8</sup> measure of teacher effectiveness that combines teacher observation scores with "growth" and "achievement" scores. The growth component for teachers of tested subjects is their TVAAS score (for details, see SAS Analytics Software, 2015). Growth scores for teachers of untested subjects are based on school- or district-wide student outcomes (e.g., accountability test scores, school-wide TVAAS scores). Achievement measures are grade-, school-, or

district-wide student achievement outcomes (e.g., ACT scores and high school graduation rates). A teacher and her school administrators receive the teacher's observation, growth, achievement, and LOE-Cont scores in advance of her first observation, as these scores partially determine the number of observations assigned by state policy. Moreover, observers can access a teacher's prior observation scores.

### *Theory of Action*

This study analyzes feedback concerning refinement areas because we are most interested in the characteristics of critical feedback. Reinforcement area feedback emphasizes practices that a teacher should continue; it is not intended to change teaching practices. In contrast, the refinement area, which we classify as critical feedback, is designed to change teachers' practices. Subsequent references to feedback refer exclusively to the refinement area of feedback.

The TEAM theory of action asserts that teacher performance as measured by the TEAM rubric will improve when teachers receive critical feedback that (a) aligns with areas of refinement, (b) references evidence, (c) explicitly sets improvement goals and times, and (d) includes actionable next steps. Written feedback for performance improvement is expected to exhibit "alignment" to a TEAM indicator because refinement feedback aims to improve some aspect of teaching measured by the TEAM rubric. As observers receive annual training about the importance of referencing evidence, identifying next steps for teacher improvement, and goal-setting, the analyzed feedback is expected to include characteristics (b) to (d) mentioned earlier. The TEAM theory of action assumes that feedback provided at any point of the year, even late in the school year, can improve teaching as measured by observation scores and may be able to improve value-added scores (i.e., TVAAS scores; Alexander, 2016). Emerging evidence corroborates this assumption (Phipps, 2018).

If feedback improves teacher performance as measured by the TEAM rubric, it is expected to improve TVAAS scores. Prior work finds that higher TEAM scores are associated with higher student achievement on standardized tests (Daley & Kim, 2010). Teachers whose teaching

improves more may raise their students' achievement scores by more, translating into even higher TVAAS scores. Similar logic undergirds recently reformed teacher evaluation systems (Steinberg & Donaldson, 2016).

## **Data and Method**

### *Data*

Our data come from administrative records obtained from TNDOE via the Tennessee Education Research Alliance. Table 2 presents descriptive statistics for variables from the 2013–2014 school year, the baseline year. Teachers' data include years of experience, age, gender, highest degree held, and race/ethnicity. We also obtained teachers' LOE scores and scores from LOE determinants.

TNDOE collects a written version of post-observation feedback regarding a teacher's area of refinement after each observation. Observers record this written feedback into a TNDOE information management system, and unique identifiers link feedback data to individual teachers. Some teachers received a single observation and post-observation feedback session, whereas others received several. In our sample, the average teacher was observed 3 times and received three post-observation conferences, although some received as many as 10.

We examine written critical feedback entries for 1,219 randomly selected early-career teachers (i.e., teachers with less than 5 years of experience) across the state. We focus on early-career teachers because Tennessee policy requires these teachers to be observed at least 4 times per year unless they receive a prior-year effectiveness score of 5, which would place them in the highest category of teacher effectiveness. In addition, prior research suggests that early-career employees are more likely to benefit from on-the-job feedback (Kimball, 2003), implying that if there are associations between feedback quality and changes in performance among any group of teachers, it would be early-career teachers, *ceteris paribus*. Our randomly chosen sample closely resembles the population of all early-career Tennessee teachers (see Table 2). The only statistically significant difference is the proportion of non-White teachers in our sample (0.05) compared with those in the state (0.13).

TABLE 2  
*Standardized Differences Between Sample and Population*

Measures	Sample		Population		SD
	<i>M</i>	<i>SD</i> or %	<i>M</i>	<i>SD</i> or %	
Prior-Yr LOE-cont	354.66	82.45	360.65	80.84	−0.07
Prior-Yr Obs Score	3.71	0.53	3.73	0.58	−0.04
Prior-Yr TVAAS	0.18	3.61	0.58	3.76	−0.11
Female	0.8	0.4	0.82	0.39	−0.03
Non-White	0.05	0.21	0.13	0.34	−0.32*
MA or Higher	0.03	0.17	0.04	0.19	−0.05

*Note.* LOE = level of effectiveness; TVAAS = Tennessee Value-Added Assessment System.  
 \*0.40 ≥ | *standardized difference* | ≥ 0.20, a “small” difference (Cohen, 1988).

### *Qualitatively Coded Written Feedback*

Two analysts independently coded approximately 5,000 individual critical feedback episodes for the characteristics identified in Table 1. If a post-observation conference included critical feedback exhibiting a specific feedback characteristic, analysts flagged the conference accordingly. For example, if one conference included critical feedback making a single reference to evidence, a second conference included critical feedback making several references to evidence, and a third conference no mention of evidence, the first two conferences are coded as one and the last zero. We repeated this process for each of the four feedback characteristics identified for our study.

We adopted two processes to ensure that analysts coded the feedback similarly. First, about 500 conferences were randomly selected and independently coded by each analyst. The mean percentage of agreement in the coding of this subsample was 88% and, as seen in the right-most column of Table 1, it ranged from a high of 91.9% agreement in the “actionable” characteristic to a low of 84.7% in the “goal-setting” attribute. All the percentages of agreement exceed the 80% minimum agreement threshold established by Marques and McCall (2005) and Belur et al. (2021). Where disagreements existed, analysts rereviewed written post-observation feedback entries, discussed differences, and adjusted subsequent coding. Second, to mitigate “coder drift” (Bartholomew et al., 2000), the two analysts conducted check-ins every 2 weeks to

maintain their shared understanding of the qualitative codes while sharing examples of the types of text representative of each characteristic (Carey & Gelaude, 2008).

### *Quantitative Analytic Strategies*

We convert qualitatively flagged conferences to teacher-level quantitative measures by calculating the share of each teacher’s conferences that included critical feedback exhibiting each of our CFCs. This resulted in four proportions per early-career teacher. Unconditional means and standard deviations (*SDs*) describe the feedback characteristics of an early-career teacher’s typical (i.e., mean) conference and variation in these characteristics.

*Associations With Baseline Differences.* We examine the relationships between each feedback characteristic and the baseline differences in early-career teacher human capital measures using Equation 1:

$$c_{ij}^k = X_{ij}A + e_{ij}, \tag{1}$$

where  $c_{ij}^k$  is the share of the  $i$ th teacher’s conferences in school  $j$  that include feedback exhibiting characteristic  $k$ . Vector  $X_{ij}$  includes teacher prior-year effectiveness score, prior-year observation score, whether the teacher holds higher than a Master’s degree, and years of teaching experience.

To facilitate interpretation, we standardize continuous measures in  $X_{ij}$ . By including all the



baseline measures of early-career teacher human capital as right-hand side variables, we can identify which measures have the strongest associations with  $c_{ij}^k$  independent of the other measures. The error term is  $e_{ij}$ . As  $c_{ij}^k$  is a proportion ranging from 0 to 1, we interpret coefficients in  $A$  in terms of probability points. We hypothesize that these coefficients will be negative, that is, higher shares of CFCs are provided to teachers with less human capital.

*Associations With Early-Career Teacher Outcomes.* To investigate the relationships between  $c_{ij}^k$  and early-career teachers' performance, we estimate Equation 2:

$$y_{ij} = \delta c_{ij}^k + X_{ij}A + e_{ij} \quad (2)$$

The variable  $y_{ij}$  represents the observation scores or TVAAS scores of teacher  $i$  in school  $j$  and  $X_{ij}$  refers to the same measure as in Equation 1. In Equation 2, we rescale  $c_{ij}^k$  so that  $\delta$  represents the associated change in  $y_{ij}$  for a 10 percentage point increase in the share of a teacher's conferences with critical feedback exhibiting characteristic  $k$ . The vector  $X_{ij}$  is included to increase the precision of  $\delta$  and because we use these variables for moderation analyses. As effectiveness scores are a function of observation scores, ordinary least square estimates the effectiveness score coefficient using variation in effectiveness scores that is not due to observation score variation. Prior work suggests that  $\delta$  will be positive.

*Moderation by Measures of Early-Career Teacher Human Capital.* We investigate whether measures of early-career teacher human capital moderate the relationships between  $c_{ij}^k$  and  $y_{ij}$  using Equation 3:

$$y_{ij} = \delta_1 c_{ij}^k + \delta_2 c_{ij}^k m_{ij} + X_{ij}A + e_{ij}, \quad (3)$$

where  $m_{ij}$  is one of the measures of teacher human capital from  $X_{ij}$ . We transform  $m_{ij}$  and its corresponding measure in  $X_{ij}$  to facilitate interpretability of the moderators, "centering" all moderators on the lowest value for each score. Prior-year effectiveness is centered at 100 and prior-year observation scores are centered at 1. Both are rescaled according to their  $SD$ s. Teacher

experience is centered at 1 year of experience but not rescaled. Standard errors in all equations are clustered at the school level.<sup>9</sup>

## Findings

### *Distributions and Correlations of CFCs*

Although there is substantial variation in the concentration of CFCs early-career teachers receive, the modal teacher tends not to receive any of the feedback examined and the mean teacher receives low concentrations of critical feedback (Figure 1). Each panel of Figure 1 shows that some teachers receive no conferences with the CFC while all other teachers' conferences include it (the range of each characteristic is zero to one).  $SD$ s also suggest ample within-characteristic variation as each  $SD$  is approximately 0.24 units. However, there are some important between-characteristic differences. Approximately 50% to 80% of conferences included aligned feedback for a relatively large number of early-career teachers, which is why 66% of the mean teachers' conferences are aligned (dashed line). Moreover, 80% of the modal teacher's conferences contained aligned feedback. However, distributions of the remaining CFCs are somewhat disappointing. None of the conferences received by the modal teacher include evidence, goal-setting, or actionable feedback, and no more than one-third of the conferences received by the mean teacher exhibited these CFCs (dashed lines). Goal-setting and actionable feedback are particularly rare, as approximately half of the teachers in our sample received no conferences with these CFCs.

Teachers whose conferences document high shares of one CFC tend not to show high shares of any other CFC (Table 3). That is, the data suggest that few teachers receive conferences including more than one CFC over a school year. The largest correlation between the shares of conferences that include any two CFCs is 0.33 (evidence and actionable characteristics), a modest relationship at best. Correlations with the alignment CFC range from 0.18 to 0.28, all of which are low enough to suggest that its provision is unrelated to the provision of any other CFC. The weakest correlations belong to the goal-setting CFC, ranging from 0.08 with the actionable CFC and a near-zero of  $-0.03$  with evidence.

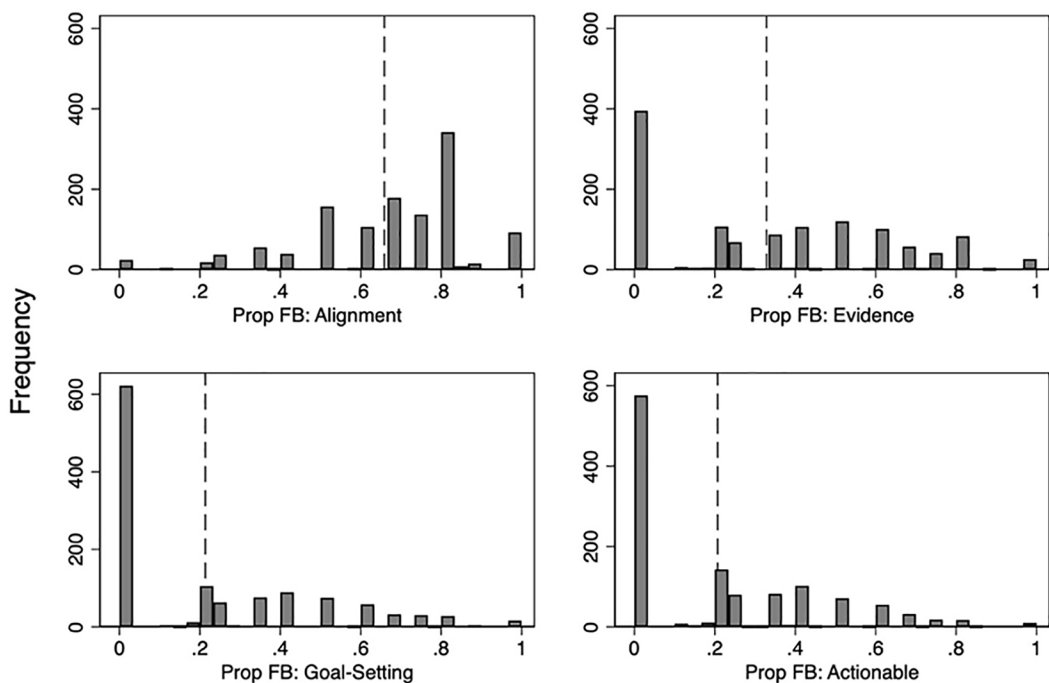


FIGURE 1. Histograms: Share of early-career teachers' feedback conferences including feedback characteristics.

Note.  $N = 1,219$ . Teachers are unit of analysis; includes first-year teachers. Dashed lines represent means. Alignment mean (SD) = 0.66 (0.21), evidence = 0.33 (0.29), goal-setting = 0.21 (0.26), and actionable = 0.21 (0.24), FB = feedback.

TABLE 3

*Correlations Between Critical Feedback Characteristics*

Characteristics	Actionable	Alignment	Evidence	Goal-setting
Actionable	1			
Alignment	.18	1		
Evidence	.33	.22	1	
Goal-Setting	.08	.28	-.03	1

Note.  $N = 1,219$  Teachers are unit of analysis; includes first-year teachers. Pearson correlations.

Taken together, findings in Figure 1 and Table 3 suggest that most teachers receive post-observation conferences that include few, if any, of the examined feedback characteristics. In addition, if a teacher receives a high share of one CFC, they are unlikely to receive a high share of another. Although most early-career teachers' conferences may not exhibit many characteristics of critical feedback, each of the four CFCs has considerable variation to exploit for the current study. To do so, we first describe who receives higher concentrations of conferences with each

of the four CFCs by examining teacher baseline human capital measures.

#### *CFCs and Measures of Early-Career Teacher Human Capital*

Table 4 reports associations between CFCs and measures of early-career teacher human capital. The only consistent relationship detected is that teachers with lower prior-year effectiveness scores tend to receive conferences with higher concentrations of each CFC (Table 4). An early-career

TABLE 4  
*Associations With Teacher Baseline Differences*

	I	II	III	IV
Baseline Characteristics	Actionable	Alignment	Evidence	Goal-setting
Level of effectiveness-Cont	−0.02* (0.01)	−0.07*** (0.01)	−0.03** (0.01)	−0.03* (0.01)
Obs Score	−0.02 (0.01)	0.01 (0.01)	0.01 (0.02)	−0.02 (0.02)
MA or Higher	−0.06 (0.04)	−0.05 (0.04)	0.00 (0.05)	−0.03 (0.05)
Yrs Exp	−0.01 (0.07)	0.03 (0.07)	−0.27** (0.09)	0.01 (0.08)
Adj- <i>R</i> <sup>2</sup>	.03	.10	.03	.02
<i>N</i>	1,175	1,175	1,175	1,175

*Note.* Standard errors clustered at school. All predictors are standardized.  
 \**p* < .05. \*\**p* < .01. \*\*\**p* < .001.

teacher with a 1*SD* lower prior-year effectiveness score is 2 probability points more likely to receive actionable feedback in all their post-observation conferences, conditional on observation scores (Column I). The relationships between prior-year effectiveness and the evidence (Column III) and goal-setting (Column IV) CFCs are similar in magnitude to the relationship with actionable feedback. Prior-year effectiveness relates most strongly with the share of conferences presenting the aligned CFC. An *SD* decrease in prior-year effectiveness increases the chance that all feedback conferences exhibit aligned feedback by 7 probability points (Column II).

Although the relationship with LOE-Cont is statistically significant across models, the largest association is between years of experience and the evidence CFC (column III). Early-career teachers with 1*SD* fewer years of experience (i.e., 1 year less experience) are 27 probability points more likely to receive a feedback conference including the evidence-referencing CFC. However, the provision of no other CFC depends on years of experience. Similarly, none of the CFCs depend on prior-year observation scores or education level.

Ultimately, less-experienced early-career teachers are more likely to receive feedback with the evidence CFC, while less effective teachers are slightly more likely to receive conferences

with all CFCs. Critically, as observation scores partially determine effectiveness scores and all models control for effectiveness and observation scores, relationships with prior-year effectiveness are effectively based on variation in nonobservation score components (i.e., student outcomes).

#### *CFCs and Early-Career Teacher Outcomes*

Although prior research suggests that receiving feedback with the CFCs improves teacher performance, little evidence substantiates this claim (Table 5). A 10 percentage point rise in the share of teachers' conferences presenting actionable feedback is associated with a near-zero and null decline in observation scores of 0.01 *SD* (0.6 units; Panel A). As the magnitude of this association is insignificant, and its standard error is precise (0.01), measurement error is unlikely to drive this null finding; we validate this below. Similar patterns exist between teacher observation scores and the alignment and evidence CFCs in Panel A. Only the goal-setting relationship is statistically significant, but it is negative (−0.02 *SD*). The bottom panel of Table 5 also tends to show near-zero and null relationships between TVAAS scores and CFCs. Most coefficients are precisely estimated but not statistically significant. Again, only the goal-setting CFC is statistically significant, but this time the relationship is positive, whereby

TABLE 5

Associations With Teacher Outcomes

CFCs				
Panel A. Observation Scores				
Actionable	−0.01 (0.01)			
Alignment		−0.02 (0.01)		
Evidence			−0.01 (0.01)	
Goal-setting				−0.02* (0.01)
<i>N</i>	1,054	1,054	1,054	1,054
Panel B. TVAAS Scores				
Actionable	0.01 (0.01)			
Alignment		0.02 (0.02)		
Evidence			−0.02 (0.01)	
Goal-setting				0.03* (0.02)
<i>N</i>	838	838	838	838

Note. Outcomes are standardized. Standard errors clustered at school level. CFCs = critical feedback characteristics; TVAAS = Tennessee Value-Added Assessment System.

\**p* < .05.

a 10 percentage point increase in the share of conferences with goal-setting feedback is associated with a rise of 0.03 *SD* in TVAAS.

Measurement Error

Measurement error is one of the most fundamental problems in education policy analysis and evaluation, particularly when relying on human coding of written feedback to operationalize complex constructs. Although the percentages of interrater agreement in Table 1 exceed the 80% minimum agreement threshold proposed by some researchers (Belur et al., 2021; Marques & McCall, 2005), error in the coding of feedback characteristics may explain the null associations with actionable, aligned, and evidence-based CFCs. We explore this issue using two conceptually different sensitivity tests, both of which refute this hypothesis.

We first examine the sensitivity of our estimates by applying *errors-in-variables* (EiV) regression,

which adjusts potentially error-prone estimated coefficients and standard errors for additive measurement error. If substantial measurement error exists, we presume that it is “additive” (for a discussion of additive measurement error, see Hardin & Carroll, 2003). Qualitative analysts coded randomly selected feedback episodes and knew nothing about the feedback providers, recipients, their schools, or their districts. As recommended by Lockwood and McCaffrey (2020), we apply bootstrapped standard errors. To specify the reliability ratio of the potentially error-prone variable—a critical component in the EiV approach—we use calculated reliabilities from the interrater agreements reported in Table 1.<sup>10</sup>

The second sensitivity test uses the *simulation extrapolation* method (SIMEX) developed by Cook and Stefanski (1994). SIMEX simulates what happens to estimated coefficients and standard errors if the suspected error-prone variable suffered from more additive measurement error. By adding additional measurement error via a

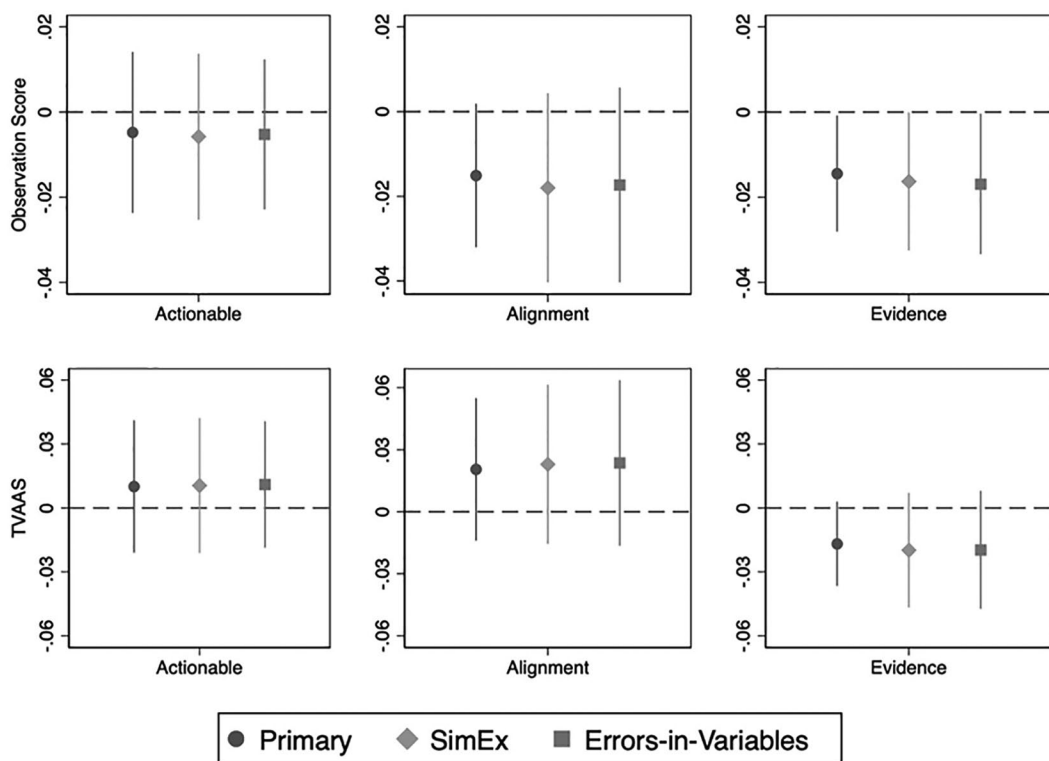


FIGURE 2. *Measurement error sensitivity tests results.*

*Note.* Point estimates and 95% confidence intervals based on bootstrapped standard errors. Each point estimate and corresponding confidence interval is generated by a different model. Associations with observation scores in top row, Tennessee Value-Added Assessment System associations in bottom row.

resampling approach, SIMEX generates a measurement error trend and then uses that trend to extrapolate back to an error-corrected estimate (i.e., an estimate with no measurement error). In our case, a meaningful difference between the noncorrected regression estimates (and standard errors) and the SIMEX estimates (and standard errors) would suggest that measurement error biased the Table 5 results.

The point estimates and confidence intervals displayed in Figure 2 indicate that the Table 5 near-zero null results are unaffected by biasing measurement error. Each cell in Figure 2 displays the EiV and SIMEX coefficients and 95% confidence intervals for an association between one of the CFCs (actionable, alignment, or evidence) and one of the outcomes (observation or TVAAS score). A separate equation generates each point estimate and confidence interval. Given that the sensitivity tests use bootstrapped standard errors, and our primary analysis used clustered standard

errors, we re-estimate Equation 2 with bootstrapped standard errors; these estimates are labeled “primary.” In Figure 2, the point estimates and confidence intervals are remarkably similar within each cell, supporting the conclusion that Table 5 results are unaffected by biasing measurement error.<sup>11</sup>

### *Moderated Associations*

*Observation Scores and Measures of Teacher Human Capital.* There is little evidence to suggest that teacher human capital measures moderate the relationship between observation scores and shares of CFCs (Table 6). None of the main or moderated associations with the actionable or aligned CFCs are statistically significant (Columns I and II). Associations between conferences including evidence-referencing CFCs and observation scores depend on prior-year effectiveness scores, independent of information



TABLE 6  
*Associations With Observation Scores Moderated by Teacher Human Capital*

	Actionable	Alignment	Evidence	Goal-setting
Moderator	I	II	III	IV
Panel A. LOE-Cont				
Main	0.00 (0.03)	−0.04 (0.04)	−0.06* (0.02)	−0.02 (0.03)
Interaction	−0.00 (0.01)	0.01 (0.01)	0.01* (0.005)	0.00 (0.01)
Panel B. Obs Score				
Main	0.03 (0.03)	−0.01 (0.01)	−0.06 (0.04)	−0.03 (0.04)
Interaction	−0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.00 (0.01)
Panel C. MA or Higher				
Main	−0.01 (0.01)	−0.01 (0.01)	−0.01 (0.01)	−0.02* (0.01)
Interaction	0.05 (0.05)	−0.01 (0.05)	−0.04 (0.04)	−0.00 (0.03)
Panel D. Experience				
Main	0.00 (0.01)	−0.01 (0.02)	−0.00 (0.01)	−0.01 (0.01)
Interaction	−0.01 (0.01)	−0.00 (0.01)	−0.01 (0.01)	−0.00 (0.01)
N	1,054	1,054	1,054	1,054

*Note.* The outcome is standardized. Standard errors, in parentheses, clustered at school level. LOE = level of effectiveness.  
 \**p* < .05.

contained in prior-year observation scores. However, these are the only moderated associations detected in Column III. The least effective early-career teachers receiving a 10 percentage point higher share of conferences with evidence-referencing CFCs have lower observation scores (−0.06 *SD*; Panel A, Column III), and the interaction increases observation scores by 0.01 with each *SD* increase in prior-year effectiveness score (~ 80 LOE-Cont units). Yet even the total association for early-career teachers with the highest prior-year effectiveness score is 0.01 *SD* and nonsignificant at the 5% level. We also find that less-educated early-career teachers who receive a 10 percentage point higher share of conferences with the goal-setting CFC have lower observation scores (−0.02 *SD*), while there is effectively no relationship among more educated early-career teachers (Panel C, Column IV).

*TVAAS Scores and Measures of Teacher Human Capital.* Thus far, the only association with teacher outcomes corroborating our hypotheses is between the goal-setting CFC and TVAAS scores. Results of Table 7 suggest that less-educated and less-experienced teachers drive this association, as expected (Column IV). Less-educated early-career teachers receiving a 10 percentage point higher share of conferences including the goal-setting CFC have 0.03 *SD* higher TVAAS score, whereas the least experienced early-career teachers receiving a higher dosage of goal-setting feedback have 0.05 *SD* higher TVAAS score. No other interactions with education and experience are statistically significant. Furthermore, we conclude that there are no other moderated relationships between TVAAS scores and actionable, aligned, or evidence-referencing CFCs. Although the interaction between experience and actionable feedback is statistically significant, the main association is not,

TABLE 7

*Associations With TVAAS Scores Moderated by Teacher Human Capital*

	Actionable	Alignment	Evidence	Goal-setting
Moderator	I	II	III	IV
Panel A. LOE-Cont				
Main	−0.01 (0.06)	0.08 (0.07)	0.04 (0.04)	0.02 (0.05)
Interaction	0.01 (0.02)	−0.02 (0.02)	−0.02 (0.01)	0.00 (0.01)
Panel B. Obs Score				
Main	0.01 (0.09)	0.09 (0.09)	0.04 (0.06)	0.01 (0.05)
Interaction	0.00 (0.02)	−0.01 (0.02)	−0.01 (0.01)	0.00 (0.01)
Panel C. MA or Higher				
Main	0.01 (0.02)	0.02 (0.02)	−0.02 (0.01)	0.03* (0.01)
Interaction	−0.10 (0.12)	0.00 (0.06)	0.05 (0.06)	−0.08 (0.10)
Panel D. Experience				
Main	0.04 (0.02)	0.04 (0.03)	0.00 (0.02)	0.05* (0.03)
Interaction	−0.02* (0.01)	−0.01 (0.02)	−0.02 (0.01)	−0.01 (0.01)
N	838	838	838	838

*Note.* The outcome is standardized. Standard errors, in parentheses, clustered at school level. TVAAS = Tennessee Value-Added Assessment System; LOE = level of effectiveness.

\**p* < .05.

which is why we infer that there is no evidence of moderation in Column I.

Discussion

Prior work suggests that observation processes influence teacher development, specifically post-observation feedback (Donaldson, 2021). Theoretically, observation processes in recently reformed teacher evaluation systems provide teachers with critical feedback that will improve their performance, directly or indirectly, by pointing them to appropriate professional learning opportunities (Donaldson, 2021). This study examined the prevalence of four CFCs (specifically, evidence-referencing, goal-setting, aligned to improvement area, actionable) in post-observation conferences within Tennessee’s reformed evaluation system. Unlike prior studies, which

predominantly rely on self-reports and self-selection for participation in studies, we qualitatively coded nearly 5,000 instances of written feedback provided to a random sample of early-career teachers. We then converted codes for quantitative analysis. Ultimately, we created a data set described by others as presenting previously unseen analytic affordances and contributions to teacher evaluation research (Fesler et al., 2019). As such, several of our findings offer new insights on teacher observation processes and challenge the findings of previous feedback research.

Our data suggest that few early-career Tennessee teachers’ conferences include CFCs. Over an academic year, nearly half of the teachers in our sample did not receive any actionable or goal-setting feedback, and nearly one third did not receive feedback referencing evidence. In most teachers’ conferences, the only CFC

present was feedback aligned to an identified improvement area, a relatively easy characteristic to manifest in the Tennessee context. Tennessee's information management system forces observers to identify an area of strength and improvement after every observation, prompting them to provide feedback aligned with the improvement area. Although a crucial policy and practice takeaway might be to increase the provision of CFCs, which we agree with in concept, we are cautious about providing a blanket declarative given the limited associations between our CFC measures and important teacher outcomes, a finding we address in greater detail later in the discussion.

Despite the low concentration of CFCs received by the modal teacher, we found substantial variation in teachers' shares of conferences presenting each CFC. Using this variation, we conclude that teachers whose conferences exhibit high shares of one CFC tend not to exhibit high shares of other CFCs. We do not know whether the correlational evidence implies that observers are unskilled at providing several characteristics of critical feedback simultaneously or if their choice to provide specific characteristics is strategic; untangling these interpretations warrants additional research.

Although the visual and correlational evidence implies opportunities to improve observation conference implementation, select analyses find that observers issue critical feedback to the early-career teachers who need it most. We examined whether teacher experience, education level, prior-year effectiveness, and observation scores played a role in determining the recipients of high shares of conferences with CFCs. As most of the relationships between CFCs and human capital measures were negative, the evidence partially supports our hypotheses that teachers with less human capital would receive higher shares of critical feedback. Simultaneously, most of these negative relationships were statistically nonsignificant. Tennessee's *de facto* teacher composite measure of effectiveness—LOE—was the only human capital measure examined that consistently affected the provision of CFCs. As hypothesized, less effective early-career teachers were more likely to receive higher concentrations of CFCs, extending strategic management of human capital theory by suggesting

that teacher effectiveness informs feedback provision.

That teachers with less human capital receive conferences with theoretically performance-enhancing feedback is promising. However, we suspect that policymakers are more interested in subsequent improvements to teaching performance. The evidence is disappointing in this regard, as the receipt of higher shares of critical feedback across conferences is not associated with subsequent teacher performance, on average. Goal-setting is the only CFC associated with subsequent observation or TVAAS scores, and in the former, the association is negative while the latter's association is positive. Other research concludes that students taught by teachers with 1.0 *SD* higher value-added are 0.82 percentage points more likely to attend college, will enjoy an adult income hike of 1.3%, and will score 0.13 *SD* higher on end-of-year tests (Chetty et al., 2014). Back-of-the-envelope calculations suggest that increasing the share of teachers' conferences presenting the goal-setting CFC by 1 *SD* (21 percentage points) may increase the likelihood of college attendance by 0.17 percentage points, raise adult income by 0.27%, and increase test scores by 0.03 *SD*. These are small but meaningful changes.

Why might CFCs associate with observation scores negatively and TVAAS scores positively? The answer may rest in the specific goals set during feedback conferences. An established body of work suggests that observation and value-added scores capture different aspects of teaching (Bacher-Hicks et al., 2017; Hill & Grossman, 2013; Kraft et al., 2020). Goal-setting CFCs may focus on teaching aspects that are more related to gains in student achievement than teaching practice as evaluated by a classroom observation rubric. Given that TVAAS scores represent a sizable component of a teacher's effectiveness score, it is plausible that Tennessee observers set goals in written feedback focused on improving student test scores. Furthermore, the negative main and moderated associations with observation scores may reflect bias. Research suggests that an observer's prior knowledge about a teacher's performance can influence observation scores (Hunter, 2020). Observers might issue critical feedback to teachers they expect will have difficulty improving; these expectations may downwardly bias subsequently assigned observation

scores independently of teachers' performance. Although such bias affects observation scores, it cannot affect TVAAS scores, which may account for the positive TVAAS associations.

In addition, moderation analyses find that goal-setting feedback provided to less-educated and less-experienced early-career teachers positively correlates with TVAAS scores. These two moderated associations are consistent with the idea that the purpose of goal-setting is to help early-career teachers understand performance expectations. Teachers may develop their understanding of performance expectations through formal education or on-the-job experience. Goal-setting may compensate for the absence of these experiences among the less-educated and less-experienced.

We recognize that our findings are at odds with some prior research. For example, teacher survey results from two states suggest that teachers receive CFCs (e.g., Cherasaro et al., 2016), while Chicago teachers and administrators say that evaluations are helpful, implying effective feedback (Sartain et al., 2020). The uniqueness of our data may largely explain these differences, further underscoring the need for further research using similar data.

Our study is not without limitations. First, we only examine four CFCs, but research identifies many feedback characteristics for improvement, including specificity and timeliness (Jawahar, 2010; Kimball, 2003; Kinicki et al., 2004). Future research might directly examine additional characteristics of feedback provided to teachers and relate those characteristics to teacher performance. Prior survey research also finds that employee reactions to feedback mediate associations between feedback characteristics and employee performance (Jawahar, 2010). During the post-observation conference, the words, tone, and body language of an observer may also affect how a teacher responds to their oral and written feedback. It would be helpful to know whether these same mediators and moderators apply to information based on the feedback directly provided to teachers during formal conferences.

Second, the associations between the proportion of teacher conferences exhibiting specific CFCs and subsequent teacher performance and effectiveness may not be causal. That we find associations between teacher baseline human

capital measures underscores this limitation. Indeed, the lack of positive associations with teacher performance and effectiveness may be explained by the fact that early-career teachers with less human capital receive conferences with higher shares of the feedback examined, which in turn may introduce negative bias.

Finally, the generalizability of our results is limited. We purposefully selected teachers in their first 5 years on the job; our findings may not generalize to mid- or late-career teachers. In addition, as findings from feedback research may depend on the observation system in which the feedback is provided, our findings may not generalize beyond the TEAM evaluation system. Machine learning techniques may be well-suited for generalizability. Computers can learn how we qualitatively coded written feedback and then use our coding procedures to other written feedback in new samples, substantially expanding the analytical data set.

Notwithstanding these potential limitations, the evidence suggests that feedback processes within Tennessee's teacher evaluation system, one of the most mature "next-generation" systems in the United States (Koedel et al., 2019; Steinberg & Garrett, 2016), are not working as intended. We end by discussing how policy and research might improve feedback processes, but we urge caution befitting a first-of-its-kind study like ours. Indeed, such prudence underscores the need for more large-scale research using externally coded feedback episodes from field settings.

Based on prior work, the lack of CFCs and unidimensionality of feedback suggests that policymakers should promote higher levels of CFCs and multidimensional feedback, which theoretically improves teacher performance. However, evidence in the current study suggests that higher concentrations of the examined CFCs do not affect teacher performance, implying that policies pushing for higher concentrations of critical feedback, *ceteris paribus*, may not be effective. Similarly, other research implies that increasing the number of observations per teacher to increase the total levels of critical feedback received, *ceteris paribus*, may be unwise without first addressing the quality of the observations themselves. For example, emerging quasi-experimental research finds that more observations do not improve student

achievement (de Barros, 2019; Hunter, 2019). Furthermore, while there is mixed evidence concerning the burdens of higher observational loads on evaluators, it seems that evaluators may cope with higher loads by reducing time spent in pre- and post-conferences, potentially undermining the developmental goals of increasing observations (Hunter & Rodriguez, 2021; Kraft & Gilmour, 2016; Rigby, 2015). Ultimately, we believe there is not yet enough evidence to inform policy regarding the provision of CFCs, highlighting the need for more research. Specifically, we urge researchers to expand on work like ours by using externally coded feedback episodes linked to subsequent teacher performance measures.

Our study implies that policymakers might take steps to ensure that certain teachers receive conferences with high concentrations of goal-setting feedback. Although the main associations between goal-setting and TVAAS are significant, they are driven by feedback received by beginning and less-educated teachers, implying that goal-setting benefits these teacher groups. Our findings suggest that teachers' prior-year effectiveness drives the provision of goal-setting feedback, not teacher experience or education. In light of this, policy-driven supports are likely needed to change how evaluators issue goal-setting feedback. For example, policies might encourage principal preparation programs to train leaders in performance management principles and the role of goal-setting in employee improvement (Gallo, 2011). Another avenue is for state or district policies to support executive coaching to transform school administrator (i.e., evaluator) practices (Hagen, 2012; Huff et al., 2013). At the same time, recent research shows workshop-style in-service is typically ineffective, whether for teachers (Garet et al., 2001; Penuel et al., 2007) or administrators (Kraft & Christian, 2021).

Our study also raises questions about other opportunities for improving feedback and observation processes. The findings suggest that the feedback provided to teachers across their observations is predominantly unidimensional, exhibiting just one of the examined CFCs. Critical feedback may be more effective when it is multidimensional, consistently exhibiting multiple CFCs. Future research should examine the effect of multidimensional critical feedback on teacher performance. As we do not find many substantive

associations, researchers might also explore the policy-manipulable factors that might suppress feedback's effects (e.g., lack of targeted teacher professional learning opportunities supporting formal goal attainment). Although psychological research identifies several employee (teacher) cognitive factors suppressing feedback's effects (Jawahar, 2010; Kinicki et al., 2004; London & Smither, 2002), we suspect that policy is ill-positioned to change cognitively based suppressors.<sup>12</sup> However, some have intimated that loose or non-existent connections between evaluation processes and professional development systems suppress teacher performance improvement (Donaldson, 2021; Papay, 2012; Weisberg et al., 2009). Large-scale research exploring the importance of these connections might eventually show that policymakers should purposefully link evaluation and professional development systems, if feedback and other observation processes are to improve teacher performance and, ultimately, the learning opportunity provided to students.

### Acknowledgments

This article is much improved from its original versions, thanks to helpful CFCs from several parties, including EEPA editors and reviewers, the Tennessee Department of Education, Tennessee Education Research Alliance, Association for Public Policy and Management, Association for Education Finance and Policy, and Inequality Seminar at The University of North Carolina at Chapel Hill. We are also grateful to Karin Gegenheimer for excellent research assistance. Corresponding author, Matthew G. Springer, can be reached at [mgspringer@unc.edu](mailto:mgspringer@unc.edu).


### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### ORCID iDs

Seth B. Hunter  <https://orcid.org/0000-0002-3051-872X>

Matthew G. Springer  <https://orcid.org/0000-0001-9759-0999>



## Notes

1. Rowan and Raudenbush (2016) offer a thorough historical and empirical treatment of teacher evaluation in American schools.

2. Work by Koedel et al. (2019) focuses on whether summative ratings influence teachers' self-reported, self-directed professional improvement activities, as measured by four items on a statewide teacher survey. Using a regression discontinuity design, they find no evidence that teachers alter their time investments in professional improvement or adjust their professional improvement activities based on evaluation feedback, or in response to their ratings.

3. This is not an exhaustive list of theoretically or empirically grounded feedback characteristics. For more characteristics, see DeNisi and Murphy (2017), Hattie and Timperley (2016), and Ilgen et al. (1979).

4. Some research also suggests that several feedback mediators exist (e.g., employee cognitive processes; DeNisi & Murphy, 2017; Jawahar, 2010; Kinicki et al., 2004). Although we might wish to see proof of these mediators in evaluation contexts, we assume that policymakers' foremost interest is in the presentation of critical feedback characteristics (CFCs), as feedback effects necessarily depend on them. Moreover, it is undoubtedly easier for a policy to support the provision of feedback characteristics than to support specific characteristics and mediational chains. Finally, research suggests net positive associations between employee performance and CFCs (DeNisi & Murphy, 2017; Jawahar, 2010; Kinicki et al., 2004).

5. Qualitative coders did not know whether the teacher, observer, or both set the goal. Although the performance benefits of self-set versus externally set goals for adults are not definitive, psychological research suggests either can lead to performance improvements (Harkins & Lowe, 2000; Locke & Latham, 2002).

6. More than 80% of teacher evaluators are principals or assistant principals. The remaining evaluators include teacher peer observers and full-time evaluators working out of central offices.

7. Emerging evidence suggests that administrators tend to assign more observations to white colleagues with more years of experience, higher prior-year administrator observation scores, and advanced degrees (Hunter & Rodriguez, 2021).

8. LOE-Cont for teachers of tested subjects in 2014–2015 is 50% observation scores, 35% Tennessee Value-Added Assessment System (TVAAS), and the remainder is determined by achievement scores. Weights applied to observation, growth, and achievement scores of teachers of untested subjects in 2014–2015 are 60, 25, and 15, respectively. LOE-Cont ranges from 100 to 500.

9. We test the robustness of our findings to district and school multiway clustered standard errors. Multiway clustered standard errors increase the precision of some estimates and decrease the precision of others (see online Supplementary Appendix B). Nonetheless, these results do not threaten our conclusions.

10. Although our use of reliability ratios is consistent with previous applications (e.g., Hardin et al., 2003), we also test whether *errors-in-variables* (EiV) estimates are sensitive to reliability ratios ranging from 0.2 to 1.0 by 0.1 increments. These estimates are qualitatively similar to our main EiV estimates, meaning that under no circumstances do EiV tests suggest that measurement error explains the null results in Table 5 (see online Supplementary Appendix C).

11. As estimated associations between observation scores and the evidence characteristic are similar, the primary findings are unaffected by measurement error. However, each of these associations is statistically significant, which seems to be an artifact of the bootstrapped standard error. We test the sensitivity of results in Tables 5–7 to bootstrapped standard errors in online Supplementary Appendix D. The only difference is the association between evidence and observation scores. We believe this is nothing more than a false positive, given that the bootstrapped result is statistically significant at the 5% level, and online Supplementary Appendix D includes more than 20 tests.

12. "Feedback orientation" is a multidimensional construct referring to an individual's receptivity to feedback, regardless of the objective information it contains (Linderbaum & Levy, 2010). Recipient's attitudes regarding negative feedback, mindful consideration of feedback, empathy, and value placed on feedback affect feedback orientation (London & Smither, 2002).

## References

- Alexander, K. (2016). *TEAM evaluator training* [Certification training, TEAM evaluator training 2016–17]. [http://team-tn.org/wp-content/uploads/2013/08/TEAM-Teacher-Training-2016\\_FINAL\\_PDF.pdf](http://team-tn.org/wp-content/uploads/2013/08/TEAM-Teacher-Training-2016_FINAL_PDF.pdf)
- Archer, J., Cantrell, S., Holtzman, S., Joe, J., Tocci, C., & Wood, J. (2016). *Better feedback for better teaching: A practical guide to improving classroom observations* (1st ed.). Jossey-Bass.
- Bacher-Hicks, A., Chin, M., Kane, T., & Staiger, D. (2017). *An evaluation of bias in three measures of teacher quality: Value-added, classroom observations, and student surveys* (No. w23478). National Bureau of Economic Research. <https://doi.org/10.3386/w23478>
- Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performance:

- Some problems in the design and implementation of evaluation systems. *Educational Researcher*, 44(2), 77–86. <https://doi.org/10.3102/0013189X15574904>
- Bartholomew, K., Henderson, A. J. Z., & Marcia, J. E. (2000). Coded semi-structured interviews in social psychological research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (pp. 286–312). Cambridge University Press.
- Belur, J., Tompson, L., Thornton, A., & Simon, M. (2021). Interrater reliability in systematic review methodology: Exploring variation in coder decision-making. *Sociological Methods & Research*, 50, 837–865. <https://doi.org/10.1177/0049124118799372>
- Bleiberg, J., & Harbatkin, E. (2020). Teacher evaluation reform: A convergence of federal and local forces. *Educational Policy*, 34(6), 918–952.
- Brinko, K. T. (1993). The practice of giving feedback to improve teaching. *Journal of Higher Education*, 64(5), 574–593.
- Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). Macmillan.
- Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal*, 55, 1233–1267. <https://doi.org/10.3102/0002831218776216>
- Cannata, M., Rubin, M., Goldring, E., Grissom, J. A., Neumerski, C. M., Drake, T. A., & Schuermann, P. (2017). Using teacher effectiveness data for information-rich hiring. *Educational Administration Quarterly*, 53(2), 180–222.
- Cannon, M. D., & Witherspoon, R. (2005). Actionable feedback: Unlocking the power of learning and performance improvement. *Academy of Management Perspectives*, 19(2), 120–134. <https://doi.org/10.5465/ame.2005.16965107>
- Carey, J. W., & Gelaude, D. (2008). Systematic methods for collecting and analyzing multidisciplinary team-based qualitative data. In G. Guest & K. M. MacQueen (Eds.), *Handbook for team-based qualitative research* (pp. 227–274). Altamira Press.
- 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.
- Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). The long term impacts of teachers: Teacher value added and student outcomes in adulthood. *American Economic Review*, 104(9), 2633–2679.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.
- Cohen-Vogel, L. (2011). “Staffing to the test”: Are today’s school personnel practices evidence based? *Educational Evaluation and Policy Analysis*, 33(4), 483–505. <https://doi.org/10.3102/0162373711419845>
- Cohen-Vogel, L., Tichnor-Wagner, A., Allen, D., Harrison, C., Kainz, K., Socol, A. R., & Wang, Q. (2015). Implementing educational innovations at scale: Transforming researchers into continuous improvement scientists. *Educational Policy*, 29(1), 257–277. <https://doi.org/10.1177/0895904814560886>
- Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428), 1314–1328.
- Cornelius, K. E., & Nagro, S. A. (2014). Evaluating the evidence base of performance feedback in pre-service special education teacher training. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, 37(2), 133–146. <https://doi.org/10.1177/0888406414521837>
- Daley, G., & Kim, L. (2010). *A teacher evaluation system that works* [Working paper]. National Institute for Excellence in Teaching.
- de Barros, A. (2019). *Evaluating teacher evaluation: Evidence from Chile. Organization of schools and systems & education in global contexts*. Society for Research in Educational Effectiveness.
- Dee, T. S., & Cohodes, S. R. (2008). Out-of-field teachers and student achievement: evidence from matched-pairs comparisons. *Public Finance Review*, 36(1), 7–32. <https://doi.org/10.1177/109114210628933032>
- DeNisi, A. S., & Murphy, K. R. (2017). Performance appraisal and performance management: 100 years of progress? *Journal of Applied Psychology*, 102(3), 421–433. <https://doi.org/10.1037/apl0000085>.supp
- Doan, S. (2019). *What do classroom observation scores tell us about student success? Capturing the impact of teachers using at-scale classroom observation scores*. Vanderbilt University. <https://ir.vanderbilt.edu/handle/1803/15448>
- Donaldson, M. L. (2021). *Multidisciplinary perspectives on teacher evaluation: Understanding the research and theory* (1st ed.). Routledge.
- Donaldson, M. L., Cobb, C. D., LeChasseur, K., Gabriel, R., Gonzales, R., Woulfin, S., & Makuch, A. (2014). *An evaluation of the pilot implementation of Connecticut’s system for educator evaluation*

- and development. UConn Center for Education Policy Analysis.
- 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>
- Feeney, E. J. (2007). Quality feedback: The essential ingredient for teacher success. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 80(4), 191–198. <https://doi.org/10.3200/TCHS.80.4.191-198>
- Fesler, L., Dee, T., Baker, R., & Evans, B. (2019). Text as data methods for education research. *Journal of Research on Educational Effectiveness*, 12(4), 707–727. <https://doi.org/10.1080/19345747.2019.1634168>
- Finster, M., & Milanowski, A. (2018). Teacher perceptions of a new performance evaluation system and their influence on practice: A within- and between-school level analysis. *Education Policy Analysis Archives*, 26, 41. <https://doi.org/10.14507/epaa.26.3500>
- Gallo, A. (2011, February 7). Making sure your employees succeed. *Harvard Business Review*. <https://hbr.org/2011/02/making-sure-your-employees-suc>
- Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, 38(4), 915–945. <https://doi.org/10.3102/00028312038004915>
- 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>
- Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy*, 11(2), 125–149. [https://doi.org/10.1162/EDFP\\_a\\_00180](https://doi.org/10.1162/EDFP_a_00180)
- Glickman, C., Gordon, S., & Ross-Gordon, J. (2018). *Supervision and instructional leadership* (10th ed.). Pearson.
- Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals’ human capital decisions and the emergence of teacher observation data. *Educational Researcher*, 44(2), 96–104. <https://doi.org/10.3102/0013189X15575031>
- Grissom, J. A., & Bartanen, B. (2018). Strategic retention: Principal effectiveness and teacher turnover in multiple-measure teacher evaluation systems. *American Educational Research Journal*, 56(2), 514–555. <https://doi.org/10.3102/0002831218797931>
- Grissom, J. A., Kalogrides, D., & Loeb, S. (2017). Strategic staffing? How performance pressures affect the distribution of teachers within schools and resulting student achievement. *American Educational Research Journal*, 54(6), 1079–1116. <https://doi.org/10.3102/0002831217716301>
- Grissom, J. A., & Loeb, S. (2017). Assessing principals’ assessments: Subjective evaluations of teacher effectiveness in low- and high-stakes environments. *Education Finance and Policy*, 12, 369–395.
- Hagen, M. S. (2012). Managerial coaching: A review of the literature. *Performance Improvement Quarterly*, 24(4), 17–39. <https://doi.org/10.1002/piq>
- Hardin, J. W., & Carroll, R. J. (2003). Measurement error, GLMs, and notational conventions. *The Stata Journal*, 3(3), 329–341.
- Hardin, J. W., Schmiediche, H., & Carroll, R. J. (2003). The simulation extrapolation method for fitting generalized linear models with additive measurement error. *The Stata Journal*, 3(4), 373–385.
- Harkins, S. G., & Lowe, M. D. (2000). The effects of self-set goals on task performance. *Journal of Applied Social Psychology*, 30(1), 1–40. <https://doi.org/10.1111/j.1559-1816.2000.tb02303.x>
- Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, 95(7–8), 798–812. <https://doi.org/10.1016/j.jpubeco.2010.11.009>
- Harris, D. N., & Sass, T. R. (2014). Skills, productivity and the evaluation of teacher performance. *Economics of Education Review*, 40, 183–204. <https://doi.org/10.1016/j.econedurev.2014.03.002>
- Hattie, J., & Timperley, H. (2016). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Hemmeter, M. L., Snyder, P., Kinder, K., & Artman, K. (2011). Impact of performance feedback delivered via electronic mail on preschool teachers’ use of descriptive praise. *Early Childhood Research Quarterly*, 26(1), 96–109. <https://doi.org/10.1016/j.ecresq.2010.05.004>
- Hill, H. C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, 83(2), 371–385. <https://doi.org/10.1017/CBO9781107415324.004>
- Hitt, D. H., & Tucker, P. D. (2016). Systematic review of key leader practices found to influence student achievement: A unified framework. *Review of Educational Research*, 86(2), 531–569.
- Huff, J., Preston, C., & Goldring, E. (2013). Implementation of a coaching program for school

- principals: Evaluating coaches' strategies and the results. *Educational Management Administration & Leadership*, 41(4), 504–526. <https://doi.org/10.1177/1741143213485467>
- Hunter, S. B. (2019). *The effects of more frequent observations on student achievement scores* (No. 2019–04; Strengthening Tennessee's Education Labor Market). Tennessee Education Research Alliance. [https://peabody.vanderbilt.edu/TERA/files/TERA\\_Working\\_Paper\\_2019-04.pdf](https://peabody.vanderbilt.edu/TERA/files/TERA_Working_Paper_2019-04.pdf)
- Hunter, S. B. (2020). The unintended effects of policy-assigned teacher observations: Examining the validity of observation scores. *AERA Open*, 6(2). Advance online publication. <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., & Ege, A. (2021). Linking student outcomes to school administrator discretion in the implementation of teacher observations. *Educational Administration Quarterly*, 57, 607–640. <https://doi.org/10.1177/0013161X211003134>
- 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, 739–758.
- Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, 64(4), 349–371. <https://doi.org/10.1037/0021-9010.64.4.349>
- Ivancevich, J. M. (1982). Subordinates' reactions to performance appraisal interviews: A test of feedback and goal-setting techniques. *Journal of Applied Psychology*, 67(5), 581–587. <https://doi.org/10.1037/0021-9010.67.5.581>
- Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non-test score outcomes. *Journal of Political Economy*, 126(5), 36.
- Jawahar, I. M. (2010). The mediating role of appraisal feedback reactions on the relationship between rater feedback-related behaviors and ratee performance. *Group & Organization Management*, 35(4), 494–526. <https://doi.org/10.1177/1059601110378294>
- Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *The Journal of Human Resources*, 46(3), 587–613.
- Kimball, S. M. (2003). Analysis of feedback, enabling conditions and fairness perceptions of teachers in three school districts with new standards-based evaluation systems. *Journal of Personnel Evaluation in Education*, 16(4), 241–268. <https://doi.org/10.1023/A:1021787806189>
- Kimball, S. M., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly*, 45(1), 34–70.
- Kinicki, A. J., Prussia, G. E., Wu, B. J., & McKee-Ryan, F. M. (2004). A covariance structure analysis of employees' response to performance feedback. *Journal of Applied Psychology*, 89(6), 1057–1069. <https://doi.org/10.1037/0021-9010.89.6.1057>
- Koedel, C., Li, J., Springer, M. G., & Tan, L. (2019). Teacher performance ratings and professional improvement. *Journal of Research on Educational Effectiveness*, 12(1), 90–115.
- Kraft, M. A., & Christian, A. (2021). Can teacher evaluation systems produce high-quality feedback? An administrator training field experiment. *American Educational Research Journal*. Advance online publication. <https://doi.org/10.3102/00028312211024603>
- Kraft, M. A., & Gilmour, A. F. (2016). Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly*, 52(5), 711–753. <https://doi.org/10.1177/0013161X16653445>
- 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.
- Ladd, H. F., & Sorensen, L. C. (2017). Returns to teacher experience: Student achievement and motivation in middle school. *Education Finance and Policy*, 12(2), 241–279. [https://doi.org/10.1162/EDFP\\_a\\_00194](https://doi.org/10.1162/EDFP_a_00194)
- Linderbaum, B. A., & Levy, P. E. (2010). The development and validation of the Feedback Orientation Scale (FOS). *Journal of Management*, 36(6), 1372–1405.
- Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, 31(1), 61–95. <https://doi.org/10.1007/s11092-018-09291-3>
- 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>
- Lockwood, J. R., & McCaffrey, D. F. (2020). Recommendations about estimating errors-in-variables regression in Stata. *The Stata Journal: Promoting Communications on Statistics and*



- Stata, 20(1), 116–130. <https://doi.org/10.1177/1536867X20909692>
- 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* [Doctoral dissertation]. The Pennsylvania State University. [https://etda.libraries.psu.edu/files/final\\_submissions/18672](https://etda.libraries.psu.edu/files/final_submissions/18672)
- Marques, J. F., & McCall, C. (2005). The application of interrater reliability as a solidification instrument in a phenomenological study. *The Qualitative Report*, 10(3), 449.
- 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>
- National Council on Teacher Quality. (2019). *NCTQ: Yearbook: State teacher policy database*. <https://www.nctq.org/yearbook/home>
- Neumerski, C. M., Grissom, J. A., Goldring, E., Drake, T. A., Rubin, M., Cannata, M., & Schuermann, P. (2018). Restructuring instructional leadership: How multiple-measure teacher evaluation systems are redefining the role of the school principal. *The Elementary School Journal*, 119(2), 28.
- Odden, A., & Kelly, J. A. (2008). *Strategic management of human capital in public education*. Strategic Management of Human Capital.
- Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, 82(1), 123–141. <https://doi.org/10.17763/haer.82.1.v40p0833345w6384>
- Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, 130, 105–119. <https://doi.org/10.1016/j.jpubeco.2015.02.008>
- Penuel, W. R., Fishman, B. J., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional strategies development effective? Strategies that foster curriculum implementation. *American Educational Research Journal*, 44(4), 921–958. <https://doi.org/10.3102/0002831207308221>
- Phipps, A. R. (2018). *Incentive contracts in complex environments: Theory and evidence on effective teacher performance incentives* [Doctoral dissertation]. LibraETD, University of Virginia. [https://libraetd.lib.virginia.edu/public\\_view/qn59q448d](https://libraetd.lib.virginia.edu/public_view/qn59q448d)
- Phipps, A. R., & Wiseman, E. A. (2021). Enacting the rubric: Teacher improvements in windows of high-stakes observation. *Education Finance and Policy*, 16(2), 283–312. [https://doi.org/10.1162/edfp\\_a\\_00295](https://doi.org/10.1162/edfp_a_00295)
- Rigby, J. G. (2015). Principals' sensemaking and enactment of teacher evaluation. *Journal of Educational Administration*, 53(3), 374–392. <https://doi.org/10.1108/JEA-04-2014-0051>
- Rowan, B., & Raudenbush, S. W. (2016). Teacher evaluation in American schools. In *Handbook of Research on Teaching* (5th ed., pp. 1159–1216). American Educational Research Association.
- Sanders, W. L., & Rivers, J. C. (1996). Cumulative and residual effects on teachers on future student academic achievement. *Research Progress Report*. Knoxville: University of Tennessee Value-Added Research and Assessment Center. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.474.3738&rep=rep1&type=pdf>
- Sartain, L., Zou, A., Gutierrez, V., Shyja, A., Hinton, E., Brown, E. R., & Easton, J. W. (2020). *Teacher evaluation in CPS: Perceptions of REACH implementation, five years in*. Research Brief. University of Chicago Consortium on School Research.
- SAS Analytics Software. (2015). *Technical documentation for 2015 TVAAS analyses 1.1*. <https://www.tn.gov/education/data/tvaas.html>
- Scheeler, M. C., Ruhl, K. L., & McAfee, J. K. (2004). Providing performance feedback to teachers: A review. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, 27(4), 396–407. <https://doi.org/10.1177/088840640402700407>
- Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*, 11(3), 340–359. [https://doi.org/10.1162/EDFP\\_a\\_00186](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, 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](https://doi.org/10.1162/EDFP_a_00173)
- Sun, M., Muteson, R. B., & Kim, J. (2016). Teachers' use of evaluation for instructional improvement and school supports for such use. In J. Grissom & P. Youngs (Eds.), *Improving teacher*



- evaluation systems: Making the most of multiple measures* (1st ed., pp. 102–116). Teachers College Press.
- Teacher and Principal Evaluation Policy. (2013). [https://web.archive.org/web/20140123232134/http://www.tn.gov/sbe/Policies/5.201\\_Teacher\\_and\\_Principal\\_Evaluation\\_Policy\\_11-5-13.pdf](https://web.archive.org/web/20140123232134/http://www.tn.gov/sbe/Policies/5.201_Teacher_and_Principal_Evaluation_Policy_11-5-13.pdf)
- Tennessee Department of Education. (2016). *Evaluation—TEAM-TN*. <https://web.archive.org/web/20161105020918/http://team-tn.org/evaluation/>
- Thurlings, M., Vermeulen, M., Bastiaens, T., & Stijnen, S. (2013). Understanding feedback: A learning theory perspective. *Educational Research Review*, 9, 1–15. <https://doi.org/10.1016/j.edurev.2012.11.004>
- Tuytens, M., & Devos, G. (2011). Stimulating professional learning through teacher evaluation: An impossible task for the school leader? *Teaching and Teacher Education*, 27(5), 891–899. <https://doi.org/10.1016/j.tate.2011.02.004>
- Tuytens, M., & Devos, G. (2013). How to activate teachers through teacher evaluation? *School Effectiveness and School Improvement*, 25(4), 509–530. <https://doi.org/10.1080/09243453.2013.842601>
- Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, 73(1), 89–122. <https://doi.org/10.3102/00346543073001089>
- Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget Effect*. [http://tntp.org/assets/documents/TheWidgetEffect\\_2nd\\_ed.pdf](http://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf)

## Authors

SETH B. HUNTER is an assistant professor of education leadership at George Mason University. His research focuses on teacher leadership and the policies and practices of educator evaluation.

MATTHEW G. SPRINGER is the Robena and Walter E. Hussman, Jr. Distinguished Professor and chair of the Educational Policy and Organizational Leadership at The University of North Carolina at Chapel Hill. His research focuses on accountability, compensation, and incentives.

Manuscript received August 15, 2019

First revision received September August 31, 2020

Second revision received September 3, 2021

Third revision received October 19, 2021

Accepted October 26, 2021