ETS

Measuring the Power of Learning.™

Research Report
ETS RR–18-18

Administrators' Uses of Teacher Observation Protocol in Different Rating Contexts

Yi Qi

Courtney A. Bell

Nathan D. Jones

Jennifer M. Lewis

Margaret W. Witherspoon

Amanda Redash

December 2018

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

Administrators' Uses of Teacher Observation Protocol in Different Rating Contexts

Yi Qi,[1] Courtney A. Bell,[1] Nathan D. Jones,[2] Jennifer M. Lewis,[3] Margaret W. Witherspoon,[1] & Amanda Redash[2]

1 Educational Testing Service, Princeton, NJ
2 Boston University, Boston, MA
3 Wayne State University, Detroit, MI

Teacher observations are being used for high-stakes purposes in states across the country, and administrators often serve as raters in teacher evaluation systems. This paper examines how the cognitive aspects of administrators' use of an observation instrument, a modified version of Charlotte Danielson's Framework for Teaching, interact with the complex and dynamic rating contexts in applied settings. Findings suggest that administrators' rating strategies and rating approaches vary as the characteristics of the rating contexts differ. Even shortly after training (and more so as time passed), raters used reasoning strategies not supported by their training to make scoring decisions. We discuss the implications of the findings for the training of raters and the development of evaluation systems in high-stakes contexts.

Keywords Observation; teacher quality; measurement of teaching; rater cognition

doi:10.1002/ets2.12205

In recent years, spurred by state and federal initiatives (e.g., Race to the Top, No Child Left Behind, Teacher Incentive Fund), states and districts across the country implemented teacher evaluation systems to improve the quality of instruction (Sawchuk, 2009). These systems include multiple measures but often include two major components: first, information derived from measures of teaching effectiveness that depend on student achievement data—student growth percentiles (Betebenner, 2011) and valued-added measures (e.g., Braun, Chudowsky, & Koenig, 2010; Lockwood et al., 2007), and second, observational measures of teachers' instructional practices in the classroom. Much of the literature has focused on the measures based on student achievement data (e.g., value-added measures and student growth percentiles) and the concerns associated with using them in teacher evaluation (Ballou & Springer, 2015; Berliner, 2013). Research on the quality of practice-based observation systems has been slower to develop, although recent work is beginning to supply needed insight (Kraft & Gilmour, 2016; Steinberg & Sartain, 2015; Whitehurst, Chingos, & Lindquist, 2014).[1]

Although the consequences associated with using observations to evaluate teachers vary, observation ratings are included in high-stakes evaluation systems in many states. For high-stakes systems to support the improvement of teaching to be seen by stakeholders as legitimate for sorting and rewarding teachers, accurate observation ratings are necessary. Therefore, various factors associated with the processes through which observation ratings are created, including those related to both raters and rating contexts, require policymakers' attention (Hill, Charalambous, & Kraft, 2012).

Little is known about the processes raters use to assign ratings to observation data. Bejar (2012) reviewed literature on rater cognition, that is, raters' mental processes in the creation of assigning scores to a performance of some type. Among other issues, the interaction of contextual factors with rater cognition when assigning classroom observation ratings remains unclear.

School administrators in the United States often serve as raters for conducting teacher evaluations, so understanding the interactions of context and cognition matters. The high-stakes environment of teacher evaluations, in contrast with research settings, presents a number of challenges: Administrators' cognitive processes likely differ from raters in research settings; rating contexts hold new complexities, such as personal and professional relationships between and among teachers and administrators; and district policies could distort observation efforts in practice. These diverse challenges impose risks in creating reliable ratings. At a general level, recent research suggests that principals are aware that their frank

Corresponding author: C. A. Bell, E-mail: cbell@ets.org

assessments of teacher quality do not always match the summative teacher evaluation scores the principal assigns (Kraft & Gilmour, 2017). At the more specific level of observation scores, in a research setting, almost always the ratee is unknown to the rater, whereas an administrator's prior experiences working with a particular teacher might influence the dynamics of an observation either positively or negatively and thus be reflected in the observation ratings assigned. And because administrators' ratings factor into human capital consequences such as promotion or retention, administrators' and teachers' attitudes toward observations could be affected and consequently shape teachers' performances or administrators' rating behaviors.

Multiple social, organizational, and political factors shape the rating contexts in high-stakes teacher evaluation systems (Darling-Hammond, 1990). It should be noted that such contexts often do not remain stable, as administrators use the evaluation tool in various ways over time in the field. For example, a district's evaluation policy may change in terms of what resources are allocated for conducting teacher evaluation, the process of the evaluation, and how evaluation ratings will be used. The extent to which administrators' rating practices vary during this dynamic process of complicated cognitive rating work amidst shifting contextual features remains unclear and warrants increased attention.

To better understand administrators' rating behaviors while they engage in observational ratings in real-world contexts, we investigated how administrators in the Los Angeles Unified School District (LAUSD) used the observation tool being implemented as a part of the district's new teacher evaluation system, known as the Teacher Growth and Development Cycle (TGDC).[2] Specifically, we focused on how the cognitive aspects of rating varied across two rating contexts, a practice year and an implementation year. These two contexts differ in terms of administrators' experiences, volume of work, rating capacity, and proximity to training, as well as consequences of evaluation. We analyzed administrators' think-aloud data from these contexts to address the following research question: To what extent do administrators' reasoning processes vary as they gain greater experience with their district observation tool?

Theoretical Framework

Using Observation Protocols as a Tool

Observation protocols serve to nominate teaching practices or characteristics to which administrators should attend while observing classroom practice. When administrators rate teachers' instructional practices, they interpret and implement an observation protocol. This process is shaped by a number of factors. Vygotsky's theory about humans' use of tools to mediate social environments has been applied to the dynamics of individuals using various objects, artifacts, or tools in education (Drijvers, Doorman, Boon, Van Gisbergen, & Gravemeijer, 2007; Lantolf & Appel, 1994; Lee & Smagorinsky, 2000). Brown (2009), for example, studied the interactions between teachers and the curriculum materials they use and argued that "curriculum materials play an important role in affording and constraining teachers' actions"; at the same time, "teachers notice and use such artifacts differently given their experience, intentions, and abilities" (p. 19). This recognition of the bidirectional influence from tool to individual and from individual to tool is relevant to administrators' use of observation protocols in teacher evaluation. Although observational protocols enable administrators to make use of a set of standards to evaluate teaching practices, they also provide boundaries to constrain rating behaviors in certain ways. Both enable administrators to produce accurate and reliable ratings.

As administrators interact with observation protocols, their use of the tool varies based on their prior experiences, their perceptions of the goals of teacher evaluation, and their knowledge about teaching. Each of these factors functions differently in the specific rating contexts in which administrators operate. And each of these factors will have an impact on the evidence administrators draw from observations and how they use the rubric to evaluate the evidence they notice. We conjecture that administrators tend to integrate their professional training, professional experiences, and own understanding about teaching with the observation protocol used when evaluating teaching. In this study, we try to understand such interactions in specific rating contexts by isolating and investigating administrators' cognitive processes when rating teaching practices.

Rater Effects and Rater Cognition

Although our theory of tool use draws from Brown (2009), curriculum materials and observation protocols differ in critical ways. Most importantly, it is acceptable (and often encouraged) for teachers to interpret and adapt curriculum in

ways responsive to their students and the wider context; that is, teachers' instructional practices should take into account the curriculum resources and the teachers' prior practices and beliefs, as well as the specific student needs and resources in the classroom. In contrast, observation protocols are designed and implemented with the intent that ratings be replicable by any rater trained on the observation rubric regardless of the rater's experience or background so that accurate and reliable information can be generated about teaching performance (Bell et al., in press). Therefore, in examining how administrators interact with observation protocols, researchers need to consider rater factors that shape the ways in which raters use protocols and how those relate to issues of reliability and accuracy.

In their work examining teacher evaluation practices in US schools, Rowan and Raudenbush (2016) discussed risk and distortion, two threats to the key qualities of performance information. Risk and distortion are sometimes referred to as the *criterion problem* in performance measurement. Risk concerns the extent to which variations in measures reflect biased or chance variations rather than the actual efforts of an employee. Distortion concerns the degree to which the measures fail to adequately measure the true value of an employee's performance. In observation-based teacher evaluation, risk is associated with observation measures that are unreliable and biased, causing teachers' performances to be evaluated with error. Much of this error is related to raters (Casabianca et al., 2013; Hill et al., 2012; Kane & Staiger, 2012). A situation one would reasonably expect is that consensus among raters regarding standards will develop gradually (Hoskens & Wilson, 2001). Studies such as the Measures of Effective Teaching project and Understanding Teaching Quality study show, however, that even with extensive training and ongoing monitoring, low inter-rater agreement remains a major concern, especially on some aspects of classroom instruction (Bell et al., 2014; Gitomer et al., 2014; Kane & Staiger, 2012). In the rating of written responses, researchers have explored various sources of error, including the relationship between the frequency or volume of rating and the severity and consistency of scores over time (Braun, 1988; Congdon & McQueen, 2000; Hoskens & Wilson, 2001; Lim, 2011; Myford & Wolfe, 2009). Researchers found that rater severity varied based on which day ratings occurred during the rating program; findings also suggested that raters tend to drift toward the center rating categories. Further, there is evidence that raters' performance improves once they reach a certain rating volume. Finally, the organizational role of the rater and the social relationship between rater and ratee has been found to influence rater bias (Ho & Kane, 2013; Lefkowitz, 2000).

Additional research has documented the issue of distortion in observation-based teacher evaluation. Specifically, impacted by the "inability to arrive at a consensual definition of the teacher 'performance' construct" in American education (Rowan & Raudenbush, 2016, p. 1172), various observation protocols being developed show a lack of consensus around what dimensions of teaching are being measured. In practical applications, as a result, any particular protocol used in an evaluation system will present some distortion, or a certain extent of divergence from the true value of teaching performance of interest. The variation in instruments' dimensionality, along with the theoretical emphases behind the constructs, creates challenges for raters and further warrants researchers' attention to rater cognition when rating teaching.

Raters account for much of the error in observation scores (Casabianca, Lockwood, & McCaffrey, 2015; Ho & Kane, 2013; Rowan & Raudenbush, 2016). Although research has documented multiple cognitive factors that may present risks to score reliability, such as halo effects, leniency or severity effects, and the central tendency effect (Leckie & Baird, 2011; Saal, Downey, & Lahey, 1980), not much is known about whether and how raters' cognitive processing could contribute to the problem of distortion.

Administrators, like raters in research settings, do not begin learning to use an observation protocol in teacher evaluation systems as if their minds are blank slates. The process of creating observation ratings is characterized by constructing an internal representation of the "abstract concepts and dynamic activities" specified by the observation protocol (Brown, 2009, p. 21). In this process, administrators' background, prior experiences, and idiosyncratic beliefs about teaching all play a part in how their own internal representation of the observation rubric is constructed. Therefore, regardless of what protocol is being used in a local evaluation system, it is likely that an administrator will hold beliefs that are not in complete alignment with the particular dimensions of teaching being measured. Although the need to discipline administrators' thinking and rating behavior is acknowledged, and trainings and monitoring are often in place, there is little empirical research on how administrators actually interact with and make sense of the observational tool they are trained to use. One such effort was Bell et al.'s (2014) study comparing trained raters in the Understanding Teaching Quality study and master raters in how they use various reasoning strategies to rate a lesson. In the research setting, they found that master raters used the observation rubric more consistently to reason to a score compared to the trained raters, who more often

used strategies other than the rubric, such as reasoning from memorable training, or calibration videos, or use of their own internal criteria.

In practical settings, administrators interact with much more complicated contextual factors than in research settings. The social and organizational contexts could vary dramatically in different evaluation environments as well as at different phases of the implementation of an evaluation system. There is limited research exploring how raters' cognitive processing varies when learning to rate or when rating contexts differ. In a study to examine rater cognition while teachers assess project work, a *configurational judgment model* was examined, that is, the evaluators assess the quality of the work as a whole first and then verify the judgment using references to criteria (Crisp, 2012; Sadler, 1989). In Crisp's (2012) study, it was hypothesized that with more experience, raters would tend to apply a more intuitive and rapid reasoning process rather than rely on the scoring criteria and engage in an analytic process to judge the features of the evidence. It is not clear whether and how such findings transfer to settings in which administrators use observation tools. Further discussion on the rating contexts below demonstrates how rater cognition may interact with the contextual factors and hence have an effect on rating accuracy and reliability.

Context

Research has documented the influences of the social and organizational contexts on teacher evaluation (Darling-Hammond, Wise, & Pease, 1983; Ellett & Teddlie, 2003; Hausman, Crow, & Sperry, 2000). Drawing on research from organizational theory and agency theory, Rowan and Raudenbush (2016) reviewed patterns in personnel evaluation, broadly conceived. They explained how the important constraints in organizations influence the development and use of performance-based evaluation systems, and as a result, "these constraints temper the weight given to objective measures of performance in evaluation practice" (p. 1167). Such constraints often stem from the characteristics of the evaluation contexts, particularly in relation to how performance goals are defined, whether there are well-developed processes to reach the performance goals, and to what extent the performance measures are subject to risk and distortion. One issue they identified in the US teacher evaluation context is the "incomplete (and potentially idiosyncratic) representation of the larger performance domain" (p. 1172).

In this study, we are interested in connecting the micro-level cognitive processes (i.e., the use of tools and cognitive reasoning) with the more macro-level context of administrators' use of district observation tools. We define *context* as

> a system of interior and exterior factors and conditions of human behavior and activity, which may affect perception, understanding and transformation of a particular situation, and which determine the meaning and sense of the situation as a whole and its comprising components. (Verbitsky & Kalashnikov, 2012, p. 118)

Based on this conceptualization, the study focuses on the internal cognitive aspects of administrators' work when assigning observation ratings and explores how this performance measure is subject to risk and distortion given the reality of external contextual factors. We identify the specific contextual factors of focus in the next section.

LAUSD's Teacher Growth and Development Cycle

This study was conducted in LAUSD, the second largest public school district in the United States, with about 800 schools serving approximately 670,000 students. This district has a racially and socioeconomically diverse student and teacher population. In the 2012–2013 school year, LAUSD began the practice year of initial implementation of its new teacher evaluation system, the TGDC. More than 1,000 administrators were trained and required to become certified to use an observation protocol, the Teaching and Learning Framework (TLF). The TLF is a modified version of Danielson's (2007) Framework for Teaching, which was developed to establish a shared understanding of effective teaching across the district aligned to California teaching standards. In the practice year of implementation, the TLF rubric included four rating levels (*ineffective, developing, effective*, and *highly effective*), and administrators were asked to rate participating teachers on 21 focus elements of the 63 total in the TLF. Administrators went through a weeklong training session that included 4 full days of instruction and 1 day of a mandatory certification test. Training generally focused on learning the content of the focus elements, taking unbiased evidence, and justifying a score with that evidence. Administrators were taught to take careful evidence, align the evidence with the focus elements, and then use the rubric to assign the score. Although the research

Table 1 Comparison of Rating Context Features

Contextual features	Practice year	Implementation year
Rater experience	Limited practice	Extensive practice
Volume of work/capacity	One teacher/fewer time constraints	Multiple teachers/more time constraints
Consequence of evaluation	No stakes	High stakes
Proximity to training	Closer to training	Further away from training

team saw only a small fraction of the training sessions, trainers used a common set of slides, assignments, and videos. Across the five trainings, researchers observed a good deal of fidelity to the slides and consistency in assignments and practice scoring. More detailed descriptions about the training process can be found in the *Understanding Consequential Assessment Systems for Teachers* (UCAST) Year 1 final report to LAUSD (Bell et al., 2016).

We now describe the two practice contexts in which administrators were learning the observation tool and honing their knowledge and skills. In 2012 – 2013, administrators conducted observations with only one teacher, who either volunteered or was requested by their administrator to participate in the practice year observation cycles. Data suggested that the volunteering teachers who participated were primarily hard-working, high-performing teachers with whom the principal had a trusting relationship (Bell et al., 2016; Strunk, Weinstein, & Makkonen, 2014). In this year, the implementation of TGDC was executed in a low-stakes environment: There were no formal consequences associated with the evaluation efforts and results.

Following the practice year, LAUSD moved to full implementation of the TGDC, and for the first time, the TGDC was used to make human capital decisions across the district, with evaluation results attached to retention and other consequences. Administrators were expected to select a subset of teachers for formal evaluation, and the selection decisions were based on district policy as well as administrators' discretion. Teachers on probationary status or those in a contractual observation year were required to be evaluated; administrators often included teachers who were new to the school, were perceived to be struggling, and those who volunteered to participate (Bell et al., 2015).

In both years, administrators went through the mandated cycle with each teacher, which included formal goal-setting with the participating teacher at the beginning of the year, two formal observations with pre- and post-observation conferences, informal observations, a final evaluation meeting, and documentation of that meeting at the end of the year. This study collected data in the form of think-aloud interviews during these 2 years and tried to understand administrators' reasoning strategies and rating approaches when they were in two different contexts of the TGDC implementation.

As illustrated in Table 1, several features differed across the 2 years of the study, creating two different rating contexts. First, the volume of the evaluation work an administrator was responsible for and the amount of practice they had varied greatly. In the practice year, each administrator was working with only one volunteer teacher, whereas in the implementation year, an administrator was working with multiple teachers and thus had opportunities to practice using the evaluation tool regularly. Second, given the different workload and capacity constraints, administrators generally could spend more time with the volunteer teacher in the practice year, whereas in the implementation year, each teacher got limited time and attention, comparatively. Third, the consequences of the evaluation ratings were very different, which likely shaped the way administrators worked with the tool. In the practice year, the evaluation ratings were not a part of any official record, but in the implementation year the evaluating ratings became a component of the teacher's file that had high-stakes consequences. Finally, the evaluation observations and our first round of think-aloud data collection were conducted closer to administrators' formal training on the observation protocol during the practice year. In the implementation year, administrators were further away from their formal training. At the same time, the rubric and the evaluation process were not as new to them as in the previous year. We hypothesize that these contextual features matter to the strategies administrators learned to use to score the TGDC.

Research Methods

This analysis was a part of the UCAST project, which collected extensive data on administrators who were being trained as raters in LAUSD in 2012 – 2013 and 2013 – 2014. The full body of data collected in the project is documented elsewhere (Bell et al., 2015; Bell et al., 2016). In addition to large-scale data collection across all administrators in LAUSD, extensive and in-depth data were collected from a group of administrators who were recruited and agreed to participate in a series

of research activities to share their experiences and thoughts about the TGDC implementation. This analysis draws almost exclusively on think-aloud data from this group of administrators, which provides rich information on how administrators cognitively approached the TGDC.

Participants

A group of administrators from 2012 to 2013 were recruited during the first TGDC training session in summer 2012. Research team members visited four training sites in LAUSD to introduce the UCAST study and recruit participants on site. Across seven training classrooms in three out of the four training sites, the study was described verbally by the UCAST researcher to the administrators, and consent forms were passed out. One hundred twenty-one administrators returned their consent forms and volunteered to participate as administrators. From the volunteers, through a stratified random selection process to balance job role and grade level, 42 administrators were chosen to participate in our study.

To recruit administrators in 2013 – 2014, recruitment emails were sent to administrators participating in summer training sessions. Similar to the process in 2012, on-site recruitment was conducted by research team members during summer training sessions at three LAUSD venues. In addition, administrators from year 2012 to 2013 were requested to continue participating in the study. Thirty-eight administrators were recruited to participate in the second year of the study, 23 of whom had participated during 2012 – 2013.

This study examines think-aloud data from 18 administrators who were trained and certified in TGDC and who participated in the UCAST project for 2 years. The study collected much more think-aloud data than are included in our sample. We narrowed first to 23 administrators for whom we had 2 years of data and then removed any administrator who did not have complete data. We did this to ensure comparability across time points. Among them, nine were principals, five were assistant principals, one was an instructional coach, and three were instructional directors — individuals who are the direct supervisors of principals and serve as key leaders on local educational service teams to coordinate instructional and evaluation activities to ensure academic success. Seven of the 18 administrators worked in elementary schools, eight worked in secondary schools, and the three instructional directors worked with more than one school.

Think-Aloud Design

Think-aloud exercises are one commonly used method to study scoring behaviors (Suto, 2012). In this study, for each think-aloud exercise, administrators were shown a 10-minute video-recorded lesson clip with the associated transcripts. Administrators rated the clip on three focus elements of the TLF and talked out loud about their thinking and reasoning process when watching the video, aligning evidence to the appropriate element and assigning ratings. For both years' think-aloud exercises, administrators were asked to align one piece of evidence from the video (they could use the transcripts or the notes they took while watching the video clip) and assign ratings to each of the following three focus elements they had been trained on: expectations for learning and achievement; quality and purpose of questions; and standards-based projects, activities and assignments. When they completed rating, they answered a researcher's clarifying questions about their ratings in the stimulated recall part of the exercise. They also were asked to reflect on the degree to which their rating experience using the video was similar to how they use the protocol when they score their own teachers. The protocol is available in the Appendix of this report.

Two different video clips were used for the two think-aloud exercises in the practice year and the implementation year. The video used in the practice year depicted the homework review and launch of a seventh-grade mathematics lesson. The implementation year video featured a first-grade mathematics lesson reviewing the concepts of odd and even numbers. Most think-aloud sessions were conducted over the phone using a screen-sharing application for research team members to guide and observe participants' note-taking and evidence-aligning activities throughout the process. In the practice year, a few think-aloud sessions were completed in person when research members visited LAUSD. Research team members took as close to verbatim notes as possible during all think-aloud exercises. All notes were imported into NVivo for coding analyses.

It is worth noting that the focus of the study is not how administrators' use of the scoring scales might differ when rating a teacher with whom they do not have a professional relationship versus when they are rating a teacher they supervise. We presume that the cognitive work administrators do during observations in which they do not know the teacher will be similar to, although not exactly the same as, the work they do when they supervise the teacher. By standardizing the

Table 2 Descriptions of Reasoning Strategy Codes

Reasoning strategy codes and subcodes	Code definition
Referencing scoring criteria	Referencing or rereading the rubric to review and evaluate the descriptions in the various scoring categories
Starting from a score	Beginning with a specific score point and then deciding whether the evidence matches or not and whether he/she needs to move up or lower to another score
Using information from other classrooms	Comparing the evidence they collect from the lesson with what they have seen or experienced in other classrooms to make a judgment
Lesson improvement	Speculating about lesson improvement through talking from the point of view of what the teacher should have done or what they could have done to make them reach a higher score point
Postobservation conversation	Discussing lesson improvement from the perspective of planning for postobservation/reflection conversation with the teacher
Using internal criteria	Using personal criteria not clearly specified or supported by the rubric to judge the teaching practice and decide on a score
Personal views of teaching	Judging according to personal beliefs on teaching, subjective preference, or personal focus on particular teaching practices that are not specified by the rubric element
Professional experience and knowledge	Making a judgment using knowledge of teaching acquired from previous professional experience; training or knowledge that is not specified by the rubric element

lessons and removing the social issues that influence administrators' behaviors when they rate teachers they know, we are better able to understand how administrators' cognitive processes vary when administrators are in different contexts (time from training, workload, and so on). Insights from this type of analysis will have implications for the cognitive aspects of observation training as well as the design and implementation of evaluation systems built on the commonplace assumption that principals reason similarly.

Using the think-aloud design, we are trying to understand what administrators' rating strategies and rating processes look like when they are using the tool regularly in two specific rating contexts. These contexts are similar to the range of contexts other districts could expect to see as they move forward implementing new teacher evaluation systems. They correspond to what many policy reformers might recommend as an early stage implementation and a later stage implementation.

Data Analysis

Qualitative data analyses were conducted to examine the 18 sets of think-aloud notes from both years. One set refers to think-aloud notes for a single administrator from the practice year and the implementation year.

Coding was conducted in several steps. After think-aloud data from the practice year were collected, the research team identified five reasoning strategies administrators used in their rating process through three rounds of open coding, code definition, refinement, recoding, and accuracy checks. Subsequently, the research team applied the same coding scheme to the think-aloud data from the implementation year. Twenty percent of the notes were double coded; inter-rater agreement was over 95% on all codes.

More detailed descriptions about the coding process and reasoning strategies can be found in Bell et al. (2016). Table 2 describes these reasoning strategy codes identified across 2 years. Among these strategies, only two — referencing scoring criteria and starting from a score — are taught in training and supported by the TLF protocol.

After completing the first round of coding on both years' data, through debriefing and discussions, the research team identified a few additional subcodes to further investigate what criteria or resources administrators drew upon when using criteria not clearly specified or supported by the rubric. Specifically, two major themes emerged and were identified as new subcodes for using internal criteria; those subcodes are *personal views of teaching* and *professional experience and knowledge*. Sometimes administrators' impressions or judgments were shaped by their personal beliefs about teaching, their subjective preferences, or a focus on particular teaching practices not necessarily specified by the rubric; for example, when rating element 3c1, *standards-based projects, activities, and assignments,* although visual aids were not part of the criteria referenced in the rubric, Bridget[3] said,

Table 3 Descriptions of Scoring Approach Typology Codes

Scoring approach categories	Scoring approach typology definition
Use of tool—use of rubric	Using the TLF rubric to guide reasoning; read or reviewed the rubric, gauged evidence against the rubric descriptions, and decided on a rating accordingly
	Using the TLF rubric to confirm or justify the rating decisions already made
	Not explicitly using the TLF rubric to facilitate reasoning and assigning ratings
Use of tool—analysis of rubric attributes	Reviewing and considering all rubric attributes when reasoning to assign a rating
	Reviewing and considering only certain rubric attributes when deciding on a rating
	Assigning a rating without explicitly reviewing or considering rubric attributes
Use of evidence	Reviewing or discussing evidence from the video to generate ideas about ratings
	Searching for evidence to support their initial decision on ratings

> I'm always looking for visual aids with projects. Part of it is it's so important that there are multiple ways the students can be explained what they are going to work on and why … [referring to the video clip] I don't see anything other than numbers.

In other cases, administrators made judgments using knowledge or understanding of teaching acquired from their previous professional training or their knowledge about instructional tools that were not specified in the rubric. For example, Scott commented: "(for the small group activity) … [t]here was ample opportunity for everyone to be successful. Going from writing to manipulating, you are meeting many different learning objectives. Gardner's multiple intelligences were being addressed." In this case, Scott supported his judgment based on a learning theory he valued, which was not specified in the rubric but drawn from his own prior knowledge.

A third subcode concerned administrators' strategy of referring to the hypothetical postconference conversation with the observed teacher in reasoning about the rating for the teacher, under the *lesson improvement* strategy. Heather, for example, talked out loud when rating element 3b1, *quality/purpose of questions*:

> Maybe she is differentiating here … I would like to elicit some more information (during the postobservation conference). It'd be a good cognitive coaching question. I think I would largely probably push for *ineffective* but start off the conference with *developing*.

It is important to note that these administrators did not know the teachers in the videos and would not provide feedback to them. This is why we refer to the conference as a hypothetical conference. These subcodes were applied to both years' data.

Although identifying the reasoning strategies administrators used provided information about the reasonableness of the administrators' judgments, we were also interested in learning how administrators went about applying the various strategies to assign ratings. Administrators' strategy applications were investigated by looking at the function of the TLF rubric in administrators' scoring process or how administrators used the rubric and analyzed the rubric attributes as well as how they made use of the evidence from the video. Descriptions of the scoring approach typology are provided in Table 3. Three research team members applied the scoring approach typology codes to all think-aloud notes. About 10% of the data were double-coded, and the exact rater agreement was 91%. After coding was completed, comparisons of administrators' scoring behaviors over the 2 years were made.

The degree to which specific strategies were related to creating accurate ratings is an important aspect of understanding administrators' thinking and the outcomes of that thinking. In order to determine accuracy, the video excerpts used for both years' think-aloud exercises were rated by master raters to create master ratings. The two master raters have facilitated many master scoring sessions for LAUSD's TGDC training and implementation. We used the master ratings they created as the true ratings and investigated the accuracy level of administrators' ratings in both years by comparing ratings between the administrators and master raters.

Findings

After the coding was complete, we conducted a series of analyses to understand the patterns of administrators' use of reasoning strategies and scoring approaches in the two different rating contexts: the practice year and the implementation

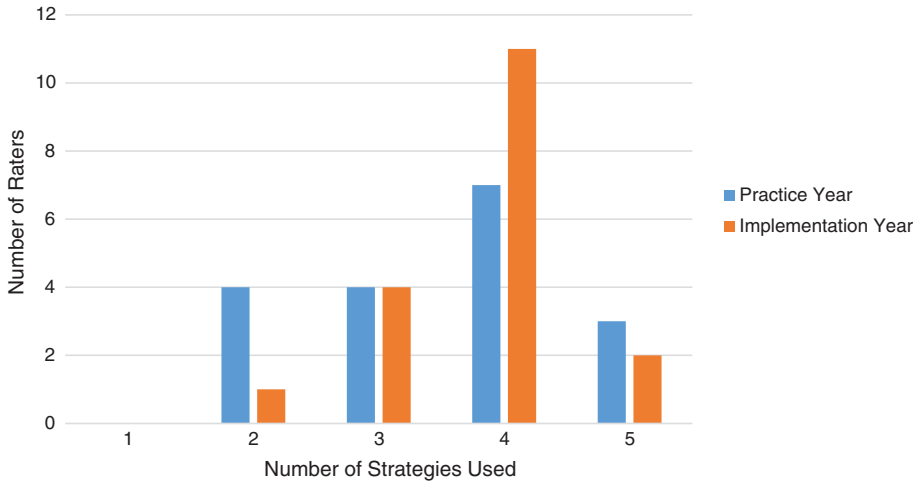

Figure 1 Distribution of administrators' use of the five reasoning strategies.

year. First, we looked at the descriptive data to examine the prevalence and distribution of each strategy used by administrators. Paired-sample *t*-tests were performed to determine differences in how administrators made use of the strategies in these 2 years. Second, using descriptive statistics and paired-sample *t*-tests, we explored administrators' overall scoring approaches used in both years from a macro perspective. Finally, we assessed and compared the accuracy levels of administrators' ratings from the think-aloud exercises from both years.

Prevalence and Distribution of Strategy Use

During the think-alouds, administrators used many strategies. When an administrator used a particular strategy to reason, that strategy was coded once, and each separate instance of use during the interview was counted separately. Each administrator's strategies were coded as described and then aggregated in two ways. For example, if there were 30 discrete instances that an administrator used a strategy to score in a think-aloud, we counted each one separately and summarized by calculating the proportion of the 30 instances that fell into each of the five strategy codes. This aggregation provides a sense of the prevalence of specific reasoning strategies for a given administrator. We also looked across all 30 instances of strategy use to determine whether each of the five types of strategies were ever used, even once. This aggregation provides a sense of the range of strategies an administrator used.

We first looked at the five basic reasoning strategies shown in Table 2 that administrators used: referencing scoring criteria, starting from a score, using information from other classrooms, making suggestions for lesson improvement, and using internal criteria. Figure 1 depicts the distribution of the number of strategies administrators used by year. In both years' think-aloud exercises, all administrators used more than one strategy to determine observation ratings. The vast majority of administrators used between three to five strategies to arrive at their ratings. This pattern was more pronounced in the implementation year, as the percentage of administrators using three or more strategies increased from 77.8% to 94.4% (Figure 1).

We found that there were three major differences in administrators' use of strategy under the two different contexts. First, although referencing scoring criteria was the most frequently used strategy in both contexts, the use of this strategy decreased in the implementation year. As shown in Table 4, in the practice year, across all strategies used by all the administrators for all their ratings, 57% of the strategies used were coded as referencing scoring criteria; however, during the implementation year, just 39% of strategies referenced the scoring criteria.

Second, administrators used internal criteria less frequently when they were closer to training than when they were further away from training. In the practice year, the use of this strategy among all strategies used by administrators was 18%. But in the implementation year, internal criteria accounted for 29% of the strategies administrators used when rating.

Third, in the implementation year, administrators speculated more about lesson improvement when they rated the video excerpt. Results showed that the percentage of the use of this strategy among all strategies used by administrators increased from 12% to 21% across all administrators. In addition, in the practice year, fewer than 60% of the administrators

Table 4 Prevalence of Administrators' Strategies and Criteria Use for Rating

	Percentage of all strategies		Percentage of administrators using strategy	
Strategy	Practice year	Implementation year	Practice year	Implementation year
Referencing scoring criteria	57	39	100	100
Starting from a score	6	5	53	32
Using information from other classrooms	8	7	53	69
Lesson improvement	12	21	58	95
Using internal criteria	18	29	84	90

Note. Percentage of all strategies equals the percentage of the total number of strategies used pooled across all administrators and all strategies used by each administrator across all ratings made by the administrator. Percentage of administrators using strategy equals the percentage of all administrators who reported using a strategy at least one time.

Table 5 Descriptive Statistics Using *t*-test for the Percentages of Individual Administrator's Strategy Use

	Practice year		Implementation year		
Strategy	Mean	*SD*	Mean	*SD*	*t*-Test ($df = 17$)
Referencing scoring criteria	.61	.21	.41	.16	4.40 *
Starting from a score	.05	.06	.05	.10	.08
Using information from other classrooms	.08	.09	.06	.07	.95
Lesson improvement	.11	.11	.21	.11	−.13 *
Using internal criteria	.16	.12	.27	.14	−2.92 *

Note. $N = 18$. Mean is the average across the 18 administrators of the proportion of strategies used by each administrator that are of the specified type. *SD* is the standard deviation of the administrator proportions.
*$p < .05$.

decided their ratings from the perspective of lesson improvement, however, in the implementation year, almost every administrator referenced this notion at least once when scoring (Table 4).

We conducted a series of paired-sample *t*-tests to see if there were differences between the practice year and implementation year when we examined administrators' strategy use.[4] For each administrator, we calculated the proportion of each strategy among the total number of strategies they used when rating. Results from the *t*-tests showed significant differences for referencing scoring criteria, using internal criteria, and lesson improvement over the 2 years (Table 5).

We further explored the specific type of internal criteria administrators drew from when using criteria not clearly specified or supported by the rubric. We found that in the implementation year, personal views of teaching were coded twice as often as in the practice year, whereas professional experience and knowledge were coded about the same amount in both years.

When we examined the extent to which administrators reasoned about ratings from the perspective of engaging in the hypothetical postobservation conversation with the teacher, we found that, although administrators speculated about lesson improvement more often in the implementation year, the number of times they directly approached it from the postobservation conversation point of view was about the same across 2 years.

Scoring Approach

Although reasoning strategies provide information about what criteria administrators use to reason and arrive at a rating, they do not necessarily tell us the general approach administrators take when scoring. In order to understand administrators' overall scoring approach, we coded how the TLF rubric functioned in administrators' reasoning, how administrators analyzed the rubric concepts, and how they made use of the evidence from the video excerpt. Table 6 counts the number of times an administrator used a specific approach when rating each of the three TLF elements. For each category of scoring approach, an administrator's scoring is characterized by only one predominant approach; therefore, for each category the number of codes should sum to 18.

Table 6 Distribution of Uses of General Scoring Approach, by Administrator, by Rating Element

	Element 2B2		Element 3B1		Element 3C1	
Use	Practice year	Implementation year	Practice year	Implementation year	Practice year	Implementation year
Use of tool—use of rubric						
Primarily rely on rubric	16	13	14	11	15	11
Use rubric to justify rating	1	1	3	3	1	1
No explicit use of rubric	1	4	1	4	2	6
Use of tool—analysis of rubric attributes						
Review most rubric attributes	4	3	7	5	9	5
Focus on certain attributes	11	9	11	9	8	6
Did not analyze specific attribute	3	6	0	4	1	7
Use of evidence						
Use evidence to generate ideas for rating	16	13	15	12	15	13
Decide a score and find evidence to support	2	5	3	6	3	5

We found that in both years, the majority of administrators relied primarily on the TLF rubric to guide their reasoning. They frequently read or reviewed the rubric, gauged evidence against the rubric, and decided on a rating. In the implementation year, we found that fewer administrators used this scoring approach. In both years, a few administrators used the TLF rubric mainly to confirm or justify the rating decisions they had already made after watching the video excerpt. There is no difference in the percentage of administrators using this approach across 2 years, and the administrators who used this approach were the same people in both years. Comparatively, this approach was used most often when administrators were rating element 3b1, *quality/purpose of questions*, the one element that administrators reported to be the easiest to score and the element that had been the focus of professional development in the district. There was also a small percentage of administrators who did not explicitly use the rubric when deciding on a rating. They usually discussed the evidence and their reasoning independent of the rubric. In the implementation year, for all three elements, we observed that more administrators used this approach when rating (Table 6). Administrators used multiple approaches within and across the two think-alouds, and the coding reflects this.

The TLF specifies multiple attributes for each element of teaching practice being evaluated. In addition to an overall explanation about an element, the rubric describes the performance level on each attribute. For example, in element 3c1, standards-based projects, activities and assignments, the four rating descriptions (*highly effective, effective, developing, ineffective*) describe the extent to which instructional projects and activities are cognitively engaging and culturally relevant; whether students initiate or adapt activities; and to what extent there is differentiation to support student learning. We examined the degree to which administrators considered all the attributes in the rubric when deciding on a rating. In both years, for elements 2b2 and 3b1, the majority of administrators focused only on certain rubric attributes when assigning a rating, instead of discussing all attributes in the rubric. For example, some might only pay attention to differentiation when judging the teacher's performance for this element. For element 3c1, the number of administrators who reviewed all rubric attributes was about the same as the number of administrators who focused only on certain rubric attributes. In the implementation year, higher percentages of administrators talked about the teachers' overall practice level without analyzing specific attributes: 33% for element 2b2, 22% for element 3b1, and 39% for element 3c1 (Table 6).

How people attend, consider, select, and interpret the evidence they collect or use in teacher performance judgment has been of interest for researchers (Nijveldt, Beijaard, Brekelmans, Wubbels, & Verloop, 2009; Schutz & Moss, 2004). In our sample, we found that most administrators reviewed the evidence (the transcripts of the video excerpt provided to them) to generate ideas about the teacher's performance level and ratings. In the implementation year, however, more administrators developed an initial interpretation or rating decision and then used evidence to support those decisions. They would go to the transcripts to search for evidence that supported the ratings they already had in mind (Table 6).

Another scoring approach we examined concerned the extent to which administrators tried to distinguish the rubric language differences among various performance levels to help them decide on a rating. We coded the number of times

Table 7 Administrators' Accuracy Level Based on True Scores

	Number of administrators matching true scores by element			
Year	2b2	3b1	3c1	Number of administrators who got two or more elements correct
Practice year	8	12	11	12
Implementation year	10	9	13	12

administrators were engaged in analyzing the rubric language differences during the think-aloud exercise. For example, when rating element 3c1, Ron said:

> I think it's mostly ineffective … (*rereading descriptions of "developing"*). There's lack of rigor, true. I am looking at the adjectives: few or some (*referring to "few or no students are cognitively engaged" in the description of "ineffective"* vs. *"some students are cognitively engaged" in the description of "developing"*). And I didn't see any differentiation.

We found that in the implementation year, administrators analyzed the rubric language differences less frequently. In the practice year, the average number of times an administrator engaged in such analysis was just over one. But in the implementation year, the average number of times administrators examined rubric language differences while scoring was only .5. A paired-sample t-test was conducted to examine the differences in the number of times administrators analyzed rubric language in the practice year and the implementation year. Results showed that there was a significant difference in the number of times administrators analyzed rubric language: $t(17) = 2.47$, $p = .024$.

Accuracy

In LAUSD's TGDC implementation, accuracy is defined as the exact agreement of administrators' ratings with master ratings; this rule is used in certification tests. We followed this definition in examining administrators' rating accuracy in our data. In both years' think-aloud exercises, administrators' ratings were compared for exact agreement with master raters' ratings. We found that in general, the overall accuracy levels were similar across the 2 years. The number of administrators who rated two or more elements accurately was the same for both years. There were some variations on the accuracy levels for individual elements. In the implementation year, slightly more administrators scored elements 2b2 and 3c1 accurately, whereas fewer administrators scored element 3b1 accurately (Table 7). In our small sample, there is no evidence suggesting the change of accuracy level in administrators' ratings in one direction or another.

Discussion

Findings from this study indicate that despite extensive training and certification, few administrators use the narrow reasoning strategies taught in the training they received to evaluate teachers. In the practice year context, when administrators worked with a single teacher, and relatively little time had elapsed between the completion of training and certification and when the first round of data were collected, administrators still used a wide variety of reasoning strategies and scoring criteria. This variety was more pronounced in the implementation year context, one in which administrators worked with an average of five teachers and were further from training and certification. Comparing administrators' rating behaviors over the two contexts, we found that during the implementation year, they relied less on the TLF, used more internal criteria, and more often approached scoring with explicit references to helping teachers improve instruction.

Although our findings are based on the small group of administrators we worked with and are subject to the limitations of our methods as discussed above, the findings show a clear pattern of variation in administrators' use of reasoning strategies and rating approaches that is unlikely to be due to random noise. Although such findings cannot suggest any causal explanation for administrators' changes in rating behaviors, there are a few hypotheses for these variations.

The finding that administrators relied on the rubric less and drew from internal criteria more often, especially from their personal views of teaching, may suggest a rater drift effect in administrators' ratings as they were further away from training. During the implementation year, they were less likely to review the rubric thoroughly and use it to guide their

reasoning and more likely to use the TLF and the evidence in a supporting manner to confirm or justify their initial ideas or rating decisions. This is consistent with literature on the use of other teacher evaluation tools, that is, portfolio assessment, showing that assessors demonstrated a tendency to seek confirmation of the initial interpretations they had in mind, and thus they may notice the same evidence differently or ignore certain types of evidence that contradict their thinking while constructing their own pattern of reasoning (Nijveldt et al., 2009; Schutz & Moss, 2004).

In addition, prior research documented a lack of systematic task analysis in teaching performance evaluation in American education. Instead, there is a history of measuring teaching performance based mainly on supervisors' personal choices, due to incomplete theories of teaching practice (Rowan & Raudenbush, 2016). Therefore, idiosyncratic preferences or criteria are likely to be a component in administrators' traditional ways of thinking during teacher evaluation. In the practice year, administrators were close to training and more constrained by the observation tool as they were in the process of learning to use a new protocol in a low-stakes, practice setting. After transitioning to the implementation year, the dynamics of the evaluation context bear a resemblance to those in the prior evaluation settings they worked in and hence might trigger the cognitive processes they used to be engaged in when measuring teaching performance, resulting in an increased use of internal criteria. Such a rating approach deviates from the trained usage of the rubric and could possibly lead to distorted scores on the performance measure.

From an information-processing perspective, there is a second hypothesis to explain the different patterns of strategy and criteria use. Administrators might become more familiar with the TLF rubric and the observation procedures as a result of the larger number of teachers with whom they worked during the implementation year. Perhaps they simply had more practice with the TLF. If this were true, we might not see the same explicit reference to the TLF because there is a greater familiarity. This is consistent with Crisp's (2012) finding that with more experience, raters tend to use the configurational model when making judgments. In addition, as G-studies were used to identify means for improving the reliability of ratings (Shavelson & Dempsey-Atwood, 1976; Shavelson & Webb, 1991), Rowan and Raudenbush (2016) reviewed research findings that showed that evaluators' ability to distinguish observation ratings will improve as more observation days are accumulated. Administrators' improved abilities in using the rubric might reflect an integration of the rubric into their existing knowledge base, creating their own mental models of the rubric or shortcuts to rating. In this analysis, it is plausible that administrators felt that they did not need to deliberately refer to the rubric as much as when they were in the initial learning and practice year.

These two hypotheses—that administrators are drifting away from the discipline of their training or that they are integrating the rubric in a more automatic way—are almost in opposition to one another. We cannot sort them out. Further, it is possible that both are happening at any single time point—perhaps some administrators have integrated the rubric and others have forgotten it. Regardless, we saw no evidence of this affecting administrator's accuracy, although the sample is very small.

As mentioned earlier, in the implementation high-stakes environment, the context of the evaluation activity is more similar to the one administrators were familiar with prior to TLF implementation. Such familiarity could have caused their reasoning to move back toward historical ways of thinking and rating. Our findings showed that administrators were more likely to think from the perspective of lesson improvement in the implementation year. It is possible that the high-stakes context triggers a shift to thinking that is grounded in historical ways of approaching evaluation and the attendant roles the administrator plays in supporting teachers in that high-stakes context. Social science theories argue that under the influences of distortion and risk in performance measurement, objective performance measures such as observation ratings are more valuable when used frequently to provide "rich and specific 'narrative' feedback about their performance along with advice about improvement" (Rowan & Raudenbush, 2016, p. 1206). Administrators' shifted approach of reasoning from the perspective of lesson improvement could be reflecting such a tendency in the organizational context.

Even though we have nominated possible hypotheses to explain the differences across the practice and implementation contexts, it is important to note that there are other potential explanations worthy of articulation and investigation. Perhaps, for example, administrators forgot what they learned during training, and the more distant it became in time, the more they drifted from it. Or perhaps administrators took the pieces most useful to them and carried only those pieces across time. This is possible, given the time constraints of carrying out observations for many teachers during limited time. Contextual factors prescribed by the particular evaluation system, such as proximity to training in time and volume of work, appear to be interacting with administrators' thinking when they assign ratings and

could potentially influence the extent to which the observation measures are subject to risk and distortion in the organization.

Our finding that there was no evidence for any change in the accuracy level may be explained by learning that had already taken place. It is possible that administrators did not rely on the TLF rubric as much during the implementation year because their understandings of the rubric were well integrated into their knowledge base and therefore did not affect the accuracy of their ratings. Meanwhile, it is worth noting that our accuracy measure may not be sensitive enough to capture a clear pattern because we are relying on only three elements where scores vary slightly. Also, in each year a single video was scored by all administrators with different videos for the 2 years. The ability to accurately assess the two different videos could differ and affect the accuracy of scores across the 2 years.

The current study is limited in its ability to sort through these alternative hypotheses. It is also limited, as previously noted, in its size and scope, and its contributions should therefore be understood clearly. It is tempting to want to see these findings as what might happen to administrators' reasoning 1 or 2 years after any large-scale training effort. But there are many features of the practice and implementation year contexts in LAUSD that may have interacted with how administrators reasoned and approached the scoring task, and those features may or may not generalize to other administrators in LAUSD. And the cognitive patterns we see among administrators may only partially map onto the ways in which they use the TLF with their own teachers. Despite these limitations, the value of this study is that it describes a range of ways practicing administrators understand and use an observation protocol in two different contexts. To our knowledge, there is no other work that looks at the same group of administrators in different contexts and tries to understand their use of an observation protocol. As such, the study is an important step forward from which others can build. Further, the study demonstrates the utility of applying research methods traditionally reserved for raters scoring essays to illuminate the cognitive work of administrators—individuals often seen as leaders, not learners doing cognitively challenging work with a new tool over time. By carefully describing the cognitive work administrators do when using teacher evaluation tools, the study nominates a target for further administrator training efforts.

Conclusions and Implications

This article has explored how school administrators' cognitive processes vary while performing a stable task of assigning observation ratings at different phases of implementation of a teaching evaluation system. We analyzed the interactions between administrators' thinking and the contextual factors of the evaluation settings, paying particular attention to the extent to which such interactions could lead to risk and distortion of the observation ratings. We offered a few hypotheses for why administrators' cognitive processes were different in the practice year and implementation year, and how their ratings might be subject to distortion. Future research might explore further how administrators make judgments on performances of teachers they evaluate in their organizations.

Based on our current findings and hypotheses, we propose a few items to be considered in the training of raters and the development of evaluation systems in high-stakes contexts. The decrease in administrators' references to the scoring rubric and the increase of their use of internal criteria during the implementation year suggest that some administrators may benefit from additional training and monitoring to make sure they apply the rubric consistently over time. Our refined coding of internal criteria suggests that the internal criteria administrators used are much more complicated than pure personal preferences or personal opinions. It is often a synthesis of professional knowledge and individual focus on teaching practices that have been developed, possibly based on administrators' previous training, expertise, and experiences. Administrators' learning about the TLF and their perceptions about the social context of rating could also play a role. Although it is hard to disentangle all these factors based on our think-aloud data, simply treating the use of internal criteria as a reasoning strategy not supported by the TLF rubric might not be accurate.

The finding that administrators reasoned from the perspective of lesson improvement much more when they were regularly working with their own teachers is consistent with what has been revealed in previous work: Administrators used the TGDC for both evaluation and improvement purposes, but overwhelmingly focused on supporting instructional improvement (Bell et al., 2016). The move from a low-stakes practice rating context to a high-stakes endeavor of full implementation was accompanied by a shift in reasoning that was more oriented toward observations as a way to help teachers improve. In doing so, the dynamics of the social context and human relationships were increasingly reflected in rater cognition. An administrator's rating decision may be shaped by (perhaps subconscious) predictions about the emotional or motivational effect of a particular rating he or she is about to assign. Administrators and raters are also more

likely to make the effort to provide specific feedback and advice about instructional improvement. It will be beneficial if policy makers and practitioners consider ways to discipline and facilitate such efforts so that ratings are not distorted by the improvement purpose and administrators are equipped with the expertise and skills to produce valid and useful feedback (Rowan & Raudenbush, 2016).

It is not clear how the use of multiple, varied reasoning strategies affects the quality of observation ratings in high-stakes contexts. Accuracy results in this study did not suggest a very clear relationship between strategy use and accuracy level, although in both years, the accuracy levels were modest. Differences in the types and frequency of administrators' reasoning strategy use over the 2 years were evident. There are at least two ways we can understand this. If we conceptualize the observation rubric as a "tool," the "affordance" it has for the rating activity decides that raters will inevitably mediate and develop the way the tool is being used. That is, when administrators learn to understand the rubric concepts and apply them in evaluating teaching practices, they are engaged in a process of synthesizing the new information with their existing knowledge and creating an idiosyncratic understanding of the rubric as well as their own approach to rating. It is important for researchers, trainers, and policymakers to acknowledge and understand the nature of interactions between administrators and the observational tool.

At the same time, the tool "not only opens doors to new experiences but also places important restrictions" on the rating activity (Brown, 2009, p. 20). This restricting function aligns with one of the main goals of training: to discourage administrators from rating based on pure personal preferences or biases and a strict adherence to the way they were trained to rate. Targeted trainings and ongoing support should be provided to regulate administrators' rating practices and assist them in integrating the rubric and their perceptions in a way that does not stray from the intention of the observation protocol, thereby minimizing the threat of distortion. As shown in our results, this is a challenging task, especially when administrators operate in socially dynamic contexts and such contexts change and evolve as dictated by practical, administrative, and policy demands.

Notes

1 There has been a good deal of research in which observation scores are created by paid raters who are not administrators (e.g., Gill, Shoji, Coen, & Place, 2016; Kane & Staiger, 2012; Steinberg & Garrett, 2016).
2 The TGDC system was renamed and modified somewhat in 2015. It is now called the Educator Development and Support: Teachers (EDST).
3 Pseudonyms are used for all participants.
4 The paired sample *t*-tests used are subject to at least two threats. First, strategy use is not normally distributed because the underlying number of strategies being used by any administrator is small and the sample of administrators is small; second, the *t*-tests presume the data have constant variance across administrators, which is not possible given the fact administrators use different numbers of strategies. To assess the robustness of the *t*-test results, we ran multiple permutation tests that do not rely on these assumptions. The results confirm the findings of the *t*-tests and are available upon request.

References

Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performance: Some problems in the design and implementation of evaluation systems. *Educational Researcher, 44*(2), 77–86.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2–9.

Bell, C., Jones, N., Lewis, J., Qi, Y., Kirui, D., Stickler, L., & Liu, S. (2015). *Understanding consequential assessment systems of teaching: Year 2 final report to Los Angeles Unified School District* (Research Memorandum No. RM-15-12). Princeton, NJ: Educational Testing Service.

Bell, C., Jones, N., Lewis, J., Qi, Y., Stickler, L., Liu, S., & McLeod, M. (2016). *Understanding consequential assessment systems of teaching: Year 1 final report to Los Angeles Unified School District* (Research Memorandum No. RM-16-12). Princeton, NJ: Educational Testing Service.

Bell, C., Qi, Y., Croft, A., Leusner, D., McCaffrey, D., Gitomer, D., & Pianta, R. (2014). Improving observational score quality: Challenges in observer thinking. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 50–97). San Francisco, CA: Jossey-Bass.

Bell, C., Qi, Y., Jones, N., Lewis, J., Kirui, D., & Stickler, L. (in press). Inside the black box: Think aloud evidence of administrators' strategies for judging teaching. Educational Assessment.

Berliner, D. C. (2013). Problems with value-added evaluations of teachers? Let me count the ways! *The Teacher Educator, 48*(4), 235–243.

Betebenner, D. W. (2011). *A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories*. Dover, NH: The National Center for the Improvement of Educational Assessment.

Braun, H., Chudowsky, N., & Koenig, J. A. (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: National Academies Press. Retrieved from http://www.nap.edu/openbook.php?record_id=12820&page=1

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational and Behavioral Statistics, 13*(1), 1–18.

Brown, M. W. (2009). The teacher-tool relationship: Theorizing the design and use of curriculum materials. In J. T. Remillard, B.A. Herbel-Eisenmann, & G. M. Lloyd (Eds.), *Mathematics teachers at work: Connecting curriculum materials and classroom instruction*. New York, NY: Routledge.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311–337.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D., Bell, C., Hamre, B. K., & Pianta, R. C. (2013). Effects of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*, 757–783.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*, 163–178.

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Assessment, 31*(3), 10–20. https://doi.org/10.1111/j.1745-3992.2012.00239.x

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

Darling-Hammond, L. (1990). Teacher evaluation in transition: Emerging roles and evolving methods. In J. Millman & L. Darling-Hammond (Eds.) *The new teacher handbook of teacher evaluation: Assessing elementary and secondary school teachers*. Newbury Park, CA: Sage.

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.

Drijvers, P., Doorman, M., Boon, P., Van Gisbergen, S., & Gravemeijer, K. (2007). Tool use in a technology-rich learning arrangement for the concept of function. In D. Pitta-Pantazi & G. Philippou (Eds.), *Proceedings of the V Congress of the European Society for Research in Mathematics Education CERME5* (pp. 1389–1398). Cyprus: Larnaca.

Ellett, C. D., & Teddlie, C. (2003). Teacher evaluation, teacher effectiveness and school effectiveness: Perspectives from the USA. *Journal of Personnel Evaluation in Education, 17*(1), 101–128.

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*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory MidAtlantic. Retrieved from http://ies.ed.gov/ncee/edlabs

Gitomer, D. H., Bell, C. A., Qi, Y., McCaffrey, D. F., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record, 116*(6), 1–32.

Hausman, C. S., Crow, G. M., & Sperry, D. J. (2000). Portrait of the "ideal principal": Context and self. *NASSP Bulletin, 84*(617), 5–14.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56–64.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel* (MET project research paper). Seattle, WA: Bill and Melinda Gates Foundation.

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the Golden State Examination. *Journal of Educational Measurement, 38*(2), 121–145.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation.

Kraft, M. A., & Gilmour, A. F. (2016). Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly, 52*(5), 711–753. https://doi.org/10.1177/0013161X16653445

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting *The Widget Effect*: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Rresearcher, 46*(5), 234–249. https://doi.org/10.3102/0013189X17718797

Lantolf, J. P., & Appel, G. (1994). *Vygotskian approaches to second language research*. Westport, CT: Greenwood Publishing Group.

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399–418.

Lee, C. D., & Smagorinsky, P. (2000). *Vygotskian perspectives on literacy research: Constructing meaning through collaborative inquiry*. Cambridge, England: Cambridge University Press.

Lefkowitz, J. (2000). The role of interpersonal affective regard in supervisory performance ratings: A literature review and proposed causal model. *Journal of Occupational and Organizational Psychology*, *73*, 67–85.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*, 543–560. https://doi.org/10.1177/0265532211406422

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47–67.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement, 46*(4), 371–389.

Nijveldt, M., Beijaard, D., Brekelmans, M., Wubbels, T., & Verloop, N. (2009). Assessors' perceptions of their judgment processes: Successful strategies and threats underlying valid assessment of student teachers. *Studies in Educational Evaluation, 35*(1), 29–36.

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 edition). Washington, DC: American Educational Research Association.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413–428.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144.

Sawchuk, S. (2009). New teacher-evaluation systems face obstacles: Stimulus funds require districts to revamp teacher yardsticks. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2009/12/11/15evaluate.h29.html

Schutz, A., & Moss, P. (2004). Reasonable decisions in portfolio assessment: Evaluating complex evidence of teaching. *Education Policy Analysis Archives,* 12(33). Retrieved from http://epaa.asu.edu/ojs/article/view/188/314

Shavelson, R. J., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research, 46*(4), 553–611.

Shavelson, R. J., & Webb, N. M. (1991*). Generalizability theory: A primer*. Newbury Park, CA: Sage.

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

Steinberg, M. P., & Sartain, L. (2015). Does better observation make better teachers? *Education Next 15*(1), 71.

Strunk, K. O., Weinstein, T. L., & Makkonen, R. (2014). Sorting out the signal: Do multiple measures of teachers' effectiveness provide consistent information to teachers and principals? *Education Policy Analysis Archives, 22*, 100.

Suto, I. (2012). A critical review of some qualitative research methods used to explore rater cognition. *Educational Measurement: Issues and Practice, 31*(3), 21–30. https://doi.org/10.1111/j.1745-3992.2012.00240.x

Verbitsky, A. A., & Kalashnikov, V. G. (2012). Category of "context" and contextual approach in psychology. *Psychology in Russia: State of the Art*, *5*, 117–130. https://doi.org/10.11621/pir.2012.0007

Whitehurst, G., Chingos, M., & Lindquist, K. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts*. Washington, DC: Brown Center on Education Policy at Brookings. Retrieved from http://www.brookings.edu/~/media/research/files/reports/2014/05/13-teacherevaluation/evaluating-teachers-with-classroom-observations.pdf

Appendix

LAUSD Focus Observer Think-Aloud Protocol

Focus Observer Think-Aloud Protocol

Directions for Interviewer:

Please read through all of the directions (the ones below and the ones you will read to the participant) prior to the think-aloud. Text in bold you should read to the participant so that our approach is standardized across interviewers.

Thank you for agreeing to help us understand how observers learn to score. Today, we are going to try to understand how you think about domains 2 and 3 of the Teaching and Learning Framework. In order to do that, we are going to watch a video together with a transcript in front of us, and then listen to you as you assign scores to what you saw.

Some of what we do today will be similar to what you do in your usual observations and some of it will be different. I will describe the whole process to you so you have a big picture view and then I'll answer any questions you have. First, we will watch the short video together. You can follow along on the transcript if you would like. You can take additional notes, but you do not have to. After the video is over, I will ask you to align the evidence (by cutting and pasting from the transcript to the scoring sheet) and then give a score to just three of the focus elements. While you do all of that, I will be

quiet, watch, and listen. I won't interrupt you but I will write down places where I don't understand something you have said. I will wait and ask you my questions at the very end.

I mentioned that we call this a think-aloud protocol. Think-aloud protocols try to understand how someone does something by listening to them talk out loud as they do the thing. So in the parts where I ask you to speak out loud, you just need to say the thoughts that are going through your mind. When we have done these think-alouds with other principals some have said they like it, others have said it feels strange. Whatever your reaction, please do not worry about what you are saying. There are no silly things to say. We know so little about how administrators do the work of observing so it is really helpful to just hear people think.

Would it help if I give you some examples of the kinds of things you might say? [pause for response] When you are aligning evidence you might say something like "I am trying to find the spot where there is evidence of the teacher's questioning." or "I'm trying to decide if I want to use this quotation as evidence for questioning and discussion or for using assessment." When you are assigning scores you might say something like "I don't think the teacher did a very good job asking questions. They were mostly low level."

After you are done scoring, I will ask you some questions about what you said. And then we'll be all done. I think this should take us about an hour. Does all that make sense? [pause for response] Do you have any questions before we get started? [pause for response].

Before we begin, you should know that this is a seventh grade "regular" mathematics classroom and the video starts part way through the homework review. Then you see about 5 minutes of the first activity he does with the students.

Begin video. Video is 10 minutes.

After video is over—Ok, the next step is for you to align and score as you normally would. You will use the electronic copy of the transcript to cut and paste at least one piece of evidence for each of the three elements. If you would like to include more than one piece of evidence you can, but you don't have to. When you decide on your overall score, you should use all the evidence, but you do not need to include all the evidence on the scoring sheet. Again, I will be quiet, watch and take notes as you talk out loud. OK? Questions?

Before you start, sometimes it is hard for me to get everything down; would it be ok if I record you so I can get complete notes? I will delete the recording once I have completed my notes.

The interviewer will have the observer proceed through the steps above. The interviewer should focus on understanding (a) the rationale that the observer uses in any aspect of describing evidence, sorting evidence, assigning and/or justifying scores and (b) the places where the observer is using information that comes from somewhere other than the video to make decisions about scores.

So that we can learn a little bit more about how administrators are thinking, we are asking everyone some clarifying questions about the elements you scored. So I will start with those:

1. For *Expectations for Learning & Achievement* you gave the teaching an xx. Can you tell me a little bit more about why you gave it xx and not an xx − 1 or an xx + 1?
    a. Probe: How did you decide what level of expectations the teacher had?
    b. Probe: How did you decide how many students the teacher had those expectations for?
2. For *Quality/Purpose of Questions* you gave the teaching an xx. Can you tell me a little bit more about why you gave it xx and not an xx − 1 or an xx + 1?
    a. Probe: How did you decide what kinds of questions were being asked?
    b. Probe: How could you tell if the teacher was differentiating the questions for students?
3. For *Standards-Based Projects, Activities & Assignments* you gave the teaching an xx. Can you tell me a little bit more about why you gave it xx and not an xx − 1 or an xx + 1?
    a. Probe: How did you decide whether the activities were rigorous and appropriate for the students?
    b. Probe: How did you decide whether students were cognitively engaged by the activities?
4. [Ask any other clarifying questions you may have]

    Now I'd like to take a step back and get a little perspective on how this teaching and scoring made sense to you.
5. How does what we just did compare with what you have been doing with the teachers at your school this year?
6. Which was the hardest element to score for this video? Why?

7. Which was the easiest element to score for this video? Why?
8. What's your overall assessment of this excerpt of teaching?
    a. (If needed) What makes you say so?
    b. At what point did you arrive at that impression in your mind?
9. Are there any other things I should know about how you think about scoring before we end today?

Ok, that is my last question. Could you please save your scoring sheet and email me a copy? I really appreciate your help. Thank you again. I know you are very busy, so thank you very much for making the time to do this with me.

Suggested citation

Qi, Y., Bell, C. A., Jones, N. D., Lewis, J. M., Witherspoon, M. W., & Redash, A. (2018). *Adminstrators' uses of teacher observation protocol in different rating contexts* (Research Report No. RR-18-18). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12205

Action Editor: Heather Buzick

Reviewers: Teresa Egan and Caroline Wylie