Research Report

Certified to Evaluate: Exploring Administrator Accuracy and Beliefs in Teacher Observation

ETS RR–21-05

Nathan Jones
Courtney Bell
Yi Qi
Jennifer Lewis
David Kirui
Leslie Stickler
Amanda Redash

December 2021

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Certified to Evaluate: Exploring Administrator Accuracy and Beliefs in Teacher Observation

Nathan Jones[1], Courtney Bell[2], Yi Qi[2], Jennifer Lewis[3], David Kirui[2], Leslie Stickler[2], & Amanda Redash[1]

1 Wheelock College of Education & Human Development, Boston University, Boston, MA
2 ETS, Princeton, NJ
3 College of Education, Wayne State University, Detroit, MI

The observation systems being used in all 50 states require administrators to learn to accurately and reliably score their teachers' instruction using standardized observation systems. Although the literature on observation systems is growing, relatively few studies have examined the outcomes of trainings focused on developing administrators' accuracy using observation systems and the administrators' perceptions of that training. Therefore, the focus of this study is on examining administrators' efforts to become accurate and reliable within the context of a comprehensive teacher evaluation reform. This study was conducted during the year-long training and implementation of a new observation system in the context of a large urban district's teacher evaluation reform. The study brings together data on the outcomes of the district training — results on a certification exercise from all administrators in the district — with two sources of data on administrators' perceptions and beliefs. Specifically, we collected fall and spring survey data from nearly 300 administrators and longitudinal interview data from a subsample of 24 administrators. Taken together, these data allowed us to investigate administrators' responses to training and low-stakes practice using the observation process over 1 year. At the end of initial training, administrators reported high levels of learning, particularly in domains aligned with the focus of training. Over the year, administrators reported increased facility with the routines of conducting observations, but they still expressed learning needs, many related to the content of the observation framework. However, results from the training certification test suggested lower than desired levels of accuracy and reliability; administrators regularly did not agree with each other or with master raters. The certification test results suggested that even with a significant investment in administrator learning, there was more to be learned and mastered. If we hope for teacher evaluation to lead to the types of changes in teaching and learning that reformers have envisioned, policymakers and practitioners alike will need to devote time and resources to supporting administrator learning in initial training and throughout administrator use in practice.

Keywords Teacher evaluation; teaching quality; teacher quality; observation; certification; administrator beliefs; accuracy

doi:10.1002/ets2.12316

Through a confluence of federal, state, and private investments, the last 10 years have seen teacher evaluation reform in almost every state in the country. Observation systems have been at the center of most states' evaluation reforms (Donaldson & Papay, 2014). In most cases, the new systems have emphasized the use of multiple measures of teacher quality, including scores based on student growth and scores drawn from observations. States have largely focused on using these data to meet two distinct goals: (a) helping teachers improve and (b) making human capital decisions (Donaldson & Papay, 2014). Because student growth measures are limited in providing meaningful feedback to teachers or administrators (Corcoran, 2016; Harris, 2011), observations appear to be the measure best suited to facilitating teacher development. And, there is evidence that observation systems, more than student growth measures, are being used to make teacher accountability decisions (Goldring et al., 2015).

In order for classroom observations to meet either of these two goals, they will need to be substantiated with evidence on their reliability and validity. Observations rely on human judgment, and the fairness of the evaluation process depends on whether administrators can create scores that are free from error or systematic bias. However, existing research suggests several factors that might impact observation score quality. First, observation scores are prone to bias and can be sensitive to various contextual factors (Garrett & Steinberg, 2015; Gill et al., 2016; Steinberg & Garrett, 2016; Whitehurst et al., 2014; Wind et al., 2019). Specifically, teachers of students who exhibit higher levels of academic performance are also more likely to score higher on observation systems (Steinberg & Garrett, 2016; Whitehurst et al., 2014). And, teachers of students in lower grades tend to score higher than those in higher grades (Mihaly & McCaffrey, 2014). There is

Corresponding author: C. Bell, E-mail: courtney.bell@wisc.edu

also evidence that raters are not consistent in how they use rating scales over time (Casabianca et al., 2013; Casabianca et al., 2015). Finally, studies by Polikoff (2015) and Garrett and Steinberg (2015) suggest that some behaviors (such as positive climate and classroom procedures) are more stable over time than others, namely domains related to instructional quality, such as a teacher's questions and discussion techniques (Polikoff, 2015).

Although informative, the existing research on Framework for Teaching (FFT) and other observation systems has largely ignored the question of how local administrators—including principals, assistant principals, and other district staff commonly tasked with carrying out observations—will rate teachers in practice. Much of the existing research on the accuracy, reliability and validity of classroom observations has been conducted in the context of research studies, where raters were hired by the research team, often had expertise in the subject they were observing, and had no preexisting relationships with the individuals they were observing. We should not presume that the validity of observation systems will carry from research to applied contexts, as documented by Liu et al. (2019). This skepticism is especially relevant when considering the unique circumstances of principals and other building administrators evaluating their teacher colleagues (for details about these circumstances see Kraft & Gilmour, 2016). That said, a small number of studies have begun to examine the quality of scores produced by administrators in teacher evaluation systems. The studies conducted by Garrett and Steinberg (Garrett & Steinberg, 2015; Steinberg & Garrett, 2016) focused on local administrators' scoring patterns across Chicago Public Schools. Similarly, the Wind et al. (2019) study examined scoring patterns from over 1,300 principals in Missouri, and the authors found evidence of differential scoring patterns based on teachers' gender, years of experience, and school level. There is also research that links administrators' ratings to other validity outcomes (e.g., Dee & Wyckoff, 2015; Ronfeldt & Shanyce, 2016), but that work uses observation scores to predict outcomes rather than an investigation of the scores themselves. In addition, a growing body of research examines how principals approach the work of conducting observations, manage the work of observations vis-à-vis other responsibilities, and use observation scores in making evaluation decisions (Donaldson & Mavrogordato, 2018; Donaldson & Woulfin, 2018; Goldring et al., 2015; Grissom & Bartanen, 2019; Harvey et al., 2019; Lochmiller & Mancinelli, 2019).

Despite these important contributions, existing research on classroom observations generally falls into one of two categories: validity evidence in research studies using trained raters or studies focused on the implementation of observation systems by administrators and the related consequences of how scores are used. Generally missing from the literature is evidence about how administrators perceive the experience of learning to rate and the outcomes of that learning. Therefore, the focus of this study is on examining the outcomes of observation system training and the administrators perceptions of that training within the context of a large urban district's comprehensive teacher evaluation reform. We collected multiple sources of data to inform this question. We examined district-wide data on administrators' performance on a formal certification exercise completed following formal training on their observation system. These performance data were coupled with data on administrators' perceptions taken from two sources: longitudinal survey data from nearly 300 administrators across the district and longitudinal interview data from a subsample of 24 administrators. To unpack the large-scale survey and administrative data, we followed this subsample from initial training through a practice year during which the administrators used the observation system with teacher volunteers.

New Observation Systems

Observations have a long history of use by schools to assess teachers. As early as the 1920s, schools were using observations to distinguish between teachers with varying personal traits (e.g., adaptability, neatness), although there was little consistency in what qualities of teachers the schools measured (e.g., Barr, 1931, 1946, 1958; Kennedy, 2010; Rowan & Raudenbush, 2016). This idiosyncratic approach—with the decision of what to measure and how left up to local administrators—continued well into the 1980s (Darling-Hammond et al., 1983) when researchers introduced efforts to shift administrators' attention to specific skills thought to be necessary for good teaching. Resulting from "process–product" research, these observation protocols used checklists to focus administrators' attention on discrete teacher behaviors (Brophy & Good, 1986; DeMoulin, 1988). Contemporary approaches to observation have generally moved away from such checklists to embrace nuanced, multidimensional perspectives on teaching quality. This shift has resulted in evaluation processes that are more detailed and require more time for administrators and teachers, and it is now common practice for evaluations to incorporate pre- and postobservation conferences to facilitate teacher reflection (Danielson, 2002; Kersten & Israel, 2005). Further, what makes the current set of reforms notable is that many districts and states are converging around similar observation tools, such as the Marzano Focused Teacher Evaluation Model, the

Classroom Assessment Scoring System (CLASS), and the FFT. The FFT, for example, by 2013 had been formally adopted or approved for use in at least 31 states and half of the nation's 20 largest districts (The Danielson Group, 2021). Thus, although the specific evaluation practices used across districts are not identical, they share similar characteristics and require similar sets of knowledge and skills from administrators. We next review the skills administrators need to carry out the work of rating and how training might be structured to support the development of such skills.

Developing Administrators' Observation Skills Through Training and Certification

Why attend to the development of administrators' observation skills? There is a long history of reforms that have failed, in part, because they did not attend to the learning of the actors involved (Cohen & Hill, 2001). In the context of learning to conduct observations, the first of these learning needs — and arguably the need on which the others depend — is to score accurately and reliably. By scoring accurately, we mean to score in ways that match an expert (a "master" rater) on the protocol, and to score reliably is to assign scores in a way that is consistent across time. It is not helpful if administrators are unable to accurately and reliably identify which teachers need further support and which specific skills an individual teacher should develop further.

Developing administrators' skills for creating accurate observation scores relies on districts and states developing and implementing high-quality training opportunities. Training is designed to ensure that administrators are developing a shared understanding of what the words in the observation system mean (Grossman & McDonald, 2008). However, there appears to be little consistency in states' and districts' current training practices (Herlihy et al., 2014; Leahy, 2012; McGuinn, 2012). Herlihy et al.' (2014) review suggested that states and districts take on a variety of training practices aimed primarily at holding administrators accountable for their scores (e.g., requiring certification tests and recalibration, auditing scores, and the use of multiple raters per teacher). The field has not adopted common certification standards to ensure that observers' scores are adequately consistent and accurate (White, 2018) or, in other words, that observers hold shared understandings of what they are watching and how to assess it. We know little about whether and how such training can increase administrators' abilities to provide accurate scores, including whether some components of an observation system are easier to train observers on than others. Neither do we know whether training can influence administrators' perceptions about the role of accuracy in creating observation scores. Thus, to better understand how administrators learn to conduct observations in practice, it is important to investigate outcomes of training, including how accurate and reliable administrators are and how administrators perceive the process of learning across time.

Developing Accurate and Reliable Observation Protocol Use

If successful implementation of observation systems depends on building capacity among the administrators who will be carrying out the work, what does it mean to build administrators' capacity? We propose that research needs to attend to at least three issues related to administrators' use of observation tools: (a) their beliefs about the observation tool and what it measures, (b) their perceptions of their observation skills and which aspects of using the tool are difficult and easy, and (c) their competence in using the observation tool.

With regard to beliefs, scholarship has demonstrated that what one believes shapes what one comes to understand (e.g., Schommer, 1990). If one administrator believes that an observation protocol captures good teaching, that administrator is going to engage one type of learning trajectory. This trajectory is likely to be different from the trajectory engaged by an administer who believes the observation protocol does not capture good teaching. People's beliefs (and perceptions) shape what they understand and how they act. This understanding has been demonstrated in a range of studies of administrators. From administrators whose support of beginning teachers varied with their beliefs (Youngs, 2007) to those who supported writing instruction differently depending on their beliefs (McGhee & Lew, 2007), beliefs matter to both understanding and actions. We anticipate administrators' beliefs will influence how they use the observation protocol.

In addition, we expect that administrators' perceptions of their learning needs — specifically which aspects of using FFT in practice are hard or easy — will impact where they dedicate their efforts and how they engage in the work of conducting observations. Painter (2000) and Reddy et al. (2018) have each explored administrators' perceptions of the value they place on evaluating teachers and their perceived ability at conducting observations, and we come to similar findings. We find that administrators recognize the value in conducting evaluations (and identifying low-performing teachers) and are confident in their own abilities to observe teachers. As Painter posits, if administrators value evaluating teacher

performance, they will be more motivated to engage in this work. However, other studies, including those conducted by Donaldson and Donaldson (2012) and Wieczorek et al. (2018) raise questions about the extent to which administrators perceive evaluation to be worth their efforts, as they navigate the challenges of managing evaluation tasks with their other responsibilities. A final consideration is administrators' actual proficiency at creating accurate observation scores. Existing studies have not examined administrators' accuracy at the conclusion of training. However, we anticipate that when they receive signals about their performance conducting observations, this will shape administrators' beliefs and perceptions about their ability to do this work and consequently the effort they exert toward their observations.

Consequently, in this study, we focus on two outcomes that shape the quality of administrators' use of observation instruments in practice, namely their accuracy and reliability upon initial training (measured as their ability to pass a certification test at the end of their training) as well as their ongoing beliefs and perceptions of the FFT and their training experiences across the school year. Specifically, we ask the following research questions:

1. Upon completing their training, to what extent are administrators able to rate accurately and reliably?
2. Throughout training and a practice year, how do administrators perceive their own capabilities related to scoring accurately and reliably?
3. What are administrators' perceptions of the kinds of training that would be necessary to score accurately and reliably?

Method

In order to situate administrators' outcomes, beliefs, and perceptions, we first provide a description of the district observation context and then describe the data sources and analytic methods used.

Setting and Participants

Our sample includes administrators and other district personnel being trained to conduct classroom observations in Los Angeles Unified School District's (LAUSD) implementation of a new evaluation system, the Teacher Growth and Development Cycle (TGDC). We collected certification data on all 1,000 administrators in the sample. These were coupled with survey data from 293 administrators and longitudinal interview data collected from 24 focus administrators. Of the administrators trained in the 2012–2013 school year for whom we have survey data, the vast majority (88% altogether) were principals (64%) or assistant principals (24%), and about 5% were central office administrators. On average, administrators had 6 years of administrative experience. They were generally certified as teachers, with most certified in elementary teaching and a small number certified in mathematics or science.

To more deeply understand administrators' beliefs and perceptions as they learned to use the observation system, we gathered data from a subsample of administrators ($N = 24$) during their initial training and during a practice year in which they used the instrument with a teacher volunteer with no consequences attached. From the pool of administrators trained before the 2012–2013 school year, we selected a stratified sample of focus administrators based on grade level (i.e., elementary vs. middle vs. high school), subject area, and role (i.e., principal vs. assistant principal vs. central office administrator), mirroring the population of administrators in the district. Focus administrators were similar to the larger population of observers who returned surveys; however, they were slightly less experienced, were slightly more likely to be certified in English/language arts and science, and were less likely to be certified in elementary education. Over half of the focus administrators were principals, with another quarter serving as assistant principals, four serving as instructional directors, and three serving as instructional coaches or instructional specialists. They had an average of 4.4 years in their current professional roles, including time in LAUSD and any other district. Initially, our posttraining subsample consisted of 42 focus observers. Of these, 24 completed the end-of-year survey. The administrators who did not complete spring interviews did not differ on observable variables than those who did; however, we discuss in the Limitations section the possible consequences of working with a self-selected sample.

We collected data at multiple time points throughout the implementation of the TGDC. First, all administrators were required to complete a 4-day training during the summer before the 2012–2013 school year. At the conclusion of this initial training, all administrators were required to participate in a certification activity. Then, all administrators were given a practice year to use the observation system in a no-stakes context with a single teacher volunteer. We collected

evidence of their accuracy and reliability upon conclusion of the certification exercise. We then recorded administrators' perceptions of their training experiences at both the end of training and the end of the practice school year, examining survey data across a large sample of administrators and richer interview data with a subsample of administrators.

Training and Certification

Teaching and Learning Framework Observation Tool

LAUSD uses a modified version of Danielson's FFT (Danielson, 1996, 2007) called the Teaching and Learning Framework (TLF). The FFT is a widely used instrument originally developed in the 1990s to provide guidance for improving teacher instruction. Predictive validity of the FFT has been established in a handful of existing studies, which have shown that FFT scores correlate with student achievement gains (e.g., Gallagher, 2004; Holtzapple, 2003; Kimball et al., 2004; Milanowski, 2004). For example, in 2003, Holtzapple found positive and significant correlations between FFT composite scores and student gains on state assessments, though correlations varied depending on the subject taught (e.g., .27 for reading and .38 for math). In the Measures of Effective Teaching Project (MET), researchers established significant but somewhat smaller relationships between the FFT and students' value added scores on math (.18) and English language arts (.11). The MET study also provided data on the reliability of FFT scores, finding that it took approximately four observations of a teacher to obtain a more stable estimate of his or her performance. While FFT does not have the same level of evidence of its measurement properties as a tool like the CLASS, it has wide acceptance among educators as well as take-up in systems of evaluation.

LAUSD's derivative of the FFT — the TLF — was aligned to the California teaching standards and vetted with and modified by LAUSD stakeholders. There are five domains, or standards, represented in the TLF:

- Standard 1: Planning and Preparation
- Standard 2: Classroom Environment
- Standard 3: Delivery of Instruction
- Standard 4: Additional Professional Responsibilities
- Standard 5: Professional Growth

Of these, all but Standard 4 (which was left to the administrators' judgment) were included in the observation cycle process. Out of the FFT's 61 total elements, the district selected 21 focus elements upon which teachers would be evaluated during 2012 – 2013.

The administrator training and certification lasted 5 days in total, with 4 days of instruction and 1 day devoted to the certification test. Training was collaboratively led by a vendor, Teaching and Learning Solutions (http://www.teachinglearningsolutions.com/) and LAUSD staff. Analyses of training documentation and interviews with district personnel suggested the formal training objectives focused on two goals: supporting a shared understanding of teaching and learning and preparing administrators to conduct reliable, accurate observations. To support these goals, the trainers provided administrators with a detailed explanation and discussion about the focus elements of TLF. Trainers helped administrators to record accurate and appropriate notes, and they also engaged administrators in conversations about how to accurately score teaching practice using the TLF. A half day of training was devoted to pre- and postobservation conversations with teachers. By the end of training, administrators were expected to pass a certification test. The certification test served a gatekeeping function for the proper use of the TLF protocol. Administrators needed to demonstrate that they could accurately and reliably score teaching practice at acceptable levels when judged against master raters' scores.

Practice Year

During the practice year, the district required administrators to conduct observations with one teacher volunteer. There were no stakes attached to the observation scores. Administrators were asked to complete two formal observation cycles, including preobservation and postobservation conferences, and to enter observation notes (and corresponding ratings) into LAUSD's online platform. Administrators were also asked to complete informal observations with teachers, to the extent that this was feasible.

In addition to giving administrators time to practice conducting observations with no stakes attached, the practice year also gave administrators additional supports. The main set of supports available to administrators was through district personnel known as teaching and learning coordinators, who worked directly with administrators and set up regular support sessions and activities. Examples of support sessions included focusing on how to distinguish between different points on the scoring scale for a given element (e.g., how to identify differences between "developing" and "effective" on the element "monitoring student learning") or guidelines for providing feedback to teachers based on observation data. In addition, administrators received supports from the central office detailing process requirements and technical aspects of the implementation of the new evaluation system.

Data

We drew on three sources of data: certification test data, surveys, and interviews. The first source, administrator certification test data, serves as a direct outcome of training and provides information on administrators' accuracy and reliability. We then measure teachers' perceptions of their training and of their scoring abilities through longitudinal surveys. These are coupled with more fine-grained interview data on administrator perceptions collected from a subset of administrators.

Certification Test Data

Certification exercises are a common outcome used to assess observer training. In our study, all trained administrators received overall certification scores of certified with distinction (CD), certified (C), preliminarily certified (PC), or not yet certified (NYC). Four individual certification measures comprised the overall certification score — alignment, objectivity, representation, and accuracy. The district defined these four certification measures as follows:

- Objectivity: The observer records evidence that is free of "bias, opinion, and subjectivity."
- Alignment: The observer correctly aligns evidence to the TLF criteria in ways that reflect the context of the evidence.
- Representation: The observer records a preponderance of evidence for scoring criteria and accurately represents the classroom and artifact data.
- Accuracy: The observer assigns numerical scores similar to the scores master observers assign.

At the conclusion of training, 89% of administrators were rated at least at the level of preliminarily certified and were allowed to conduct observations. The remaining 11 percent were only allowed to observe with a certified administrator. Whereas the majority of administrators (68%) were trained during the summer before the 2012–2013 school year, others received training during the school year. We reported survey results for the full sample of trained administrators because analyses examining differences in certification rates by time of training yielded no differences.

Surveys

During the 2012–2013 school year, we administered surveys to administrators at the end of the training and one at the end of the school year. The posttraining survey had eight questions and focused on assessing administrators' learning, confidence, and satisfaction in the training. We also administered an end-of-year survey to capture self-reported administrator understanding of the TGDC process, experiences conducting observations during the 2012–2013 school year, impressions of the TLF, and their beliefs about the feasibility of integrating this work with their other professional responsibilities. The posttraining survey had a response rate of 65.5% and the end-of-year survey had a response rate of 33.6%. To allow us to track the same set of administrators over the course of the year, we limited our focus to the 293 administrators who completed surveys at both time points. Analyses conducted with the full sample did not differ significantly for any of the descriptive findings presented here. Altough the survey nonresponse is a potential limitation of the study, it is worth noting that it is not low relative to other nonmandatory survey-based studies published in the administrator literature (e.g., Koedel et al., 2017; Rockoff et al., 2012).

Interviews

Paralleling the broader survey efforts, we conducted interviews with 24 focus administrators, once after initial training and once at the end of the practice year. The focus administrators participated in interviews over the course of the study,

one after completing their initial training and the other at the end of the practice year. The interviews were meant to provide more substantive details to the data collected in the certification exercise and surveys.

Training interviews focused on administrators' beliefs of good teaching, their perceptions of their ability to score accurately and reliability, and their perceptions about how well their training prepared them for conducting observations. End-of-year interviews — which typically lasted 45 – 60 minutes — again asked about administrators' perceptions of their ability to score but also focused on their experiences and unmet learning needs, relationships with volunteer teachers and how TGDC affected these relationships, and perceptions of the evaluation policy. Interviewers took verbatim notes during all interviews; these notes were checked against audio recordings for accuracy. In order to understand what aspects of training administrators identified as necessary and useful, we asked in the first interview, "What challenges do you expect to face in conducting the observations that are a part of TGDC?" During our end-of-year interview, we asked focus administrators to reflect on their learning experiences during the practice year: "What parts of the TGDC have been the easiest for you to learn? Why? Which have been the hardest?"

Analyses

As noted above, we presented descriptive data from a variety of sources. We looked first at the outcomes of training, as measured in administrators' certification results, and then we tracked on administrators' beliefs and perceptions around the instrument and their preparedness as a result of their training. In analyzing the certification data, we calculated accuracy by examining the proportion of administrators in the sample who "matched" a master rater's score. Given that the TLF observation instrument only includes four scoring points, we only counted exact matches. Match percentages were calculated for each focus element on the TLF and at the aggregated domain level. Because the certification exercise was a one-time test, we cannot assess within-person reliability. Instead, we also calculated the percent exact match between trained administrators, which allowed us to see how much, after training, administrators agreed with each other.

After presenting these certification results, we provided a descriptive account of administrators' perceptions regarding their competence at scoring accurately and of the contributions of the training to their ability to do the work of observing their teachers. Survey responses were presented descriptively, as we reported the distribution of sample responses on questions pertaining to their competence and perceptions of additional learning needs.

The interviews with focus administrators were analyzed to further understand survey findings. Qualitative analysis of all interview responses were completed using a thematic approach, where we identified and analyzed patterns and clusters of responses (Braun & Clarke, 2006; Ritchie et al., 2013). Themes were derived inductively through multiple analyses of the entire data set; as patterns of meaning emerged in the data, they were organized into higher-level themes. The overall thematic framework was revised and reapplied to the data multiple times. A subset of responses (20%) were double coded to assess interrater reliability, and the two coders agreed on approximately 93% of coding decisions.

The list of final codes is as follows:

- Using the observation instrument,
- Time, resources, and logistics
- Technology
- Coaching and feedback conversations
- Managing personal relationships.

Given the emphasis on challenges and learning needs related to accurate use of the TLF, a more fine-grained set of codes within the *Using the observation instrument* code was developed:

- A1. Recording evidence
- A2. Aligning evidence
- A3. Scoring
- A4. Practice and training

Illustrative quotes were chosen to convey the essence of these themes.

Table 1 Administrator Certification Rates on the Teaching and Learning Framework Observation Tool (Overall and by Component)

$N = 1{,}009$	Overall (%)	Objectivity (%)	Alignment (%)	Representation (%)	Accuracy (%)
Certified with distinction (DC)	.00	<.01	<.01	.00	.05
Certified (C)	.11	.77	.44	.29	.29
Preliminarily certified (PC)	.82	.18	.44	.56	.35
Not yet certified (NYC)	.07	.05	.12	.14	.32

Results

Outcomes of Administrator Training

We assessed the outcomes of the administrator training by looking at how accurately administrators in LAUSD matched master raters and each other. In the following tables, we provide an overview of how administrators fared on the certification standards adopted by the district. Table 1 illustrates overall certification rates and certification rates by component (i.e., objectivity, alignment, representation, and accuracy). Overall, the majority of administrators (82%) scored PC; administrators with PC or higher were allowed to conduct observations in 2012–2013. Eleven percent of administrators scored NYC whereas 7% of the administrators scored C. When examining individual components of certification, we saw that administrators struggled most on the representation and accuracy categories. It is noteworthy that 32% of administrators rated in the NYC category on accuracy, which was a far higher proportion of the sample than any other certification component.

Table 2 summarizes our analyses of the accuracy results. We found that administrators struggled on accuracy (match with master raters) and with developing a shared understanding (match with each other). Less than 50% of administrators matched master raters' scores across all standards except for one standard. The highest levels of agreement were in Standard 2 (Designing Coherent Instruction, 69% exact agreement), which is principally concerned with behavior management, organization, and the classroom environment. The lowest levels of agreement were with elements in Standards 1 (Planning, 42%) and 5 (Professional Growth, 43%). Standard 3 (Standards-Based Learning Activities) was somewhat higher, with 48% agreement with master raters. On the Planning, Standards-Based Learning Activities, and Professional Growth standards, administrators tended to score teachers higher than master raters. On the remaining standard, Designing Coherent Instruction, administrators tended to score teachers slightly lower. Across Standards 2, 3, and 4, agreement among administrators was even lower. These results suggest that, upon completion of training, most administrators continued to struggle with accuracy and likely did not yet know the TLF deeply. This final point warrants attention. It suggests that the kinds of training that districts and states are likely to be providing to administrators may not adequately prepare them for the task of conducting observations in high-stakes evaluation systems; specifically, they struggle to translate evidence into scores in consistent ways.

Administrator Perceptions at the Conclusion of Training

End-of-Training Surveys

Administrators' own interpretations of their preparedness told a somewhat different story than the certification results. Table 3 summarizes administrators' perceptions of their ability to conduct observations at the conclusion of training. Administrators' responses indicated that they felt like they learned a lot about conducting observations. Eighty-four percent of training participants believed they learned a lot about how observations fit into the TGDC evaluation system. Sixty-two percent reported they learned a lot about using the district's technology platform for scoring, and 55% said they learned how to score the observations accurately. Forty-five percent of administrators said they learned a lot about how to take good notes for observation. Many administrators also reported learning how to give feedback to teachers following the observations, with 43% saying they learned a lot. Conversely, administrators reported low levels of learning with respect to incorporating TGDC observations into their existing work. For example, only 12% reported that they learned a lot about managing the TGDC with their other responsibilities (with 37% reporting not at all), and a similarly low percentage of administrators reported learning how to help people who are resistant to change learn to improve their practice and how to effectively manage all the strengths and weaknesses of my staff. These survey results indicated that

Table 2 Administrator Rates of Accuracy on the Teaching and Learning Framework Observation Tool ($N = 1{,}009$)

Standard/element	Description	% Exact (Master)	% Exact (Others)
Standard 1	Planning and Preparation	.42	.42
Element 134	Analysis & Use of Assess. Data for Planning	.61	.44
Element 1d1	Standards-Based Learning Activities	.52	.39
Element 1d3	Purposeful Instructional Groups	.65	.47
Element 1d4	Lesson and Unit Structure	.30	.44
Element 1e1	Aligns with Instructional Outcomes	.41	.39
Element 1e2	Criteria and Standards	.30	.38
Element 1e3	Design of Formative Assessments	.17	.44
Standard 2	Designing Coherent Instruction	.69	.52
Element 2a1	Teacher Interactions with Students	.69	.53
Element 2a3	Classroom Climate	.66	.49
Element 2b2	Expectations for Learning and Achievement	.69	.51
Element 2c1	Management of Routines, Procedures, and Transitions	.70	.55
Element 2d2	Monitoring and Responding to Student Behavior	.72	.55
Standard 3	Standards-Based Learning Activities	.48	.44
Element 3a1	Communicating the Purpose of the Lesson	.20	.48
Element 3b1	Quality and Purpose of Questions	.38	.39
Element 3b2	Discussion Techniques	.65	.47
Element 3c1	Standards-Based Projects, Activities, and Assignments	.63	.46
Element 3c2	Purposeful and Productive Grouping of Students	.60	.42
Element 3d1	Assessment Criteria	.10	.38
Element 3d3	Feedback to Students	.67	.49
Element 3e1	Responds and Adjusts to Meet Student Needs	.59	.41
Standard 5	Professional Growth	.43	.32
Element 5a2	Use of Reflection to Inform Future Instruction	.43	.32

Table 3 Administrators' Self-Reports of Learning During Training, End-of-Training Survey Sample ($N = 293$)

Administrator self-report	Not at all (%)	A little (%)	Some (%)	A lot (%)
How observations fit into the TGDC	0.00	3.02	12.56	84.42
How to score observations more accurately	0.00	5.03	40.20	54.77
How to manage the TGDC with my other responsibilities	37.19	23.62	27.14	12.06
How to take good notes when observing	3.03	14.65	37.37	44.95
How to use MyPGS, the online teacher evaluation tool	1.51	7.04	29.65	61.81
How to give teachers feedback based on observed performance	0.51	13.78	42.35	43.37
How to match teachers to appropriate professional development based on their observed performance	22.84	24.87	36.55	15.74
How to help people who are resistant to change learn to improve their practice	27.78	29.29	35.35	7.58
How to effectively manage all the strengths and weaknesses of my staff	27.60	29.11	34.78	8.51

Note. TGDC = Teacher Growth and Development Cycle.

administrators left the training feeling ready to serve as reliable, accurate raters but at the same time felt ill-equipped to handle the implementation challenges associated with the work of conducting observations. Survey responses suggested that administrators felt comfortable with the content of the observation protocol, regardless of when they were asked.

Table 4 summarizes administrators' beliefs about the observation tool; administrators expressed high levels of support of the observation instrument and expressed high confidence in using it. Ninety-nine percent of the administrators responded that they agreed or somewhat agreed that the TLF aligned with their own views of teaching. Meanwhile, 70% agreed or somewhat agreed that they felt comfortable with the observation instrument.

Table 4 Administrators' Views of the Teaching and Learning Framework, End-of-Training Survey Sample ($N = 293$)

Administrator view	Disagree (%)	Somewhat disagree (%)	Somewhat agree (%)	Agree (%)
The Teaching and Learning Framework covers important domains of teaching	0.00	0.52	21.65	77.84
The view of instruction underlying the Teaching and Learning Framework is similar to my view of instruction	0.00	1.03	29.23	69.74
The Teaching and Learning Framework contains the proper amount of detail for observing teaching	2.06	7.22	42.27	48.45
Teaching behaviors are adequately specified in this instrument	1.55	11.34	44.33	42.78
The Teaching and Learning Framework meets the teaching and learning needs specific to my school	1.03	7.69	46.67	44.62
The online system makes the instrument easy to use	3.09	16.49	45.88	34.54
I am comfortable with this observational instrument	9.28	20.10	39.18	31.44

Table 5 Specific Codes Pertaining to Observation Instrument Use

Code	Training interviews	End-of-year interviews
Recording evidence	8 instances	22 instances
	7/24 interviews	14/24 interviews
	29%	58%
Aligning evidence	2 instances	5 instances
	2/24 interviews	3/24 interviews
	8%	12.5%
Scoring	2 instances	4 instances
	2/24 interviews	4/24 interviews
	8%	17%
Additional training and practice	6 instances	15 instances
	5/24 interviews	11/24 interviews
	21%	46%

End-of-Training Interviews

Interview responses allowed us to investigate the survey responses more deeply. Responses across the two interview time points are summarized in Tables 5 and 6. Consistent with the large-scale survey data, focus administrator interviews at the end of training suggested that few focus administrators (less than half) expressed concerns surrounding learning to become accurate, reliable raters, despite their struggles on the certification exercise.

Among focus administrators who did express challenges or learning needs related to becoming accurate raters, what kinds of needs did they nominate? Of the three tasks taught to them during training—objectively recording evidence, aligning this evidence with TLF elements, and scoring each element using specific TLF criteria—focus administrators' instrument-related challenges and learning needs largely fell in the realm of the first of these three tasks. Approximately 29% of focus administrators expressed concerns or challenges about recording evidence during an observation. Focus administrators were concerned about their abilities to script, or type verbatim, everything that occurs during each lesson, a key expectation of observations completed as part of TGDC. When asked about her perspective of the TGDC process, Sara P. indicated, "My worry is just keeping up with the lesson and getting everything down."

Table 6 Overall Codes for Training and End-of-Year (EOY) Interviews

Code	Training interviews	EOY interviews
Time, resources, and logistics	22 instances 14/24 interviews 58%	23 instances 16/24 interviews 67%
Technology	5 instances 5/24 interviews 21%	10 instances 7/24 interviews 29%
Coaching and feedback conversations	3 instances 2/24 interviews 8%	7 instances 5/24 interviews 20%
Managing personal relationships	3 instances 3/24 interviews 12.5%	3 instances 3/24 interviews 12.5%
Using the observation instrument	19 instances 12 12/24/40 interviews 50%	46 instances 18/24 interviews 75%

Focus administrators were not yet confident that they could record evidence in a way that was representative of what was occurring in the classroom during an observation. There was a concern that, while scripting, focus administrators would miss pieces of evidence needed to evaluate the lesson. Kyle A. explained:

> The challenge comes from the need to develop proficiency in gathering enough representative evidence, more details. In scoring a particular element or standard, it's not based on very small things, kind of a total overview, from the beginning of the class to the end of the class. Have some way to be mentally present, get as much evidence as possible, and be fair. From what I see, when I type, I miss things, lose important information. The challenge might be to recognize that, and to improve. To make sure every time I am in the classroom, I get as much evidence as possible.

Several focus administrators expressed concerns about the process of scripting interfering with their ability to gain an accurate sense of what was going on in a classroom they were observing. Rose B. said:

> Another challenge, even with watching the videos, is not being able to see the big picture, since I am busy with gathering the evidence. I am missing what's going on with the children. I want the teachers to know what's going on in that classroom all the time. If you are teaching, Jonny's playing, what are you doing? Are you noticing that this kid's not understanding. I need to see all that. That means I don't get to script everything since I need to see what's going on in the classroom. I think that's going to be a challenge. The whole idea behind this is finding what the kids do. How is this affecting the children?

In terms of the next two tasks required for accurate observations using the instrument — aligning collected evidence with the TLF elements and scoring each element — very few focus administrators expressed concerns or learning needs. Two focus administrators (8% of interviews) expressed concern over their own ability to accurately align evidence from the observation to the rubric, and two administrators (8% of interviews) expressed concerns with their ability to accurately score a teacher using the rubric. In two of these interviews, focus administrators were concerned over their ability to create scores of teaching that would be in agreement with other administrators acting as raters. For example, Ethan J. stated, "Part of it is the wording. The differences between developing, effective, and highly effective. You would wonder am I scoring too low or too high? There are differences amongst us. .. ."

Rather than focusing on learning needs related to scoring accurately, focus administrators were more likely to express needs related to implementing the observations in practice. To this end, 21% of focus administrators indicated needing further practice or training with the observation instrument in order to implement it in their own schools. Focus administrators indicated the need for further training in terms of formal professional development sessions, as well as in the form of ongoing support, coaching, and feedback during the impending practice year. Danita explained:

> I am concerned that I need to practice, practice, practice. Right now it's a no-stakes environment, but if this were to be implemented in the district. This has major implications for a teacher and you want to get it right. You are setting the teacher up for a disaster. So I feel like I would like consistent help from Cara (the district trainer) to just make sure that I'm staying on top of things and keeping my skills going on.

When asked about their concerns in open-ended ways, the vast majority of focus administrators discussed the amount of work involved in implementing the observation policies in their schools. Just as they and their peers indicated in their end-of-training surveys, a full 58% of the focus administrators expressed concerns about the amount of time required to do the observations adequately while juggling other tasks essential to their work in interviews conducted at the end of training. In particular, focus administrators were concerned with the time that would be required when they would be expected to evaluate a larger proportion of their staff using the TGDC in the future.

Aside from accurate use of the observation instrument, focus administrators expressed some concerns around the coaching and feedback conversations that occurred after lesson observations and scoring were complete. In the end-of-training interviews, 8% of focus administrators expressed concerns about having these types of conversations with teachers, particularly because they anticipated that some teachers would experience difficulty receiving the type of feedback they would be providing as part of the TGDC. Finally, 21% of focus administrators indicated concerns with using technology to complete observations, including using the computer to script lessons during observations and learning how to use the online platform for entering evidence and teacher scores.

To summarize focus administrators' perceptions at the end of training: they were leaving initial training with some concerns about their skills at accurately observing and scoring classroom teaching; however, they were primarily concerned about what it would require to implement the larger observation system in practice. Few referenced ways in which the initial training prepared them for this aspect of the work, and a majority identified this as a future challenge or concern about eventual school-wide implementation.

Administrator Perceptions at the End of the School Year

End-of-Year Survey Results

In the intervening year following initial training, some beliefs remained consistent while others changed. As summarized in Table 7, in end-of-year survey responses, most administrators continued to express comfort in using the TLF. Ninety-five percent of administrators somewhat agreed or agreed with the statement: "I understand the standards, components, and elements in the LAUSD Teaching and Learning rubrics." Similarly, they expressed confidence in their ability to collect objective evidence, align evidence, and distinguish between levels of performance on the rubric. They also continued to stress that their greatest challenge was the time and technical requirements of implementing observation policies, and they expressed concern about the additional time demands that would be introduced by having to conduct observations with multiple teachers in their schools. In summary, based on large-scale survey results at the end of initial training and at the end of their practice year, administrators continued to express confidence in their ability to score accurately and their stated learning needs surrounded how to manage the time demands of conducting observations at scale.

End-of-Year Interviews

In comparison to following their initial training, when interviewed at the end of the practice year, focus administrators were more likely to nominate learning needs related to accurate use of the observation instrument than they had been after their initial training. Overall, 75% of focus administrators indicated challenges or learning needs related to use of the observation instrument at the end of the practice year (Table 6). As in the end-of-training interviews, the majority of these identified challenges were related to recording evidence during instructional observations. Just over half of the administrators (58%) indicated challenges or concerns with recording evidence at the end of the year.

Challenges and learning needs in this area were, again, centered on issues with scripting during lesson observations. Administrators were concerned with their own typing skills and their ability to record adequate evidence for the teachers they evaluate. When asked about what was hardest about implementing the TGDC, Portia explained:

Table 7 Administrators' Views of Training Outcomes, End-of-Year Survey Sample ($N = 293$)

Administrator view	Disagree (%)	Somewhat disagree (%)	Somewhat agree (%)	Agree (%)
I feel confident in my ability to use MyPGS, the online observation tool	5.53	18.97	50.99	24.51
I feel confident in being able to collect objective evidence of instruction	1.61	6.83	46.59	44.98
I feel confident in my ability to align evidence	3.17	11.11	55.56	30.16
I understand the standards, components, and elements in the LAUSD teaching and learning rubrics	0.19	3.57	40.48	54.76
I understand the differences between the levels of performance (highly effective, effective, developing, and ineffective) in the teaching and learning rubrics	0.79	5.16	38.89	55.16
I feel confident conducting pre-observation meetings with teachers	0.4	5.53	37.55	56.52
I feel confident conducting post-observation meetings with teachers	0.8	4.78	36.25	58.17
LAUSD's training prepared me adequately for my observation work this year	7.2	20.4	52	20.4

Note. LAUSD = Los Angeles Unified School District.

> Observing the lesson and typing. Typing is hard for me. And that is hard for me. I cannot possibly get everything that the teacher is saying and I don't know how well I'll do putting it onto the platform because I might not be able to have the required evidence to adequately complete it.

At the end of the practice year, focus administrators identified challenges and learning needs in the second and third tasks required for accurate use of the TLF aside from recording evidence: aligning evidence with the TLF and deciding on a score for each element. Three administrators (13%) identified aligning evidence to the rubric as a challenge or learning need, and four administrators (17%) identified scoring as such. Concerns in these areas were strongly related to administrators' understanding of and familiarity with the TLF itself. For example, Bridget explained, "The most difficult thing is aligning the evidence. Because again, it's the common terminology that been actually … would this be evidence for element two or three or both." Similarly, in his discussion of the scoring process, Ron R. said: "Hard: determining what highly effective looks like. When you dig into the rubric, it's not clear what that looks like."

Accordingly, the administrators who nominated learning needs related to scoring expressed a lack of confidence in their own ability to accurately score instruction. Rob M. said, "Scoring is a little more difficult than the norm because we weren't doing that, we were just giving feedback. The precise scores has been difficult—do you give them a 2 or 2.5? Am I being too hard or too easy? You have self-doubt."

This perspective was reflected in their expressed desire for additional training and practice with the observation instrument. At the conclusion of the practice year, 46% of focus administrators wanted more practice or training with the observation instrument. A number of them indicated that, in order to learn the instrument well, they simply needed more practice using it. As Bridget expressed, "I think I would want to go through another full cycle before I really start doing this. For me what would help most would be more practice. Because it was a year but in reality it wasn't."

Although some administrators requested additional training in the form of professional development, others were eager to network with and learn from other administrators in the district. This choice may have been related to a logistical training issue brought up by several administrators. Because additional training sessions were held during the school day, many of the administrators were unable to attend, given their responsibilities at their schools.

Across these data sources, administrators expressed similar perceptions of their competence and perceived learning needs related to the observation instrument over time, however, there was an increase in the proportion of administrators expressing concerns about using the instrument by the year's end. It is possible that administrators may have become more cognizant of their lack of familiarity, or comfort, with the TLF over the course of their practice year. It is also possible that, as the year progressed, they had forgotten things learned in training. Implementation concerns dominated participants'

responses across the year, but administrators appeared to become increasingly aware of their own needs related to accuracy. That said, of the expressed learning needs related to scoring, administrators were more likely to nominate challenges associated with scripting rather than engaging in the content of the observation rubric.

Discussion and Conclusions

In this study, we examined administrators' ability to conduct observations accurately and reliably at the conclusion of a 4-day training conducted with all administrators across a large, urban district. We then examined administrators' perceptions of their competence at conducting observations as well as their perceptions of how well the training prepared them for this work. To do so, we collected longitudinal survey data as well as more intensive interview data with a subset of focus administrators across two timepoints—immediately following completion of the training and at the conclusion of a no-stakes practice year where the administrator had observed one teacher volunteer and had received support sessions from district administrators. Two findings are noteworthy.

First, the certification results suggested that, at the conclusion of training, administrators struggled to perform the skills they had been taught in their training. There was significant disagreement between and among administrators and master raters. This disagreement was most pronounced in the administrators' ability to apply the rubrics accurately. The 4- to 5-day learning opportunity provided by the district is more rigorous than many of the administrator trainings currently being provided by districts. The certification test results suggested that even with this significant investment in administrator learning, there was more knowledge to be learned and used in order to prepare administrators to begin rating the teaching in their schools.

When examining survey and interview responses, we found that administrators felt that they learned skills proximal to the use of the evaluation tool (e.g., unbiased note-taking, rating). In other words, they gained important knowledge of the observation tool and they valued that knowledge. But the knowledge was partial in two ways. First, administrators reported learning less about the process of carrying out the full cycle of an evaluation (e.g., having improvement conversations with teachers) in the context of their full responsibilities (e.g., time management, juggling an entire staff). So although they learned about the protocol, they did not learn how to use the protocol in their own schools; and it was this learning that they nominated as critical for their success in using the observation protocol at scale. The second way the learning was partial concerned accuracy. Administrators perceived a great deal of learning during training and felt that they had the skills necessary to conduct observations accurately and with reliability, which stood in contrast to the certification results.

Over the course of the practice year, administrators reported increased facility with using the TGDC to conduct observations, including transcribing teachers' lessons and conducting pre- and post-observation conferences. Still, administrators expressed a number of additional training needs at the conclusion of the practice year. Many of these focused on needing more detailed knowledge of the rubrics, in particular aligning evidence with specific TLF elements and differentiating between levels on the rubric. They also requested additional training on using TGDC at scale, something not covered well in the initial training.

Administrators' self-reports showed that they both valued the TLF and understood how to use it. Yet their performance on certification exercises, and their self-reports in subsequent interviews and surveys, indicated that beyond the initial week of training, administrators needed additional training opportunities. Some of the stated training needs were focused on deeper knowledge of the protocol (e.g., learning to score teaching accurately and aligning evidence from the observation with the TLF). Others focused on issues using the TLF at scale—for example, having time to talk and think with other administrators doing the work, being concerned about how to manage time when TGDC moved to full implementation, and juggling the learning needs of an entire staff. These training needs are not surprising if we consider the literature on how best to support teacher learning. Perhaps if we hope to see accurate and reliable scores at scale, the same types of needs that have been documented for teacher learning should pertain to administrator learning. Perhaps administrators need learning opportunities that are ongoing, situated in practice, conducted collaboratively with colleagues, and structured for repeated cycles of trying the TGDC, reflecting on that trial with knowledgeable others, and trying it again.

There is a second finding worth noting from these data. If we step back and consider administrators' perceptions at the end of training, a few facts stand out. The administrators felt they increased their knowledge of TGDC from the training. They reported high levels of comfort with many aspects of the process and details of scoring. So although administrators identified areas in which they would like continued training opportunities, in general, they felt positive and prepared with knowledge of the TGDC. However, the certification results suggested that administrators were somewhat inaccurate

in applying the TLF scoring scales at the end of training. They also struggled with adequately drawing on evidence from across the lesson to assign a score. But accuracy and representativeness were not the things administrators wanted to learn more about at the end of training. This reality—that administrators' perceptions of their proficiency and more objective evidence regarding proficiency did not match—is a common and important problem of teaching and learning. How does any teacher help a student achieve a learning goal when the student and teacher do not agree about what needs to be learned? In this specific case, perhaps administrators' extensive experience with observations generally makes them see their understanding of the TLF in a more positive way than certification results suggest. Or perhaps the training, though extensive and detailed, did not cover the aspects of the TGDC on which the principals most wanted support (e.g., integrating the policy into already full professional lives), so that is what they emphasized most stridently in their end of training surveys. Or perhaps the training itself did not give them enough formative feedback about their success in achieving the goals of the training, they did not know there was still a lot to be learned. Our data do not allow us to determine why administrators felt confident and yet did not achieve a level of accuracy that was commensurate with that confidence; however, this is an important issue for districts to understand and address in future trainings.

Limitations

This study has noteworthy limitations. First, the study design draws on a large sample of administrators in LAUSD as well as a subgroup of focus administrators. This is a design frequently used to understand both larger scale patterns and more fine-grained details about those patterns. However, in each case, there was missing data on the instruments administered at the end of the school year. Only 34% of administrators across LAUSD completed the spring survey. The survey response rates potentially undermine our ability to make claims that generalize from the sample of survey respondents to those from the population of administrators in the district. At the same time, survey responses from large-scale studies are notoriously difficult to gather (e.g., recently published survey data on the population of teachers across Tennessee had response rates of approximately 33%; see Koedel et al., 2017). Although our responses rate of 34% is not ideal, it is within the range of previous research. Further, nonrespondents looked similar to those who did respond, at least with respect to their performance on the certification exercise (i.e., the differences between the two groups were not significant).

A second shortcoming of the study is that we only had data from a single certification exercise; we did not, for example, assess administrators' accuracy and reliability at the conclusion of their practice year. This required us to rely on administrators' perceptions of their ongoing training needs. To be sure, administrators' perceptions are important, as they provide us with information on how they are interpreting their preparedness to conduct observations in practice. They also likely guide actions taken by administrators to seek improvement. But given that LAUSD only conducted its certification at the conclusion of initial training, we lack information on whether administrators became more or less accurate conducting observations as they conducted observations throughout the practice year. This reality is not uncommon among training regimes in large districts; rarely do districts provide repeated certification exercises (or calibration exercises) over time. Instead, we recommend that future research studies take up the issue of how administrators' accuracy changes across learning experiences over time.

A final shortcoming involves the unique training context of LAUSD. Few large districts have included the kind of low-stakes practice year and additional ongoing support to administrators that LAUSD offered, potentially undermining our ability to apply these findings to other districts. At the same time, our results suggest that even with a gradual rollout of new observation responsibilities, administrators still expressed a need for further learning at the end of the school year. Thus, it is likely that shorter training experiences and more immediate high-stakes use would only exacerbate the concerns about whether administrators are able to create accurate, reliable observation scores.

Policy Implications

The findings from this study press us to consider what any central office can and should do in order to train administrators in the best way possible. We will not speculate as to what LAUSD could or should have done. Instead, we offer up a learning puzzle that all districts might consider when trying to teach administrators to know and use their new observation protocol. If we accept that administrators' professional roles and prior knowledge may influence the degree to which they achieve the learning goals of the training session, then the puzzle is how does the central office create a training experience in which both the administrator's strengths and weaknesses are productively engaged toward acceptable levels of accuracy

and reliability. Should central offices try to integrate prior professional knowledge during training? Should training begin with a diagnostic assessment of an administrators' knowledge of how to score teaching and then group administrators for differentiated instruction? Should training focus more narrowly on the elements of the protocol most likely to be misunderstood?

Teachers face these instructional questions with students many times a day. For example, fourth graders know things about how to read and understand a story, but there is more to learn, more to question, new ways to approach the text. The implementation of teacher evaluation is no different. How can the rollout of the policy acknowledge that administrators have important knowledge worth building on and have made progress in understanding the TLF but realize there is more progress that needs to be made? These questions do not have simple answers. But if we hope for teacher evaluation to lead to the types of changes in teaching and learning reformers envision, policy makers and practitioners alike will need to develop answers that can be sustained over time.

Acknowledgments

Courtney Bell is currently director of the Wisconsin Center for Education Research and professor of learning sciences at University of Wisconsin – Madison. This study was supported by a grant from the the William T Grant Foundation [181068]. We would like to thank the W.T. Grant Foundation for its generous support of this work. We are also deeply grateful to the administrators in Los Angeles who graciously allowed us to learn from them. Finally, we thank the anonymous reviewers whose guidance and suggestions have improved the manuscript.

References

Barr, A. S. (1931). *An introduction to the scientific study of classroom supervision*. D. Appleton.

Barr, A. S. (1946). The measurement and prediction of teaching efficiency. *Review of Educational Research*, *16*(3), 203 – 208. https://doi.org/10.2307/1168779

Barr, A. S. (1958). Characteristics of successful teachers. *The Phi Delta Kappan*, *39*(6), 282 – 284.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77 – 101. https://doi.org/10.1191/1478088706qp063oa

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. *Handbook of research on teaching* (3rd ed., pp. 328 – 375). Macmillan.

Casabianca, J. M., Lockwood, J. R., McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, *75*(2), 311 – 337. https://doi.org/10.1177/0013164414539163

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement*, *73*(5), 757 – 783. https://doi.org/10.1177/0013164413486987

Cohen, D. K., & Hill, H. C. (2001). *Learning policy: When state education reform works*. Yale University Press. https://doi.org/10.12987/yale/9780300089479.001.0001

Corcoran, S. P. (2016). Potential pitfalls in the use of teacher value-added data. In J. A. Grissom & P. Youngs (Eds.), *Improving teacher evaluation systems: Making the most of multiple measures* (pp. 51 – 26). Teachers College Press.

Danielson, C. (1996). *The framework for professional practice. In enhancing professional practice: A framework for teaching* (pp. 60 – 119). Association for Supervision and Curriculum Development.

Danielson, C. (2002). *Enhancing student achievement: A framework for school improvement*. Association for Supervision and Curriculum Development.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching*. Association for Supervision and Curriculum Development.

Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research*, *53*(3), 285 – 328. https://doi.org/10.3102/00346543053003285

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267 – 297. https://doi.org/10.1002/pam.21818

DeMoulin, D. F. (1988). Staff development and teacher effectiveness: Administrative concerns. *Focus*, *81*, 36 – 38.

Donaldson, M. L., & Donaldson, G. A. (2012). Strengthening teacher evaluation. What district leaders can do. *Educational Leadership*, *69*(8), 78 – 82.

Donaldson, M. L., & Mavrogordato, M. (2018). Principals and teacher evaluation: The cognitive, relational, and organizational dimensions of working with low-performing teachers. *Journal of Educational Administration*, *56*(6), 586 – 601. https://doi.org/10.1108/JEA-08-2017-0100

Donaldson, M. L., & Papay, J. P. (2014). Teacher evaluation for accountability and development. In H. F. Ladd & M. R. Geortz (Eds.), *Handbook of research in education finance and policy* (pp. 174 – 193). Taylor & Francis Group.

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

Gallagher, H. (2004). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education*, *79*(4), 79 – 107. https://doi.org/10.1207/s15327930pje7904_5

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

Gill, B., Shoji, M., Coen, T., & Place, K. (2016). *The content, predictive power, and potential bias in five widely used teacher observation instruments* (REL 2017 – 191). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. http://ies.ed.gov/ncee/edlabs

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. (2019). 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

Grossman, P., & McDonald, M. (2008). Back to the future: Directions for research in teaching and teacher education. *American Educational Research Journal*, *45*(1), 184 – 205. https://doi.org/10.3102/0002831207312906

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know*. Harvard Education Press.

Harvey, M. W., Boyland, L. G., & Quick, M. M. (2019). An investigation of teacher evaluation practice in Indiana: PL 90 implementation and issues for administrators. *International Journal of Educational Reform*, *28*(1), 24 – 47. https://doi.org/10.1177/1056787918824191

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record*, *116*(1), 1 – 28. https://www.tcrecord.org/Content.asp?ContentId=17292

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education*, *17*(3), 207 – 219. https://doi.org/10.1007/s11092-005-2980-z

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher*, *39*(8), 591 – 598. https://doi.org/10.3102/0013189X10390804

Kersten, T. A., & Israel, M. S. (2005). Teacher evaluation: Principals' insights and suggestions for improvement. *Planning and Changing*, *36*, 47 – 67.

Kimball, S., White, B., Milanowski, A., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education*, *79*(4), 54 – 78. https://doi.org/10.1207/s15327930pje7904_4

Koedel, C., Li, J., Springer, M., & Tan, L. (2017). The impact of performance ratings on job satisfaction for public school teachers. *American Educational Research Journal*, *54*(2), 241 – 278. https://doi.org/10.3102/0002831216687531

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

Leahy, C. (2012). *Teacher evaluator training: Ensuring quality classroom observers*. Education Commission of the States. www.ecs.org/clearinghouse/01/01/14/10114.pdf

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*, 61 – 95. https://doi.org/10.1007/s11092-018-09291-3

Lochmiller, C. R., & Mancinelli, J. L. (2019). Principals' instructional leadership under statewide teacher evaluation reform. *International Journal of Educational Management*, *33*(4), 556 – 568. https://doi.org/10.1108/IJEM-02-2018-0076

McGhee, M. W., & Lew, C. (2007). Leadership and writing: How principals' knowledge, beliefs, and interventions affect writing instruction in elementary and secondary schools. *Educational Administration Quarterly*, *43*(3), 358 – 380. https://doi.org/10.1177/0013161X06297202

McGuinn, P. (2012). *The state of teacher evaluation reform*. Center for American Progress. https://doi.org/10.12698/cpre.2012.stateteacherevaluation

Mihaly, K., & McCaffrey, D. F. (2014). Grade level variation in observational measures of teacher effectiveness. In K. Kerr, R. Pianta, T. Kane (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 9 – 49). Jossey-Bass. https://doi.org/10.1002/9781119210856.ch2

Milanowski, A. (2004, April 12–16). *Relationships among dimension scores of standards-based teacher evaluation systems, and the stability of evaluation score–student achievement relationships over time* (CPRE-UW Working Paper Series TC-04-02). [Conference presentation]. American Educational Research Association Conference, San Diego, CA, United States. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.568.1037&rep=rep1&type=pdf

Painter, S. R. (2000). Principals' efficacy beliefs about teacher evaluation. *Journal of Educational Administration*, *38*(4), 368–378. https://doi.org/10.1108/09578230010373624

Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education*, *121*(2), 183–212. https://doi.org/10.1086/679390

Reddy, L. A., Dudek, C. M., Peters, S., Alperin, A., Kettler, R. J., & Kurz, A. (2018). Teachers' and school administrators' attitudes and beliefs of teacher evaluation: A preliminary investigation of high poverty school districts. *Educational Assessment, Evaluation and Accountability*, *30*(1), 47–70. https://doi.org/10.1007/s11092-017-9263-3

Ritchie, J., Lewis, J., Nicholls, C. M., & Ormston, R. (Eds.). (2013). *Qualitative research practice: A guide for social science students and researchers*. Sage.

Rockoff, J. E., Staiger, D. O., Kane, T. J., & Taylor, E. S. (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review*, *102*(7), 3184–3213. https://doi.org/10.1257/aer.102.7.3184

Ronfeldt, M. C., & Shanyce, L. (2016). Evaluating teacher preparation using graduates' observational ratings. *Educational Evaluation and Policy Analysis*, *38*(4), 603–625. https://doi.org/10.3102/0162373716649690

Rowan, B., & Raudenbush, S. W. (2016). Teacher evaluation in American schools. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 1159–1216). American Educational Research Association. https://doi.org/10.3102/978-0-935302-48-6_19

Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of Educational Psychology*, *82*(3), 498–504. https://doi.org/10.1037/0022-0663.82.3.498

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, *38*(2), 293–317. https://doi.org/10.3102/0162373715616249

The Danielson Group. (2021). *Our history*. https://danielsongroup.org/our-story

White, M. C. (2018). Rater performance standards for classroom observation instruments. *Educational Research*, *47*(8), 492–501. https://doi.org/10.3102/0013189X18785623

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observers: Lessons learned in four districts*. Brown Center on Education Policy at Brookings. https://www.brookings.edu/wp-content/uploads/2016/06/Evaluating-Teachers-with-Classroom-Observations.pdf

Wieczorek, D., Clark, B., & Theoharis, G. (2018). Principals' perspectives of a race to the top-style teacher evaluation system. *Journal of School Leadership*, *28*(5), 566–595. https://doi.org/10.1177/105268461802800501

Wind, S. A., Jones, E., Bergin, C., & Jensen, K. (2019). Exploring patterns of principal judgments in teacher evaluation related to reported gender and years of experience. *Studies in Educational Evaluation*, *61*(8), 150–158. https://doi.org/10.1016/j.stueduc.2019.03.011

Youngs, P. (2007). How elementary principals' beliefs and actions influence new teachers' experiences. *Educational Administration Quarterly*, *43*(1), 101–137. https://doi.org/10.1177/0013161X06293629

Suggested citation:

Jones, N., Bell, C., Qi, Y., Lewis, J., Kirui, D., Stickler, L., & Redash, A. (2021). *Certified to evaluate: Exploring administrator accuracy and beliefs in teacher observation* (Research Report No. RR-21-05). ETS. https://doi.org/10.1002/ets2.12316

Action Editor: Don Powers

Reviewers: Daniel Fishtein and Gary Sykes