

Characteristics of Effective Feedback in Teacher Evaluation

Jihyun Kim¹ · Xintong Li² · Christi Bergin³

Jihyun Kim and Xintong Li are co-equal first authors. *

Jihyun Kim jihyunkim@sungshin.ac.kr

Xintong Li lixint@missouri.edu

Christi Bergin berginc@missouri.edu

¹ Sungshin Women's University, C836, Soojung Building, 2. Bomun-Ro 34da-Gil, Seongbuk-Gu, Seoul 02844, Korea

² University of Missouri, 2800 Maguire blvd, Columbia, MO 65211, USA

³ University of Missouri, 109 Hill Hall, Columbia, MO 65211, USA

Abstract

One key mechanism through which evaluation facilitates teachers' growth is principals' feedback during teacher evaluation. In this study, we examined the relationship between teachers' perceptions of feedback quality and improvement in their instructional practices. We used student report – a meaningful but rarely used measure – to measure quality of instruction of two important instructional practices; promotion of cognitive engagement and critical thinking. We used teacher report to measure five characteristics of feedback. We found that feedback that focused on strengths was associated with higher quality of both instructional practices, whereas focus on ways to improve teaching was associated only with promotion of critical thinking. Whether feedback was immediate and face-to-face did not predict higher quality instructional practices.

The *Every Student Succeeds Act* (ESSA) was signed into law in 2015, replacing *No Child Left Behind* (NCLB) in the United States. A major change in the new act was that states were given more autonomy and responsibility in building their own accountability systems and identifying evidence-based interventions to improve student learning (Darling-Hammond et al., 2016; Duff & Wohlstetter, 2019; Egalite et al., 2017). To address this responsibility, many states sought ways to improve principals' capacity as instructional leaders (Reid et al., 2020; Cherasaro et al., 2016). One area of focus is principals' feedback to teachers during teacher evaluation because this is a key mechanism that facilitates teachers' growth (Donaldson & Papay, 2015).

In a variety of settings, performance feedback is an effective form of intervention with large effect sizes, yet the quality of feedback is uneven (Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Sleiman et al., 2020). In teacher evaluation settings, such variable quality of feedback may be one explanation for why the effects of teacher evaluation on various educational outcomes are still inconclusive (Feeney, 2007; Stecher et al., 2018). This study seeks to advance understanding of the characteristics of principal feedback associated with high-quality instruction, which has been rarely studied (Lavigne & Good, 2015). The field needs such understanding in order to inform principals' instructional leadership. We first briefly review literature on how performance feedback in general shapes the action of feedback recipients. We then review literature specific to teacher evaluation and principal feedback with a focus on psychological aspects of teacher evaluation.

Performance Feedback in General Settings

Feedback is considered an important tool for improving individual's performance in many contexts. Indeed, feedback is an integral element of several theories of learning and motivation, such as goal setting theory (Locke & Latham, 1990), social cognitive theory (Bandura, 1991) and expectancy-value theory (Eccles et al., 1983). Yet, there are few theories that specifically explain how feedback during a performance evaluation improves performance, and therefore how to optimize feedback. This is important because feedback is not always effective. The success of feedback depends on characteristics of feedback, tasks, feedback providers and receivers, and other factors (Alvero et al., 2001; Kluger & DeNisi, 1996; Lechermeier & Fassnacht, 2018; Thurlings et al., 2012; Thurlings et al., 2013).

In one common view, feedback is regarded as a motivated learning process that aims to reduce discrepancies between current performance and a desired performance goal (Hattie & Timperley, 2007; Kluger & DeNisi, 1996). From this perspective, feedback quality is influenced by how well three questions (i.e., “where am I going,” “how am I going,” and “where to next”) are answered (Hattie & Timperley, 2007). The first question “where am I going” is related to goal setting, which provides information regarding the criteria for success (Locke & Latham, 1990). The second question “how am I doing” is related to strategies to achieve goals and progress toward goals. The third question “where to next” provides information regarding next steps after the current task (Hattie & Timperley, 2007).

Feedback is processed in four steps: perceived feedback, acceptance of feedback, desire to respond to feedback, and the intended response. Characteristics of feedback can influence effectiveness at each step. Perceived feedback is influenced by timing in that feedback delay reduces effectiveness in behavior-feedback chains due to the limited memory of one's behavior

(Ammons, 1956, Ilgen et al., 1979). Feedback is more readily perceived and accepted if it is positive and matches one's self-concept (Ilgen et al., 1979; Lechermeier & Fassnacht, 2018). This positive reinforcement increases recipients' desire to respond to feedback (Thorndike, 1913) and lead to recipients' setting specific and challenging goals that would lead to improved performance (Locke, 1968; Locke & Latham, 1990). Recipients' response to feedback is also influenced by their expectation for success (Eccles et al., 1983). While feedback from a trusted other may shape recipients' expectancies for success, so does their own self-efficacy. This, in turn, affects the likelihood of taking action based on feedback.

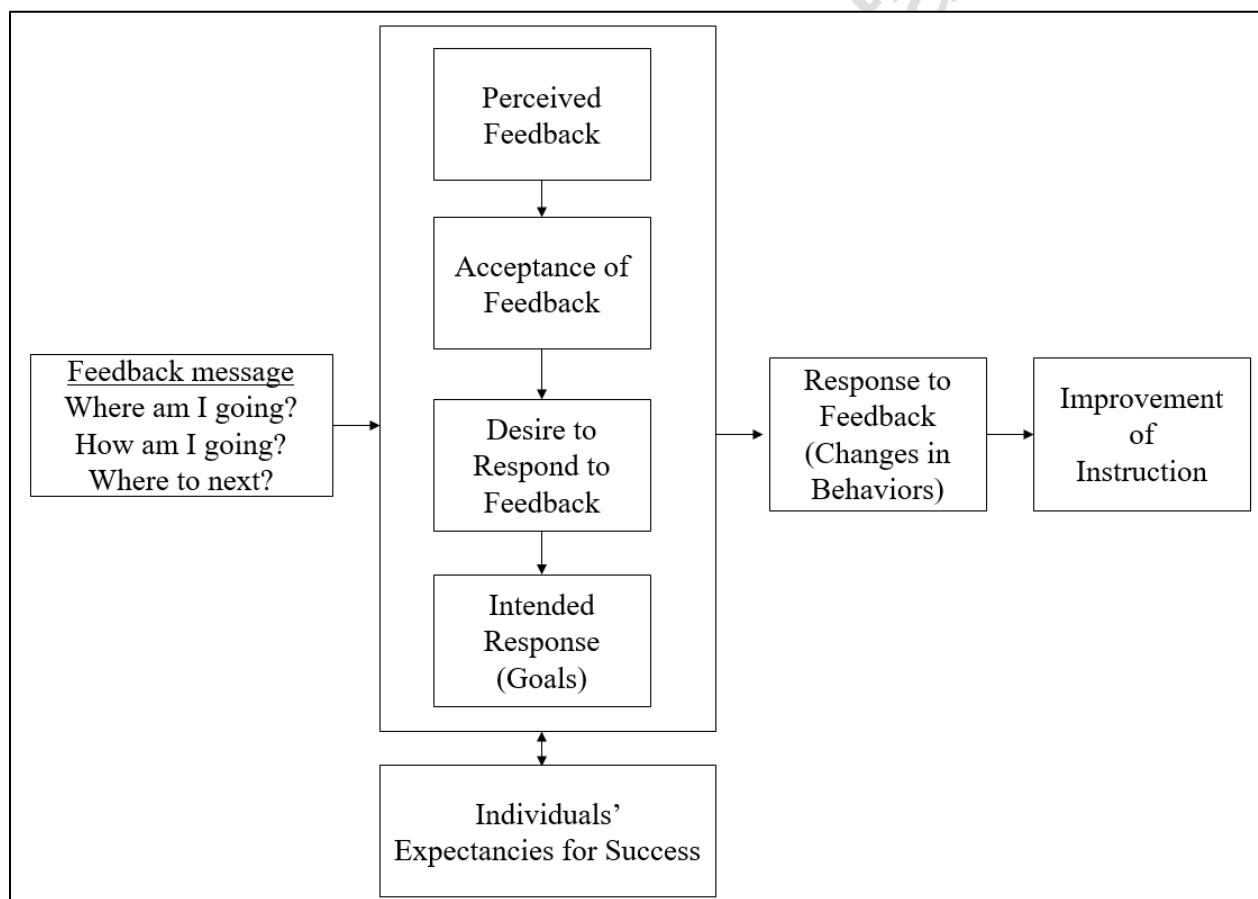


Figure 1. Framework for feedback effects on improvement of instruction.

Note. This is modified from Ilgen and colleagues (1979) and Hattie & Timperley (2007).

Performance Feedback in Teacher Evaluation

We combine these perspectives in a framework (see Figure 1) to suggest how feedback from principals may improve teachers' practices, which we discuss next. In the present study, we do not test the framework per se, but rather present it to plausibly explain the effect of feedback on teaching quality.

Many teacher evaluation systems in the U.S. and other countries use classroom observation by principals as a primary source of data and require post-observation conferences wherein principals provide feedback to teachers (e.g., Doherty & Jacobs, 2013; Shaked, 2018). Principal observations are a potentially powerful tool for improving teaching effectiveness when used to inform teacher professional development through feedback (Marzano & Toth, 2013; Phipps & Wiseman, 2021; Toch & Rothman, 2008). Empirically, rich, face-to-face feedback given to teachers after detailed classroom observations is regarded as an important component of effective teacher evaluation (Taylor & Tyler, 2012). It can increase teachers' knowledge of effective teaching (Lavigne & Good, 2015). Indeed, teachers report that feedback following classroom observations is the most valuable process for their professional development (Reinhorn et al., 2017).

In most cases, the format of feedback sessions, such as the number of meetings between evaluator and teachers and when the sessions must happen, is determined by district- or state-level policies (Reid et al., 2020). Observers generally discuss what they observed and provide suggestions for improvement based on a given observation rubric. Although the content of the feedback sessions may be shaped by the observation rubric (Halverson et al., 2004), the substance of sessions (i.e., what to focus on during the feedback session) largely depends on the evaluator, which leaves room for substantial variation in effectiveness.

Accordingly, there is a growing interest in understanding what makes feedback effective in teacher evaluation. Previous studies suggest that focused and frequent feedback promotes improved instruction quality and/or student achievement (Donaldson, 2021; Garet et al., 2017; Steinberg & Sartain, 2015). Previous studies also suggest that feedback should be based on descriptive and observable data (Danielson & McGreal, 2000; Hunter & Springer, 2022), promote reflective inquiry supported by evidence (Glickman, 2002), clarify attributes of high-quality instruction (Danielson, 1996; Marzano et al., 2001), and be immediate, specific, actionable, and non-penalizing (Curtis & Wiener, 2012; Delvaux et al., 2013; Hunter & Springer, 2022; Kraft & Christian, 2019; Tuma et al., 2018).

However, the extant literature generally relies on teacher perception data to measure both characteristics of feedback and teaching improvement, which could involve source bias (e.g., Cherasaro et al., 2016; Delvaux et al., 2013). That is, teachers who feel positive about their evaluation may perceive both the quality of feedback and the impact of the evaluation positively, regardless of objective feedback quality or improvement in instruction quality. An exception is Hunter and Springer (2022) who analyzed the association between written feedback that teachers received, their classroom observation scores, and their value-added measures. Unlike prior studies, they found little association between characteristics of feedback and teachers' improvement in instruction quality. Furthermore, the only feedback characteristic that had significant associations with instruction quality—goal-setting—was negatively associated with observation scores and positively associated with value-added measures. This suggests that the relationship between feedback characteristics and improvement in instruction quality found in previous studies may not be straightforward. A plausible explanation is that previous studies used limited indicators of both feedback quality and teaching practices. Both are multi-faceted

constructs which necessitate various measures (Hill & Grossman, 2013; Kraft et al., 2018). The field needs more empirical studies that focus on multiple fine-grained characteristics of feedback quality and use different measures of instruction quality.

In the present study, we measured the quality of feedback through teacher self-report, as most previous studies have done. We choose self-report because feedback quality may not be a fixed attribute of the feedback itself, but may depend instead on characteristics of feedback providers and recipients (Shute, 2008). That is, teachers actively assess feedback from their evaluators and decide whether to accept the feedback and desire to respond (Donaldson, 2021; see Figure 1). Empirically, for example, Kinicki and colleagues (2004) found that, if a feedback provider is viewed as trustworthy and competent, the recipients are likely to perceive the feedback as more accurate. This is important because perceiving feedback as accurate is the first step in a cognitive process that determines recipients' acceptance of feedback and, in turn, their future performance (Kinicki et al., 2004). Cherasaro and colleagues (2016) reported that accuracy and evaluator credibility predicted teachers' perceived usefulness of the feedback, and evaluator credibility most powerfully shaped teachers' responses to feedback. Credible evaluators have rich experiences in teaching, subject-matter knowledge, and enough opportunity to observe teachers (Delvaux et al., 2013). Similarly, Kraft and Christian (2019) found that teachers' perceptions of mutual respect and trust between themselves and their evaluators, as well as how much they enjoyed working with their evaluators, were predictive of teachers' perceptions of feedback quality. In addition, source credibility predicts recipients' motivation to or desire and intent to respond to the feedback and improve performance (Roberson & Stewart, 2006). Furthermore, feedback quality can be fluid, as it is based on the unfolding discussion between provider and receiver during feedback sessions. For example, Thurlings and colleagues

(2012) found that teachers were capable of transforming ineffective feedback patterns to effective ones as they gave feedback to one another in face-to-face peer groups. Taken together, teachers' perceived quality of feedback is important to understand what constitutes effective feedback.

Based on the extant literature and the theoretical framework in Figure 1, our study focused on five characteristics of feedback: (1) face-to-face and immediate, (2) specific regarding ways to improve teaching, (3) useful and relevant regarding ways to improve teaching, (4) specific regarding strengths, and (5) useful and relevant regarding strengths. We hypothesized that face-to-face and immediate feedback would lead to greater improvement of instruction because it occurs while memory is fresh and because in-person communication facilitates clarifying questions and explanations. We hypothesized that characteristics 2 and 3, which reflect feedback on ways to improve teaching, would also lead to greater improvement of instruction because they address the questions "where am I going, how am I doing, and where do I go next?" However, they may be less effective than characteristics 4 and 5 which reflect feedback on a teacher's existing strengths. A focus on strengths may increase acceptance of feedback and boost individuals' expectancies for success. Exploring these hypotheses is important because principals need guidance on whether it is more effective to point out what teachers need to do to improve versus pointing out what they are already doing well. Characteristics 2 and 4 compared to 3 and 5 reflect teachers' evaluation of feedback's specificity versus relevance and usefulness. Such nuances in teachers' perceptions of feedback have not been previously studied as they co-occur, yet are relevant to teacher's acceptance of feedback and expectations for success, which should lead to improving teacher evaluation practice.

Student Surveys as a Measure of Instruction Quality

We measured the quality of instructional practices through student report. In contrast, previous studies typically used teacher self-report (potentially resulting in source bias), but sometimes used student achievement or classroom observation data (Garet et al., 2017; Steinberg & Sartain, 2015). It is important to note that various measures of teaching quality, such as classroom observation, value-added scores, and student survey measures often assess different aspects of teaching. However, use of achievement data such as student test scores or teachers' value-added scores to measure teaching effectiveness have well-documented problems. Among these are technical issues of how estimates are derived, the non-random assignment of students to teachers, the limited availability of tests in all subjects and grades, and the inability to provide useful feedback to teachers to guide instructional improvement (Briggs & Domingue, 2011; Hallinger et al., 2014; Hill et al., 2011; Rothstein, 2010). As a result, organizations such as the National Association of Secondary School Principals and the American Statistical Association have taken a stand against value-added measures.

Student ratings have several advantages. They are based on more data points over time and aggregated across many students, which helps attenuate measurement error. In contrast, classroom observations are necessarily limited to brief snapshots of teaching, and the presence of an observer can alter classroom dynamics. Students have first-hand experience with teachers' instructional practices and can provide information not available to other observers. As a result of these advantages, student ratings are generally reliable and predictive of student learning (Tsai et al., 2022; Downer et al., 2015; Feldlaufer et al., 1988; Fraser & O'Brien, 1985; Kane & Staiger, 2012; Little et al., 2009; Marsh et al., 2019; Polikoff, 2015). For example, the Measures of Effective Teaching study found that student ratings of teachers predicted students' growth in

test scores and that they are often more reliable than achievement data or classroom observation data (Kane & Staiger, 2012). Recognizing these advantages, many states and districts have included student surveys as a component of their teacher evaluation systems (e.g., Missouri, Memphis, Chicago, Pittsburgh, and the New Teacher Project). In the coming years, we expect to see an increase in educational research that uses student survey data to measure instruction quality.

Study Context: Network for Educator Effectiveness

Our data came from a teacher evaluation system called the Network for Educator Effectiveness (NEE; www.neeadvantage.com), which is used in 295 districts in Missouri enrolling a total of 35,683 teachers and 360,056 students. NEE is a comprehensive, teacher evaluation system with the mission of promoting growth in teaching effectiveness. NEE meets the recommendations for an optimal evaluation system, according to Darling-Hammond and colleagues (2012), because it collects multiple sources of data with multiple observations by trained evaluators who provide meaningful and timely feedback (Jones & Bergin, 2019).

Classroom Observations and Feedback Sessions

Principals score teachers based on the NEE Classroom Observation (NEE-CO) rubric, which was designed for use in preK-12 classrooms across all subjects. The NEE-CO rubric is based on the nationally recognized InTASC Standards for Teaching (Interstate Teacher Assessment and Support Consortium, 2011) as adapted by the Missouri Department of Elementary and Secondary Education. It includes 27 observable teaching practices, such as promoting students' cognitive engagement (CE) and problem-solving and critical thinking (PCT). Districts select only four to six teaching practices to focus on, based on their priorities. The NEE-CO rubric is analytic, rather than holistic, which means principals award separate

scores for each teaching practice, so a teacher may score a “2” on *Problem-Solving and Critical Thinking* but a “6” on *Cognitive Engagement*. Each teaching practice is scored on a scale ranging from “0” (not present) to “7” (perfect exemplar) based on evidence the principal observes (Li & Baker, 2018). Higher scores on the rubric reflect greater frequency of high-quality instructional practices involving more students (i.e., “almost all” versus “only a few” students) when observed during a teaching session. NEE also provides principals with language to use when providing feedback to teachers. For example, for *Cognitive Engagement*, examples of feedback include the teacher “reviews frequently and spirals content” and “consistently encourages extension of discovery/play.”

NEE recommends that principals make several brief (approximately 10 minutes) unannounced visits every year for every teacher, scoring the focal four to six teaching practices per visit. The principal is expected to have a feedback conversation with the teacher within a few days of each visit. The scores are used formatively by principals and teachers to discuss each targeted teaching practice. A summative report is generated for every teacher at the end of each year, where a summative classroom observation score, a student survey score (if available), and written comments from the principal are included.

In this study, we focus on two teaching practices: promotion of cognitive engagement (CE) and promotion of problem solving and critical thinking (PCT). We chose CE and PCT because (1) they are high-impact teaching practices that predict student learning, and (2) they were the most frequently assessed teaching practices by NEE school districts, meaning that many districts prioritized improving these teaching practices.

Cognitive Engagement (CE)

In the NEE system, *Cognitive Engagement* refers to active mental involvement by students in learning activities, such as meaningful processing, strategy use, concentration, and metacognition (Wang et al., 2014; Fredricks et al., 2004; Wang & Degol, 2014). It predicts students' well-being and academic achievement (Li & Baker, 2018; Fredricks et al., 2004; Metallidou & Vlachou, 2007; Pietarinen et al., 2014; Reyes et al., 2012). Teachers can promote CE by using advance organizers, KWL (Know, Want, Learned) charts, share-outs, shoulder-partner talk, and authentic examples, as well as by connecting instruction/activities with students' lives, showing relevance, presenting a puzzling problem, and inviting responses from all students.

Problem Solving and Critical Thinking (PCT)

In the NEE system, *Problem Solving and Critical Thinking* refers to students skillfully applying, analyzing, synthesizing, and evaluating information to reach a conclusion or solve a problem (McCormick et al., 2015). PCT predicts students' achievement across age groups and subject areas (Fong et al., 2017; Giancarlo et al., 2004; Von Secker & Lissitz, 1999; Yu et al., 2010). Teachers can promote PCT by having students explain or justify their thinking, evaluate others' thinking, formulate challenging questions, make predictions, develop creative solutions, determine what makes an argument valid, assess possible solutions, categorize problems, and map concepts (Wirkala & Kuhn, 2011).

Principal Training

Principals are required to attend a three-day face-to-face certification training before they are given access to the NEE system. Principals' training focuses on how to use the NEE-CO rubric for accurate and consistent scoring as well as how to provide constructive feedback based on classroom observations. Principals return annually for a one-day, face-to-face re-certification

training. Principals are given an end-of-training video-based test of scoring accuracy. Those who receive low scores can still conduct in-field observations, but they are required to use additional training materials and are provided with in-field support.

The training follows three approaches to increase accuracy of observation ratings (Chafouleas, 2011; Woehr & Huffcutt, 1994). First, in the “rater error” training approach, raters learn to recognize and avoid leniency and halo errors and to use the full scale. Raters are trained to begin with a rating of “3” and then only move up or down the scale if the evidence clearly justifies doing so. Second, in the “performance dimension” training approach, raters learn to understand specific teaching practices through didactic research review and discussion. Third, in the “practice-with-feedback” training approach, raters watch and rate selected videos of authentic classes. They then share their ratings and evidence in small-group and whole-group discussions. Discrepancies are then discussed with trainer guidance. A previous study found that trained principals using the NEE-CO rubric successfully differentiate among teachers (Jones & Bergin, 2019).

Method

Participants

We focused on districts where students rated teachers on CE and/or PCT in the 2017-2018 and 2018-2019 school years. We used two separate samples based on student survey outcomes (CE or PCT) with 617 overlapping cases. The CE sample included 81 schools, 1,010 teachers, and 43,318 student surveys. The PC sample included 103 schools, 1,214 teachers, and 50,132 student surveys. Teachers’ scores are the mean aggregation of their students’ ratings. In smaller buildings, teachers were typically rated by all of their students, but due to logistical

issues not all students in larger buildings rated their teachers. In those cases, student raters were randomly selected from a set of students that was determined by building leaders.

Participating districts in Missouri were diverse, serving both high- and low-income students in urban, suburban, and rural areas. Across the state, 73.0% of students were non-Hispanic White; 50.0% were eligible for free/reduced-priced lunch; and the mean proficiency rate on the English language arts (ELA) and math state tests was 45.3%. Our sample schools were representative of Missouri schools with 78.6% of students being non-Hispanic white and 37.4% of students had FRPL status, and the average proficiency rates in ELA and mathematics were 50.6%, which is slightly better than the state average.

Procedures and Measures

Feedback Characteristics - Teacher Survey

NEE administers the teacher surveys (NEE-TS) online at the end of each school year. Teachers were able to access the survey when they signed into the online NEE portal. They were notified at the beginning of the survey that their responses were “completely anonymous and no one in the district can view individual responses from any teacher.”

NEE-TS asks teachers about their perceptions of school leadership. Items are based on the 10 Professional Standards for Educational Leaders (PSEL; Reston, 2015) and the research literature on the characteristics of effective feedback (e.g., Baeder, 2017; Cherasaro et al., 2015; Cherasaro et al., 2016; Feeney, 2007; Kraft et al., 2018; Ovando, 2005; Reinhorn et al., 2017; Wayne et al., 2016).

Feedback characteristics were measured by five items on NEE-TS (Cronbach's alpha = .92) and rated on a four-point scale from “0” (strongly disagree) to “3” (strongly agree). Teachers’ responses to these five items were included in the models separately. The items

include “This principal typically provides me with face-to-face feedback within two working days of observing my classroom;” “This principal provides specific feedback to me regarding ways my teaching can improve (i.e., focused, detailed, concrete);” “This principal provides useful and relevant feedback to me regarding ways my teaching can improve;” “This principal provides specific feedback to me regarding areas of strength in my teaching (i.e., focused, detailed, concrete);” and “This principal provides useful and relevant feedback to me regarding areas of strength in my teaching.”

Quality of Instructional Practices – Student Survey

NEE administers the student surveys (TESS) online at the end of each school year to students in 4th through 12th grades. The online survey interface was designed to be intuitive and easy for students to use. Students complete the survey on internet-enabled devices during a specified time window. An adult other than the evaluated teacher administers the survey using standard scripts provided by NEE. Students are assured that their responses are anonymous and voluntary. Although NEE is not able to collect information on incomplete surveys, very few students refuse to evaluate their teachers according to principals. Students were encouraged to ask questions or request definitions of difficult words, but to avoid possibly influencing students’ responses the proctors were instructed not to interpret any survey items for the students.

The TESS is based on the same InTASC standards as the NEE-CO rubric and includes 25 teaching practices observable by students. Teaching effectiveness for Cognitive Engagement (CE) was measured by four items (Cronbach’s alpha = .91) on a 4-point scale from 0 (not at all true) to 3 (very true). Higher scores indicate the students report higher quality teaching. The items include “this teacher expects us to think a lot and concentrate in this class,” “this teacher’s lessons make us think deeply,” “this teacher’s lessons make us think the whole class time,” and

“this teacher wants us to ask questions during lessons.” Teaching effectiveness for PCT was also measured by four items (Cronbach's alpha = .90). The items include “this teacher waits a while before letting us answer questions, so we have time to think,” “this teacher makes us compare different ideas or things,” “this teacher makes us use what we learn to come up with ways to solve problems,” and “this teacher asks ‘how?’ and ‘why?’ questions to make us think more.” Previous research shows that CE and PCT can be distinguished with TESS by 4th graders and older (Tsai et al., 2022) and both predict students’ performance on state proficiency tests (Li, 2022). Moreover, neither has shown a biased measure of illusory halo or teacher-student relationships (Li et al., 2022).

Covariates

Teachers’ general teaching effectiveness and personal attributes could be correlated with both feedback quality and their CE and PCT ratings, which could introduce bias into our estimation (Frank, 2000). Hunter and Springer (2022) noted that it is plausible that principals adjust their feedback quality for teachers with different backgrounds. It is also plausible that teachers perceive the feedback differently based on their teaching experiences. To mitigate this concern, we included TESS scores in the 2017-18 school year (i.e., pre-measure) as a control for such potential confounders to reduce potential bias. Pre-measures are regarded as one of the most powerful ways to control for unobservables in observational studies (Cook et al., 2008). Additional covariates include teachers’ years of experience, grade level taught, and a binary indicator of whether a teacher taught a subject tested by state-standardized tests (i.e., mathematics, English language arts, science, or social studies).

An additional covariate is principals’ support for professional development (PD). Support for PD may play a role in improving teachers’ instructional practices and shaping their

perceptions of feedback quality (Kim et al., 2019). We used a latent factor score calculated based on four items from the Teacher Survey that measure principals' support for comprehensive professional development for staff using a four-point scale from "0" (strongly disagree) to "3" (strongly agree). The survey items include "This principal provides or locates resources I need for my teaching;" "This principal provides me with valid and meaningful professional development opportunities;" "This principal monitors the application of professional development in my instruction;" and "This principal knows what professional development I need" (Cronbach's alpha = .89).

Finally, we included school characteristics because teacher evaluation implementation can be affected by school context (Cohen et al., 2020; Marsh et al., 2017). Specifically, we included the percentage of students eligible for free or reduced-price lunch, total student enrollment, the percentage of White students, and the percentage of students with Individualized Education Programs (IEP) at each school. We approximated students' overall achievement levels using their proficiency rate on the Missouri Assessment Program (i.e., the average percentage of students at or above the proficiency-level at the school). This data is school-level because the student survey is anonymous; thus, we did not have data on student- or classroom-level demographics.

Analysis

We used two-level hierarchical linear modeling (HLM) where teachers are nested in school buildings. Student level was not included, as CE and PCT measures were aggregated at the teacher level. We ran the analysis for the CE and PCT samples separately. We analyzed two models beginning with an unconditional model, which only included the outcome variables and the residuals. The second set of analyses included feedback quality items one-by-one at the

individual teacher level. In all models, teachers' responses to five survey items were included separately.

$$\text{TeachingPractice}_{19ij} = \beta_{0j} + \beta_{1j}*(\text{FeedbackQuality}_{ij}) + \beta_{2j}*(\text{PD}_{ij}) \\ + \beta_{3j}*(\text{TeachingPractice}_{18ij}) + \beta_{4j}*(\text{TeachingGrade}_{ij}) + \beta_{5j}*(\text{TestedSub}_{ij}) + \beta_{6j}*(\text{Exp}_{ij}) + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(\text{FRL}_j) + \gamma_{02}*(\text{Enrollment}_j) + \gamma_{03}*(\text{WhitePCT}_j) \\ + \gamma_{04}*(\text{ProfRate}_j) + \gamma_{05}*(\text{IEPRate}_j) + \gamma_{06}*(\text{FeedbackQuality}_j) + u_{0j}$$

where

$\text{TeachingPractice}_{19ij}$ = TESS Teaching practice scores in CE or PCT of teacher i in school j received in 2018-2019 school year, which are factor scores based on the means of students' responses

$\text{TeachingPractice}_{18ij}$ = TESS Teaching practice scores in CE or PCT of teacher i in school j received in 2017-2018, which are factor scores based on the means of students' responses (i.e., pre-measure)

$\text{FeedbackQuality}_{ij}$ = Teachers' responses to the five survey items regarding feedback quality in 2017-2018 (included in the models separately)

PD_{ij} = NEE-TS Factor scores of four professional development support items

$\text{TeachingGrade}_{ij}$ = The mean grade level reported by students minus four so that the 4th grade, the lowest grade in the samples, was recoded as zero, and consequently the intercepts represent the ones for teachers rated by 4th graders

TestedSub_{ij}	=	A binary variable indicating whether the teacher taught one of the four subjects tested by the state-standardized tests, i.e., ELA, mathematics, social studies, or science
Exp_{ij}	=	Teachers' reported years of teaching experience
FRL_j	=	The school-level percentage of students eligible for free-or-reduced-priced lunch in 2019, provided by the Missouri Department of Elementary and Secondary Education (DESE). Missing school data in the 2019 DESE report were imputed using the most recent available data from previous years.
Enrollment_j	=	The number of K-12 enrollment of schools obtained from DESE
WhitePCT_j	=	The school-level percentage of White students obtained from DESE. Missing school data in 2019 DESE report were imputed using the most recent available data from previous years.
ProfRate_j	=	The average of school-level ELA and mathematics proficiency rates of the state test (Missouri Assessment Program) obtained from DESE
IEPRate_j	=	School-level incidents rate of IEPs obtained from DESE
$\text{FeedbackQuality}_{ij}$	=	School means of teachers' responses to five principal feedback quality items
r_{ij}	=	Level 1 residuals
u_{0j}	=	Level 2 residuals

All Level 2 variables are grand mean centered. The parameter tests were based on cluster-robust standard errors with HLM 8, which can adjust biases due to observation

dependence within clusters (Esarey & Menger, 2019). Table 1 summarizes the descriptive statistics of the variables that we included in the analysis for CE and PCT samples, respectively. The two samples of teachers are largely comparable. Teachers in the PCT sample tend to work at slightly smaller schools with slightly more white students and students with IEPs.

Table 1.
Descriptive Statistics

	CE sample				PCT sample			
	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX
	CE Level 1 (Teacher level) Variables (N=1,010)				PCT Level 1 (Teacher level) Variables (N=1,214)			
Feedback quality items								
Face-to-face and immediate	2.45	0.73	0	3	2.46	0.72	0	3
Specific regarding improving teaching	2.45	0.68	0	3	2.45	0.66	0	3
Useful and relevant regarding improving teaching	2.42	0.68	0	3	2.43	0.67	0	3
Specific regarding strength	2.50	0.65	0	3	2.52	0.62	0	3
Useful and relevant regarding strength	2.49	0.65	0	3	2.50	0.62		
PD	0	.95	-3.94	1.05	0	.95	-3.91	1.06
Teaching practice (CE) in 2019	0	.97	-4.06	2.11	0	.95	-3.75	2.78
Teaching practice (CE) in 2018	0	.96	-3.89	2.51	0	.95	-3.20	2.78
Teaching grade	2.92	1.77	0	6	2.77	1.83	0	6
Teaching a tested-subject	.58	.49	0	1	.58	.49	0	1
Years of teaching experience	11.87	8.05	0	46	12.57	7.88	1	46
	CE Level 2 (School level) Variables (N=81)				PCT Level 2 (School level) Variables (N=103)			
% of students eligible for FRL	41.48	19.56	5.90	85.60	41.62	18.50	7.50	91.30
Total student enrollment	644.98	442.8	83	2183	592.75	42.57	83	2011
% of White students	77.63	23.93	0	100	82.77	18.68	0	100
% of students with IEP	11.57	4	3.92	21.06	12.12	3.86	3.92	23.25
% of students scored at or above proficient level	47.98	13.08	14.2	73.75	48.03	11.67	14.70	75.90
Number of teachers per school	12.47	13.92	2	75	11.79	11.32	2	72

Results

Before we move on to our main HLM model, we used unconditional models for the outcome variable—teaching effectiveness measured at the end of the 2018-2019 school year for the CE and PCT samples (Table 2). For both samples, Level 2 (school level) random effects were significant, and ICCs (Interclass Correlation Coefficients) were .120 and .141, respectively, which justifies including Level 2 to analyze the outcomes.

Table 2.
Unconditional Models

CE Sample				
	Coeff	SE	df	p
Intercept, γ_{00}	-.068	.052	80	.192
Random effect	SD	Variance	df	p
Level 1, r	.914	.834		
Level 2, u_0	.338	.114	80	<.001
PCT Sample				
	Coeff	SE	df	p
Intercept, γ_{00}	-.032	.047	102	.494
Random effect	SD	Variance	df	p
Level 1, r	.885	.783		
Level 2, u_0	.359	.129	102	<.001

Tables 3 and 4 present the findings from the CE and PCT samples, respectively. In each table, Models 1 through 5 only included individual survey items one-by-one.

For both the CE and PCT samples, some aspects of teachers' perceived feedback quality was related to teachers' instructional practices after controlling for their pre-measures and other individual teacher- and school-level characteristics. Specifically, for both outcomes, feedback focused on strengths had positive and significant associations with students' ratings of teachers' instructional practices. For CE sample, when a teacher perceived that their principals' feedback

Table 3.*The Association Between Feedback Quality and Instructional Practices (CE Sample)*

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
Level 1 (Teacher)					
Face-to-face and immediate	0.003 (0.04)				
Specific regarding improving teaching		0.08 (0.04)			
Useful and relevant regarding improving teaching			0.07 (0.05)		
Specific regarding strength				0.12** (0.04)	
Useful and relevant regarding strength					0.09* (0.04)
PD	0.005 (0.03)	-0.03 (0.03)	-0.03 (0.04)	-0.05 (0.02)	-0.04 (0.03)
Pre-measure of teaching practice	0.57*** (0.03)	0.57*** (0.03)	0.57*** (0.03)	0.57*** (0.03)	0.57*** (0.03)
Grade level	0.05 (0.03)	0.05 (0.03)	0.05** (0.02)	0.05 (0.03)	0.05 (0.03)
Teaching a tested-subject	0.17*** (0.05)	0.18*** (0.05)	0.18*** (0.05)	0.18*** (0.05)	0.18*** (0.05)
Years of teaching experience	-0.007* (0.003)	-0.007* (0.003)	-0.007* (0.003)	-0.007* (0.003)	-0.007* (0.003)
Level 2 (School)					
% of students eligible for FRL	-0.002 (0.003)	-0.002 (0.003)	-0.002 (0.003)	-0.002 (0.003)	-0.002 (0.003)
Enrollment	0.00003 (0.0001)	0.00004 (0.0001)	0.00003 (0.0001)	0.00004 (0.0001)	0.00004 (0.0001)
% of White students	-0.001 (0.002)	-0.001 (0.002)	-0.001 (0.002)	-0.001 (0.002)	-0.001 (0.002)
% of students with IEP	-0.002 (0.01)	-0.003 (0.01)	-0.002 (0.01)	-0.003 (0.01)	-0.003 (0.01)
% of students scored at or above proficient level	-0.0001 (0.006)	-0.001 (0.006)	-0.001 (0.006)	-0.001 (0.006)	-0.001 (0.006)
Intercept	-0.19 (0.12)	-0.37** (0.12)	-0.35** (0.12)	-0.47*** (0.12)	-0.40** (0.13)

Note. The number of teachers = 1010, the number of schools = 81.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 4.*The Association Between Feedback Quality and Instructional Practices (PCT Sample)*

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
Level 1 (Teacher)					
Face-to-face and immediate	-0.01 (0.03)				
Specific regarding improving teaching		0.10* (0.04)			
Useful and relevant regarding improving teaching			0.03 (0.04)		
Specific regarding strength				0.15*** (0.04)	
Useful and relevant regarding strength					0.08* (0.04)
PD	0.05* (0.02)	-0.01 (0.03)	0.02 (0.03)	-0.03 (0.02)	0.001 (0.03)
Pre-measure of teaching practice	0.55*** (0.03)	0.55*** (0.03)	0.55*** (0.03)	0.55*** (0.03)	0.55*** (0.03)
Grade level	0.08*** (0.02)	0.08*** (0.02)	0.08*** (0.02)	0.08*** (0.02)	0.08*** (0.02)
Teaching a tested-subject	0.14*** (0.04)	0.14*** (0.04)	0.14*** (0.04)	0.14*** (0.04)	0.14*** (0.04)
Years of teaching experience	-0.01** (0.002)	-0.01*** (0.002)	-0.01** (0.002)	-0.01*** (0.002)	-0.01*** (0.002)
Level 2 (School)					
% of students eligible for FRL	-0.001 (0.002)	-0.001 (0.002)	-0.001 (0.002)	-0.001 (0.002)	-0.001 (0.002)
Enrollment	-0.00001 (0.0001)	-0.00001 (0.0001)	-0.00001 (0.0001)	-0.00001 (0.0001)	-0.00000 (0.0001)
% of White students	-0.003 (0.002)	-0.003 (0.002)	-0.003 (0.002)	-0.003 (0.002)	-0.003 (0.002)
% of students with IEP	-0.002 (0.01)	0.002 (0.01)	-0.001 (0.01)	-0.001 (0.01)	-0.001 (0.01)
% of students scored at or above proficient level	-0.0002 (0.004)	-0.0001 (0.004)	-0.0003 (0.004)	-0.001 (0.004)	-0.001 (0.004)
Intercept	-0.49** (0.16)	-0.76*** (0.15)	-0.60*** (0.16)	-0.90*** (0.17)	-0.74*** (0.18)

Note. The number of teachers = 1214, the number of schools = 103.

* $p < .05$. ** $p < .01$. *** $p < .001$.

is specific regarding their strength in teaching, the teacher tended to involve more students in learning activities. This association is relatively stronger ($\beta = .12$) than one between teachers' perceptions of feedback being useful and relevant regarding strengths and their CE practice ($\beta = .09$). Similarly, there were significant associations between teachers' perceptions of feedback being specific, useful and relevant regarding strength in teaching and how much students actively involved in solving problems using various strategies and critical thinking during the class. Again, teachers' perceptions of feedback being specific has a stronger relationship with the practice ($\beta = .15$) compared to their perceptions of feedback being useful and relevant regarding strength in teaching ($\beta = .08$). Another important finding is that for the PCT sample, feedback focused on improving teaching also was significantly associated with instructional practice ($\beta = .10$). Contrary to our expectations based on previous studies, face-to-face and immediate feedback and useful and relevant feedback regarding improving teaching were not associated with teachers' instructional practices.

As expected, the pre-measure of teaching effectiveness had a strong association with the outcome measures. Other covariates at the individual teacher level also had significant association with the outcome variables. Teachers who taught higher grades or a subject tested on the state exam, and had more experience, tended to receive higher student ratings of instruction quality. In contrast, none of the school-level covariates were significant in any of the models such as percentage of students who are White, have an IEP, or are eligible for free and reduced lunch price.

Discussion

Teacher evaluation is a potentially powerful tool for improving teaching effectiveness. Therefore, it is logical that evaluation policies have been at the center of educational reform

efforts. However, research suggests that the effects of implementation of teacher evaluation systems have been mixed (Goldhaber, 2015; Stecher et al., 2018; Taylor & Tyler, 2012). There are likely multiple explanations for the uneven effects of teacher evaluation on teaching effectiveness, but our findings suggest that variation in quality of principals' feedback following classroom observations may be one explanation. We found that some characteristics of feedback are associated with instruction quality. Based on our theoretical framework, we focused on the five characteristics of feedback that are likely to shape the recipients' response to feedback: (1) specific regarding strengths, (2) and ways to improve teaching, (3) useful and relevant regarding strengths and (4) ways to improve teaching, and (5) face-to-face and immediate.

We found that specific (i.e., focused, detailed, concrete) feedback that focused on existing strengths of teachers was more strongly associated with instruction quality than specific feedback that focused on ways to improve teaching. In addition, teachers' perceptions of whether feedback was relevant and useful when focused on strengths was more strongly associated with instruction quality than when focused on ways to improve teaching. These findings are consistent with previous studies that support the effects of positive feedback (e.g., Kinicki et al., 2004; Scheeler et al., 2004; Sleiman et al., 2020; Thurlings et al., 2013). That is, teachers were more likely to engage in high-quality instruction after receiving confirmation that they were doing a good job, as it raised their expectancies for success.

Another key finding is that specific feedback regarding ways to improve teaching was significantly associated with higher ratings of use of problem-solving and critical-thinking instructional practices, although there was no association with cognitive engagement instructional practices. That is, the effect of one characteristic of effective feedback varied by teaching practice. While it is logical that specific suggestions from principals on how to improve

in any teaching practice might be effective, it is most likely that it would have an effect on difficult instructional practices. Promotion of critical thinking is a complex, advanced instructional practice that is challenging to implement consistently in typical classrooms (Fox, 1962; Van der Lans et al., 2018). Of course, the promotion of critical thinking is not always appropriate because there are times when students should be practicing and over-learning skills that are foundational to critical thinking, yet this challenging teaching practice is too rare in classrooms (Willingham, 2008).

Teachers' perceptions of whether feedback was useful and relevant regarding ways to improve were not related to quality of instruction in any aspect. A plausible explanation is that even if a teacher believes feedback is useful and relevant, feedback related to their weakness is more challenging to translate into practice. This obstacle to benefiting from negative feedback, compared to positive, has been documented in feedback research in other non-educational settings (Audia & Locke, 2003). Another plausible explanation, aligned with our hypotheses and theoretical framework (Figure 1), is that useful and relevant feedback regarding ways to improve may not have been as effective in terms of answering the questions "where am I going, how am I doing, where do I go next," compared to specific feedback regarding ways to improve. Certainly, we need more research on these nuanced aspects of feedback.

Our findings provide important practical implications for training principals as evaluators. First, teachers' perceptions of feedback quality do predict how well they teach based on aggregated student perceptions. Thus, providing high-quality feedback should be a priority among principals. Second, principals' focus on teachers' strengths is more likely to improve their instruction quality than principals' focus on ways to improve teaching. Thus, principals may want to prioritize feedback on teachers' strengths, but also add specific feedback on ways to

improve teaching when evaluating especially complex, challenging-to-implement instructional practices.

An unexpected finding was that face-to-face and immediate feedback was not associated with higher ratings of either teaching practice. In contrast, previous research suggested these are useful characteristics of feedback (Hunter & Springer, 2022; Kraft & Christian, 2019). One explanation for these conflicting results is that our study used student ratings to measure instruction quality whereas most previous studies used teacher report which may be influenced by source bias. This is important because (as discussed above) teachers who have positive reports of their evaluation may perceive both the quality of feedback and the impact of the evaluation positively, despite objective feedback quality or improvement in instruction quality.

Another explanation is that in the Network for Educator Effectiveness every teacher receives feedback multiple times per year. When the evaluation context is so normative and familiar, having feedback be face-to-face and immediate may be less important. This finding has practical implications because it may be more cost-effective to communicate feedback via written notes and it gives busy principals more flexibility for scheduling feedback sessions. Further research is needed, but our study expands the current discourse around feedback quality by adding a critical source of data – student ratings – about instruction quality.

Limitations

The findings from this study have some limitations. First, we focused on teachers' perceptions of feedback but did not have access to the content of the feedback, which presumably also influences teaching improvement. However, given that the perceptions of the recipients of feedback affect whether they take actions on that feedback, we think this analysis was an important first step.

Second, we cannot fully rule out the possibility of a potential selection bias because the districts in the samples are self-selected based on their simultaneous use of both the TESS and the NEE-TS. These districts may have more resources and/or place a higher priority on teacher growth and student achievement compared to other districts.

A third limitation of the study is that we focused on two instructional practices: CE and PCT. This decision was made based on previous literature that found these teaching practices to affect student learning and based on the fact that many districts focused on these instructional practices. However, it is possible that feedback quality affects other dimensions of instruction differently, such as teacher-student relationships, classroom management, and use of motivational strategies. Examining the effects of feedback on different aspects of instruction quality will deepen our understanding of the teacher evaluation feedback process.

Conclusion

In the current study, we found three characteristics of principals' evaluative feedback are associated with teachers' increased instruction quality, as measured by student ratings, after controlling for previous instruction quality. Focusing on strengths, but not ways to improve teaching, predicted higher quality of one instructional practice – promoting cognitive engagement. However, focusing on both strengths and ways to improve teaching predicted higher quality of another instructional practice – promoting critical thinking. Promotion of critical thinking is an infrequent, complex teaching practice that may benefit more from information on how to improve. Finally, we found that feedback did not need to be immediate and face-to-face to predict improvement for either instructional practice. Both evaluating teachers and training principals to be effective evaluators are resource-intensive activities. Thus, it is important for educational researchers and policy makers to better understand what

constitutes effective evaluative feedback and refine that feedback process so as to maximize benefits for teachers and their students.

Preprint for NEEadvantage.com

Note

1. NEE-CO and TESS are proprietary, which precludes the disclosure of all the items in the observation rubric or surveys.

Data Availability Statement

The datasets analyzed during the current study are not publicly available because the data were collected from students and teachers under the agreement between the Network for Educator Effectiveness (neeadvantage.com) and its member districts. The authors of the study can answer questions about the data or provide further analysis of the data on reasonable request.

Conflicts of Interest

All authors declare that they have no conflicts of interest.

References

- Ammons, R. B. (1956). Effects of knowledge of performance: A survey and tentative theoretical formulation. *The Journal of general psychology*, 54(2), 279-299.
- Alvero, A. M., Bucklin, B. R., & Austin, J. (2001). An objective review of the effectiveness and essential characteristics of performance feedback in organizational settings (1985-1998). *Journal of Organizational Behavior Management*, 21(1), 3-29.
- Audia, P. G., & Locke, E. A. (2003). Benefiting from negative feedback. *Human Resource Management Review*, 13(4), 631-646.
- Bandura, A. (1991). Social theory of self-regulation. *Organizational Behavior and Human Decision Processes*, 50, 248-287.
- Briggs, D., & Domingue, B. (2011). *A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the Los Angeles Times*. National Education Policy Center, University of Colorado.
- Chafouleas, S. M. (2011). Direct behavior rating: A review of the issues and research in its development. *Education & Treatment of Children*, 34(4), 575-591.
- 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). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Central. Retrieved from <http://ies.ed.gov/ncee/edlabs>.
- Cherasaro, T. L., Brodersen, R. M., Yanoski, D. C., Welp, L. C., & Reale, M. L. (2015). The examining evaluator feedback survey. REL 2016-100. Regional Educational Laboratory Central.

- Cohen, J., Loeb, S., Miller, L. C., & Wyckoff, J. H. (2020). Policy implementation, principal agency, and strategic action: Improving teaching effectiveness in New York City middle schools. *Educational Evaluation and Policy Analysis*, 42(1), 134–160.
- Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, 27(4), 724-750.
- Curtis, R., & Wiener, R. (2012). *Means to an end: A guide to developing teacher evaluation systems that support growth and development*. Aspen Institute.
- Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Association for Supervision and Curriculum Development.
- Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice*. Association for Supervision and Curriculum Development.
- Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, 93(6), 8-15.
- Darling-Hammond, L., Bae, S., Cook-Harvey, C. M., Lam, L., Mercer, C., Podolsky, A., & Stosich, E. L. (2016). *Pathways to new accountability through the Every Student Succeeds Act*. Learning Policy Institute.
- Delvaux, E., Vanhoof, J., Tuytens, M., Vekeman, E., Devos, G., & Van Petegem, P. (2013). How may teacher evaluations have an impact on professional development? A multilevel analysis. *Teaching and Teacher Education*, 36, 1–11.
- Doherty, K. M., & Jacobs, S. (2013). State of the States 2013: Connect the Dots--Using Evaluations of Teacher Effectiveness to Inform Policy and Practice. *National Council on*

Teacher Quality.

- Donaldson, M. (2021). *Multidisciplinary Perspectives on Teacher Evaluation: Understanding the Research and Theory*. Routledge.
- Donaldson, M. L., & Papay, J. P. (2015). Teacher evaluation for accountability and development. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of Research in Education, Finance and, Policy* (2nd ed., pp. 174–193). Routledge.
- Downer, J. T., Stuhlman, M., Schweig, J., Martínez, J. F., & Ruzek, E. (2015). Measuring effective teacher-student interactions from a student perspective: A multi-level analysis. *The Journal of Early Adolescence*, 35(5-6), 722-758.
- Duff, M., & Wohlstetter, P. (2019). Negotiating intergovernmental relations under ESSA. *Educational Researcher*, 48(5), 296-308.
- Eccles (Parsons), J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., et al. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.). *Achievement and achievement motivation* (pp. 75–146). San Francisco, CA: W. H. Freeman.
- Egalite, A. J., Fusarelli, L. D., & Fusarelli, B. C. (2017). Will decentralization affect educational inequity? The Every Student Succeeds Act. *Educational Administration Quarterly*, 53(5), 757-781.
- Esarey, J., & Menger, A. (2019). Practical and effective approaches to dealing with clustered data. *Political Science Research and Methods*, 7(3), 541-559.
- 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.

- Feldlaufer, H., Midgley, C., & Eccles, J. S. (1988). Student, teacher, and observer perceptions of the classroom environment before and after the transition to junior high school. *The Journal of Early Adolescence*, 8(2), 133-156.
- Fong, C. J., Kim, Y., Davis, C. W., Hoang, T., & Kim, Y. W. (2017). A meta-analysis on critical thinking and community college student achievement. *Thinking Skills and Creativity*, 26, 71-83.
- Fox, R. B. (1962). Difficulties in developing skill in critical thinking. *The Journal of Educational Research*, 55(7), 335-337.
- Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, 29(2), 147-194.
- Fraser, B. J., & O'Brien, P. (1985). Student and teacher perceptions of the environment of elementary school classrooms. *The Elementary School Journal*, 85(5), 567-580.
- Fredricks, J. A., Blumenfeld, P. C., & Alison, H. P. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1), 59-109.
- Garet, M. S., Wayne, A. J., Brown, S., Rickles, J., Song, M., & Manzeske, D. (2017). The impact of providing performance feedback to teachers and principals. NCEE 2018-4001. *National Center for Education Evaluation and Regional Assistance*.
- Giancarlo, C. A., Blohm, S. W., & Urdan, T. (2004). Assessing secondary students' disposition toward critical thinking: Development of the California measure of mental motivation. *Educational and Psychological Measurement*, 64(2), 347-364.
- Glickman, C. D. (2002). *Leadership for Learning: How to Help Teachers Succeed*. ASCD.
- Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher*, 44(2), 87-95.

- Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability*, 26(1), 1-24.
- Halverson, R., Kelley, C., & Kimball, S. (2004). Implementing teacher evaluation systems: How principals make sense of complex artifacts to shape local instructional practice. *Educational administration, policy, and reform: Research and measurement*, 153-188.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112.
- Hill, H., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, 83(2), 371-384.
- Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, 48(3), 794–831.
- Hunter, S. B., & Springer, M. G. (2022). Critical Feedback Characteristics, Teacher Human Capital, and Early-Career Teacher Performance: A Mixed-Methods Analysis. *Educational Evaluation and Policy Analysis*.
- 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.
- Jones, E., & Bergin, C. (2019). Evaluating teacher effectiveness using classroom observations: A Rasch analysis of the rater effects of principals. *Educational Assessment*, 24(2), 91–118.
[https:// doi. org/ 10. 1080/ 10627 197. 2018. 15642 72](https://doi.org/10.1080/10627197.2018.1564272)

- Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Research Paper. Measures of Effective Teaching Project. Bill & Melinda Gates Foundation.
- Kim, J., Sun, M., & Youngs, P. (2019). Developing the “will”: The relationship between teachers’ perceived policy legitimacy and instructional improvement. *Teachers College Record*, 121(3), 1–44.
- 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.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254.
- Kraft, M.A., Blazar, D., Hogan, D. (2018). The effect of teaching coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, 88(4), 547-588.
- Kraft, M.A. & Christian, A. (2019). In search of high-quality evaluation feedback: An administrator training field experiment (EdWorkingPaper No.19-62). Retrieved from Annenberg Institute at Brown University: <http://edworkingpapers.com/ai19-62>
- Lavigne, A. L., & Good, T. L. (2015). *Improving teaching through observation and feedback: Beyond state and federal mandates*. Routledge.
- Lechermeier, J., & Fassnacht, M. (2018). How do performance feedback characteristics influence recipients’ reactions? A state-of-the-art review on feedback source, timing, and valence effects. *Management Review Quarterly*, 68(2), 145-193.

- Locke, E. A. (1968). Toward a theory of task motivation and incentives. *Organizational behavior and human performance*, 3(2), 157-189.
- Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance*. Englewood Cliffs, NJ: Prentice Hall.
- Li, Q., & Baker, R. (2018). The different relationships between engagement and outcomes across partici-pant subgroups in Massive Open Online Courses. *Computers & Education*, 127, 41–65.
- Li, X. (2022). *Authentic teacher evaluation scores and student achievement: Comparing principals' and students' ratings*. AEFPP 47th Annual Conference, Denver, Colorado.
- Li, X., Bergin, C., & Olsen, A. A. (2022). Positive teacher-student relationships may lead to better teaching. *Learning and Instruction*, 80, 101581.
- Little, O., Goe, L., & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness*. National Comprehensive Center for Teacher Quality.
- Marsh, H. W., Dicke, T., & Pfeiffer, M. (2019). A tale of two quests: The (almost) non-overlapping research literatures on students' evaluations of secondary-school and university teachers. *Contemporary Educational Psychology*, 58, 1-18.
- 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.
- Marzano, R. J., Pickering, D., & Pollock, J. E. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement*. Association for Supervision and Curriculum Development.

- Marzano, R. J., & Toth, M. D. (2013). *Teacher evaluation that makes a difference: A new model for teacher growth and student achievement*. ASCD.
- McCormick, N. J., Clark, L. M., & Raines, J. M. (2015). Engaging students in critical thinking and problem solving: A brief review of the literature. *Journal of Studies in Education*, 5(4), 100-113.
- Metallidou, P., & Vlachou, A. (2007). Motivational beliefs, cognitive engagement, and achievement in language and mathematics in elementary school children. *International Journal of Psychology*, 42(1), 2-15.
- 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.
- Pietarinen, J., Soini, T., & Pyhältö, K. (2014). Students' emotional and cognitive engagement as the determinants of well-being and achievement in school. *International Journal of Educational Research*, 67, 40-51.
- Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education*, 121(2), 183-212.
- Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104(1), 1-15.
- Reid, D., Galey-Horn, S., & Kim, J. (2020). How states use EESA to support principal preparation, development, and quality. In Youngs, P., Kim, J., & Mavrogatoro, M. (Eds.), *Exploring principal development and teacher outcomes: How principals can strengthen instruction, teacher retention, and student achievement* (pp. 207–220). Routledge.

- Reinhorn, S. K., Johnson, S. M., & Simon, N. S. (2017). Investing in development: Six high-performing, high-poverty schools implement the Massachusetts Teacher Evaluation Policy. *Educational Evaluation and Policy Analysis*, 39(3), 383-406.
- Reston, V. A. (2015). *National Policy Board for Educational Administration*. Professional Standards for Educational Leaders, American Association of Colleges of Teacher Education.
- Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology*, 104(3), 700-712.
- Roberson, Q. M., & Stewart, M. M. (2006). Understanding the motivational effects of procedural and informational justice in feedback processes. *British Journal of Psychology*, 97(3), 281-298.
- Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125(1), 175-214.
- Scheeler, M. C., Ruhl, K. L., & McAfee, J. K. (2004). Providing performance feedback to teachers: A review. *Teacher Education and Special Education*, 27(4), 396-407.
- Shaked, H. (2018). Why principals often give overly high ratings on teacher evaluations. *Studies in Educational Evaluation*, 59, 150-157.
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153-189.
- Sleiman, A. A., Sigurjonsdottir, S., Elnes, A., Gage, N. A., & Gravina, N. E. (2020). A quantitative review of performance feedback in organizational settings (1998-2018). *Journal of Organizational Behavior Management*, 40(3-4), 303-332.

- Stecher, B. M., Holtzman, D. J., Garet, M. S., Hamilton, L. S., Engberg, J., Steiner, E. D., Robyn, A., Baird, M. D., Gutierrez, I. A., Peet, E. D., Brodziak de los Reyes, I., Fronberg, K., Weinberger, G., Hunter, G. P., & Chambers, J. (2018). *Improving teaching effectiveness: Final report: The intensive partnerships for effective teaching through 2015–2016*. RAND Corporation.
- 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*, 27(4), 793-818.
- Tsai, C. L., Bergin, C., & Jones, E. (2022). Students in 4th to 12th grade can distinguish dimensions of teaching when evaluating their teachers: A multilevel analysis of the TESS survey. *Educational Studies*, 1–16
- Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, 102(7), 3628–3651.
- Thorndike, E.L. (1913). *Educational Psychology. Vol. 1. The Original Nature of Man*. Columbia Univ.: New York.
- Thurlings, M. C. G., Kreijns, K., Vermeulen, M., Bastiaens, T. J., & Stijnen, P. J. J. (2012). Development of the teacher feedback observation scheme: Evaluating the quality of feedback in peer groups. *Journal of Education for Teaching*, 38(2), 193-208.
- Thurlings, M., Vermeulen, M., Bastiaens, T., & Stijnen, S. (2013). Understanding feedback: A learning theory perspective. *Educational Research Review*, 9, 1-15.
- Toch, T., & Rothman, R. (2008). Rush to Judgment: Teacher Evaluation in Public Education. Education Sector Reports. *Education Sector*.

- Tuma, A. P., Hamilton, L. S., & Tsai, T. (2018). *A Nationwide Look at Teacher Perceptions of Feedback and Evaluation Systems*. Rand Cooperation. Retrieved from https://www.rand.org/content/dam/rand/pubs/research_reports/RR2500/RR2558/RAND_RR2558.pdf
- Van der Lans, R. M., Van de Grift, W. J., & Van Veen, K. (2018). Developing an instrument for teacher feedback: using the rasch model to explore teachers' development of effective teaching strategies and behaviors. *The Journal of Experimental Education*, 86(2), 247-264.
- Von Secker, C. E., & Lissitz, R. W. (1999). Estimating the impact of instructional practices on student achievement in science. *Journal of Research in Science Teaching*, 36(10), 1110-1126.
- Wang, M.-T., & Degol, J. (2014). Staying engaged: Knowledge and research needs in student engagement. *Child Development Perspectives*, 8(3), 137-143.
- Wang, Z., Bergin, C., & Bergin, D. A. (2014). Measuring engagement in fourth to twelfth grade class-rooms: The Classroom Engagement Inventory. *School Psychology Quarterly*, 29(4), 517
- Wayne, A. J., Garet, M. S., Brown, S., Rickles, J., Song, M., Manzeske, D., & Ali, M. (2016). Early implementation findings from a study of teacher and principal performance measurement and feedback year 1 report. Technical report, American Institutes of Research.
- Willingham, D. T. (2008). Critical thinking: Why is it so hard to teach? *Arts Education Policy Review*, 109(4), 21-32.

- Wirkala, C., & Kuhn, D. (2011). Problem-based learning in K–12 education: Is it effective and how does it achieve its effects? *American Educational Research Journal*, 48(5), 1157-1186.
- Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, 67(3), 189-205.
- Yu, W. F., She, H. C., & Lee, Y. M. (2010). The effects of web-based/non-web-based problem-solving instruction and high/low achievement on students' problem-solving ability and biology achievement. *Innovations in Education and Teaching International*, 47(2), 187-199.

Preprint for NEEadvantage.com