Building a More Complete Understanding of Teacher Evaluation Using Classroom Observations

Julie Cohen[1] and Dan Goldhaber[2]

Improving teacher evaluation is one of the most pressing but also contested areas of educational policy. Value-added measures have received much of the attention in new evaluation systems, but they can only be used to evaluate a fraction of teachers. Classroom observations are almost universally used to assess teachers, yet their statistical properties have received far less empirical scrutiny, in particular in consequential evaluation systems. In this essay, we highlight some conceptual and empirical challenges that are similar across these different measures of teacher quality. Based on a review of empirical research, we argue that we need much more research focused on observations as performance measures. We conclude by sketching out an agenda for future research in this area.

Keywords: accountability; classroom research; educational policy; policy analysis; teacher assessment

Improving teacher evaluation is one of the most pressing and contested contemporary educational policy issues. There is compelling evidence that teachers represent a key leverage point for improving student outcomes in both the short and long term and that teachers vary substantially in their effectiveness (Chetty, Friedman, & Rockoff, 2014; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). Despite this consequential variation, until recently, most formal teacher evaluations have been cursory and not very discriminating (Toch & Rothman, 2008; Weisberg, Sexton, Mulhern, & Keeling, 2009). Most agree with the high-level assertion that teacher evaluation ought to be meaningful, which entails reforming the content and structure of evaluations. Despite this general consensus, there has been a great deal of controversy surrounding the substance of proposed reforms.

In particular, many have questioned the use of value-added measures (VAMs) in newer evaluation systems[1] (cf. Rothstein, 2009). These measures are controversial for a number a reasons. First, they have low face validity among educators who question whether standardized tests represent the broader construct of interest, student learning. Teachers also do not always know how to interpret such measures, nor do they provide information teachers can use to identify specific areas for instructional improvement (Kupermintz, 2003). Second, they create a forced distribution of performance, undercutting norms of collective success and collaboration (Johnson, 2015). Finally, their measurement properties have been scrutinized and found wanting by some given the likelihood of misclassifying an "effective teacher" as "ineffective" or vice versa (Baker et al., 2010; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; McCaffrey, Sass, Lockwood, Mihaly, 2009; Papay, 2012; Rothstein, 2009; Schochet & Chiang, 2013). But, for all the scrutiny, VAMs can only be directly used to assess the performance of a modest fraction of teachers, typically about 20% to 30% (e.g., Watson, Kraemer, & Thorn, 2009).

Classroom observations, on the other hand, are used nearly universally to assess teachers. They have high levels of face validity because they assess teaching practices that teachers themselves can observe. For those striving to become better practitioners, this information can provide timely and actionable formative feedback. Despite these potential benefits, one of the critiques leveled against observations is the precedent of not differentiating among teachers. Observation instruments are criterion-referenced measures, not necessarily leading to a distribution of ratings, and historically most teachers have been deemed effective or highly effective. This lack of differentiation is often referred to as the "widget effect" (Weisberg et al., 2009). The distribution of evaluations using classroom observations does not align with the distribution suggested by value-added measures or by principals' perceptions that teachers vary significantly in effectiveness (Jacob & Lefgren, 2007).

[1]University of Virginia, Charlottesville, VA
[2]University of Washington, Bothell, WA

Educational Researcher, Vol. 45 No. 6, pp. 378–387
DOI: 10.3102/0013189X16659442

Interestingly, given their prevalent use, we know surprisingly little about the statistical properties of classroom observations in consequential personnel decisions. Indeed, much of what we know is derived from extensive research of a large-scale research study—the Measures of Effective Teaching (MET) Project (cf. Kane & Staiger, 2012)—and it is unclear how these findings might translate when evaluation reform is put into practice (Goldhaber, 2015). Will real-world classroom observations differentiate among teachers? Will they be reliable? Will teachers receive actionable feedback, leading them to seek and receive high-quality professional development? The answers to questions like these are key to understanding how the use of observational measures of performance will affect the quality of the teacher workforce.

In this article, we highlight some conceptual and empirical challenges that are similar across different measures of teacher performance. We focus on the existing validity evidence around classroom observations as evaluation measures, much of which is derived from research studies as opposed to real-world evaluation systems. We highlight what we know about the stability of observational measures across raters and educational contexts. We speculate on the implications of the extant literature, and we ask what additional kinds of evidence we would need about observations to feel confident observations could be featured in fair and useful evaluation systems. We attempt to answer some of these questions while describing the conceptual issues that arise when measuring teachers' classroom performance based on observations. Based on a review of empirical research, we argue that we need more research focused on observations as performance measures, particularly from authentic settings where observational performance measures are used in consequential evaluation systems, and we sketch out an agenda for future research. Given space constraints, our goal is not to provide a comprehensive synthesis of the extant literature around classroom observations. Such a review would be helpful but goes beyond the scope of our brief essay, which is instead intended to outline key issues.

We focus on validity issues that we consider particularly salient given what we know about other forms of evaluation. For example, we know that value-added measures tend to fluctuate over time and across student populations, so we highlight corresponding questions about classroom observations. Other issues, such as raters, are particularly important to consider given the inherent subjectivity of observational measures of performance. We use the literature cited in the following to surface broader issues in research on classroom observations that merit further study in service of supporting the development of rigorous and fair teacher evaluation systems.

We include studies where observations were low/no stakes for participating teachers (e.g., the MET project) as well as those in which observations were consequential for teachers (e.g., Steinberg & Sartain, 2015). Our next steps for research includes discussion about how research generated from high- and low-stakes observations of teachers may engender different conclusions.

A Brief History and Summary of Current Use of Classroom Observations

It is ironic to focus on the field's lack of knowledge about the properties of classroom observations given their longstanding history in public schools. Structured classroom observations date back at least to the 1950s, when research began to focus on discrete teacher behaviors and schools began to implement classroom-based observations on a systematic large-scale basis[2] (Brophy & Good, 1986). For the past 60 years, principals have engaged in "walk-throughs" with checklists of readily assessable features of classrooms, though these were not necessarily consequential evaluations. The 1983 publication of *A Nation at Risk* (Gardner, 1983) spurred greater standardization of evaluation systems for personnel decisions (Darling-Hammond, Wise, & Pease, 1983; Hazi & Rucinski, 2009; Toch & Rothman, 2008). While performance evaluations were ostensibly used for high-stakes purposes, evidence suggests few teachers ever received less than stellar evaluations (Weisberg et al., 2009).

Contemporary observation instruments are comprised of scales representing a range of quality—a departure the binary checklists historically employed—and are designed to be more nuanced and rigorous. Reports from a number of states, however, suggest the vast majority of teachers still receive high ratings even after implementing evaluation systems designed to produce more realistic distributions. For example, more than 95% of teachers in Florida, Michigan, and Tennessee were rated "effective" or better in revamped evaluation systems (Anderson, 2013).[3]

The nationally representative Schools and Staffing Survey (SASS) provides some broader insights into how teachers have recently been evaluated. Analysis of principal survey data from the most recent (2011–2012) wave of the SASS shows that more than 95% of teachers were evaluated based on formal classroom observation(s). By contrast, only 40% of principals reported using student test scores to gauge teacher performance.[4] More recently, a report from the National Council on Teacher Quality shows the rapid increase in the number of states requiring measures of student achievement be included in teacher evaluation: from 15 states in 2009 to 43 in 2015 (Dorety & Jacobs, 2015). Classroom observations continue to be an important teacher performance measure, with 48 states requiring some formal observations (Dorety & Jacobs, 2015).

It is not clear how evaluation reform will evolve over time. Many of the recent teacher evaluation initiatives that focused on student test-based measures were spurred on by Race to the Top, along with waivers to No Child Left Behind (NCLB) requirements. The recent reauthorization to the Elementary and Secondary Education Act prohibits this type of federal involvement in states' teacher evaluation systems, and states may correspondingly move away from using student test-based measures to evaluate teachers (Goldhaber & Brown, in press; Sawchuk, 2016). This would then redouble the imperative to learn more about how observational measures work to identify true differences between teachers and drive improvement in teacher practice.

What Do We (and Don't We) Know About Observational Measures of Teaching?

In the following discussion, we highlight some existing evidence about observational measures as both research and evaluation tools, as well as some of the potential sources of error and conceptual challenges inherent in such measures. We ask: To what extent have we built a comprehensive validity argument about

observational measures as reliable assessments of teacher performance? We focus first on what is being measured by different observation instruments. Then we summarize what we know about reliability and the information needed for accurate measurement. Given that teaching is an inherently situated activity, we analyze existing evidence around how contextual features are associated with measures of practice. Finally, we conclude with a discussion of raters and the objectivity of observational measures. We do not focus on the predictive validity of observations because this topic has been well addressed elsewhere.[5]

What Is Being Measured?

There are multiple dimensions by which a teacher might be deemed "effective." Teachers support students in myriad ways in service of multiple outcomes, including social and emotional outcomes (Blazar & Kraft, 2015). Even if we focus on the narrower construct of teacher effectiveness at supporting student learning on academic outcomes, value-added estimates culled from a single test may only represent a portion of the broader construct of interest.

Likewise, any specific observational instrument also measures only a small portion of the broader construct of "teaching quality," a construct around which we do not yet have consensus. Some suggest that instructional quality is distinguished by practices that are responsive to and sustaining of students' diverse cultural backgrounds (Ladson-Billings, 1995; Paris & Alim, 2014). Others conceptualize good teaching by foregrounding the clarity and accuracy of teachers' representation of academic content and their instructional explanations (Hill, Ball, & Schilling, 2008) or the extent to which teachers foster a positive climate in the classroom and promote warm and respectful relationships with students (Pianta & Hamre, 2009). Given these divergent perspectives, it is perhaps unsurprising that different research teams and districts alike have constructed very different measures to assess instructional quality.[6]

Each instrument rests on a theory of instruction, which may or may not be supported by robust empirical evidence. Some, such as the commonly used Framework for Teaching (FFT; Danielson, 2007) and the Classroom Assessment Scoring System (CLASS; Pianta, Hamre, Haynes, Mintz, & LaParo, 2006), feature elements of teaching that theoretically cut across content areas, including classroom management and organization and use of materials. Other instruments used primarily for research studies, such as the Mathematical Quality of Instruction (MQI; Hill, Kapitula, & Umland, 2011) or the Protocol for Language Arts Teaching Observation (PLATO; Grossman, Loeb, Cohen, & Wyckoff, 2013), privilege content-specific aspects of instruction and require raters have subject specific knowledge and/or teaching experience. We still lack empirical clarity about the affordances and constraints of these divergent approaches to measuring instruction, for formative feedback and for consequential summative assessments.

Another component of this work is developing more robust evidence around the quality of enactment of teaching practice (and corresponding score points) necessary to support student learning. Looking across the observation protocols used in the MET study, where observations had no stakes for participating teachers, Polikoff (2015) found substantial fluctuation around theoretically meaningful cut scores. This suggests that not only do we lack clarity about what is quality practice, but we also know little about what constitutes quality demonstration of a practice and how to ensure that observers recognize these potentially meaningful distinctions.

Information Needed for Accurate Measure

Just as VAMs necessitate a certain amount of information to minimize measurement error, so do observation instruments rely on appropriate sampling of classroom practice. Indeed, there is now a significant amount of evidence about the stability of value added across years (Goldhaber & Hansen, 2013; McCaffrey et al., 2009), and the information has led some (e.g., Baker et al., 2010) but not all (Glazerman, Loeb, Goldhaber, Raudenbush, & Whitehurst, 2010) to conclude that the magnitude of temporal fluctuations undermines the potential use of VAMs in personnel decisions. We know comparatively little about the degree to which observational measures are stable from year to year or whether stability is a desirable attribute. Today's observations are likely quite stable given consistently high scores but probably not reflective of actual distributions in teaching quality.

Stability within years would likely be higher than across years because teachers teach different students and often content and courses from one year to the next. These differences are likely to be associated with changes in instructional practices. Thus, to the extent that teacher assignments change from one year to the next, we would expect to see large fluctuations in observation scores across years. We might also expect that new teachers' practices would fluctuate more substantially than veteran teachers' instruction (Atteberry, Loeb, & Wyckoff, 2015), though little empirical work has confirmed these hypotheses.

Historically, studies found year-to-year correlations in teaching practices in the 0.5 to 0.7 range, with a great deal of variability across the scales for different practices (cf. Brophy, Coulter, Crawford, Evertson, & King, 1975). Only a few recent studies have investigated this empirically. Polikoff (2015) and Garrett and Steinberg (2015) find year-to-year correlations ranging from 0.4 to 0.56 for all but one of the observational instruments used in the MET study, with a great deal of heterogeneity across individual scales.[7] In other words, some teaching practices are quite stable over time while others tend to fluctuate substantially.

Not all observational instruments necessitate the same amount of information for accurate measurement. Instruments focused on teacher-student relationships suggest more moderate within-year stability, with cross-lesson correlations in the 0.5 to 0.9 range (Pianta, Mashburn, Downer, Hamre, & Justice, 2008; Smolkowski & Gunn, 2012). Other studies suggest that classroom management is the most stable instructional practice across a range of different protocols (Gitomer et al., 2014; Polikoff, 2015). In contrast, even within a single subject, such as language arts, evidence suggests variation in practice in reading versus writing instruction (Grossman, Cohen, & Brown, 2014). Some features of "good teaching" might be more stable across instruction, but others might naturally fluctuate over time. We might expect productive, orderly classrooms regardless of what was taught, while the quantity of teacher feedback or classroom

discussion might vary substantially depending on lesson content and structure.

A single observation is unlikely to reflect a teacher's broader repertoire of practices, and multiple observations sampled across time and content would likely better assess instructional quality. In practice, districts vary widely in how they sample instruction in terms of content focus, duration, and frequency. The SASS data suggest untenured teachers are observed an average of 3.4 times per year and tenured teachers an average of 2.3 times per year. The standard deviations for untenured and tenured teachers are 3.1 and 2.8 times per year, respectively. For untenured teachers, the mean classroom observation length is an estimated 45.9 minutes with a standard deviation of 22.8 minutes, while for tenured teachers, the mean classroom observation length is 49.7 minutes with a standard deviation of 26.9 minutes. The large standard deviations suggest a great deal of variability in the frequency and duration of observations. State-level information confirms this overall variability: Tennessee mandates a minimum of four formal annual observations and several informal walkthroughs for all teachers, with even more formal observations for teachers with lower scores on prior observations ("Suggested Observation Pacing," n.d.). In contrast, Louisiana requires a single formal annual evaluation coupled with one informal evaluation ("Compass Overview," n.d.).

Unfortunately, we do not have much, if any, empirical evidence based on consequential evaluation systems from which to inform policymakers' decisions on whether four observations provide additional information about teacher performance beyond what three observations might provide. Most instrument developers acknowledge natural fluctuation in teaching practices, and generalizability studies have allowed researchers to determine the stability of measures of practices and identify sources of fluctuation or potential error. In low-stakes research studies, such studies have demonstrated variability by lesson, time within a lesson, time of year, and content of instruction (Cor, 2011; Hill, Charambolous, & Kraft, 2012; Joe, McClellan, & Holtzman, 2014; Pianta & Hamre, 2009). As a result, most research instruments mandate multiple observations of a certain duration of time, separated by several weeks or months to maximize the likelihood of comprehensive and stable estimates of practice (Pianta & Hamre, 2009; Hill et al., 2012; Ho & Kane, 2013). Few districts to our knowledge have structured their observational systems around such parameters, and furthermore, we know little about whether the findings of generalizability studies would be consistent in high-stakes contexts.

Contextual Factors

All measures of teacher performance include sources of error. If this error is systematic, such that a measure does not provide an accurate picture of teachers' true performance but instead systematically advantages or disadvantages specific kinds of teachers (e.g., special education) teaching in specific contexts (e.g., high poverty schools), then we would consider the measure to be biased. Value-added measures have received much attention for the potential of bias. Indeed, despite some evidence to the contrary, one of the primary critiques of evaluation systems that utilize value-added measures is that they are biased in ways that make them inappropriate for consequential personnel decisions (Baker et al., 2010).

In contrast to the empirical focus on the potential bias of value-added measures, there has been comparatively little research focused on the possibility of the same kinds of biases in observational measures. Observational scales, like any other measure, could also over- or underestimate the "true instructional quality" of a classroom, and teachers may be assigned to classrooms in which it is systematically more difficult to engage in high-quality teaching practices (Bell et al., 2015; Whitehurst, Chingos, & Lindquist, 2014).[8] Such differences would not be based on the individual teacher's instructional "skill." So too might less readily measurable group dynamics make teaching more or less challenging.

Part of the challenge is that instructional quality is inherently situated. Good teaching likely varies in response to contextual factors, including school and district leadership, curricula, and collegial support (Little, 2001; McLaughlin & Talbert, 2006). So too might high-quality instruction vary for specific populations of students, such as English learners or special education students. Indeed, prior research has demonstrated that student demographics are associated with both VAMs (Loeb, Soland, & Fox, 2014; Master, Loeb, Whitney, & Wyckoff, 2012) as well as instructional practices (Blazar, Litke, & Barmore, 2016; Grossman et al., 2014).

The limited existing evidence around stability of instructional quality across student subpopulations suggests this is an important next step for research on observational instruments, particularly as they are used in high-stakes contexts. From process-product research (Brophy et al., 1975; Rosenshine, 1970) to more contemporary work (Abedi, Hofstetter, & Lord, 2004; Cohen & Grossman, 2016; Goldenberg, 2008; Grossman et al., 2014; Ladson-Billings, 2009), studies have found teachers' instructional approaches differ depending on the students. Responsive teaching would likely vary depending on a teacher's students, and this variance is at odds with the standardization of quality practice underlying observational instruments.

Recent studies of consequential district evaluations do indicate that student characteristics may be associated with observational ratings, and teachers with students with higher prior achievement receive higher observation ratings (Chaplin, Gill, Thompkins, & Miller, 2014; Steinberg & Garrett, 2016; Whitehurst et al., 2014). Unfortunately, it is difficult to empirically determine whether these differences are related to teachers responding to their students or that teachers with different teaching practices (and potentially of differing quality) are systematically assigned to different kinds of students. Steinberg and Garrett (2016) capitalize on the random assignment of students to teachers in the MET database to demonstrate that the typically nonrandom assignment of teachers to classes of students does bias classroom observation scores.

Studies using the Measures of Effective Teaching data also conclude that scores on several observation protocols are systematically lower in middle grades than in elementary grades (Grossman et al., 2014; Mihaly & McCaffrey, 2014), and Mihaly and McCaffrey (2014) provide evidence that these differences were not attributable to teacher or student characteristics or to raters. This suggests another possible characteristic of students, age or developmental level, which may be associated with differences in instructional practices and scores on

observational measures. Elementary-aged students may be easier to engage in high-quality instruction, just as Whitehurst and colleagues (2014) assert that higher achieving students might be "easier to teach." However, as Mihaly and McCaffrey caution, lower scores may also be attributed to less quantifiable differences between middle school and elementary school teachers, who are often trained and supported differently.

We might hypothesize that some observation instruments might be less prone to bias based on the composition of students. For example, a measure that focuses primarily on teacher-specific practices, such as the quality of a teacher's instructional explanations or the accuracy and richness of teacher feedback, might be less responsive to student demographics than a measure that assesses practices that are inherently interactional (Bell et al., 2012; Gitomer et al., 2014). The CLASS, for example, theoretically focuses on the quality of the student-teacher interactions, while the MQI theoretically focuses more on teacher-specific practices. Relatively little is known about whether more interactive measures of practice are more prone to bias based on the composition of students or the degree to which these measures actually assess teacher-specific or more relational/interactive practices (Whitehurst et al., 2014). Some studies find that measures of classroom climate, teacher-student interactions, and behavior management are particularly sensitive to students' prior academic achievement (Lazarev & Newman, 2015; Steinberg & Garrett, 2016). Instructional skills such as the use of questioning and discussion techniques have been shown to be less sensitive to student characteristics (Lazarev & Newman, 2015).

Features of schools may also be associated with the instructional practices teachers employ. Schools foster different instructional cultures that have been shown to influence the nature of teaching (Bryk, Sebring, Allensworth, Easton, & Luppescu, 2010; Johnson, Kraft, & Papay, 2012; McLaughlin & Talbert, 2006; Sarason, 1996), and school leaders and colleagues can support individual instructional quality. Different school contexts provide teachers with distinct material, curricular, and intellectual resources. Environments in which teachers trust each other can facilitate the diffusion of those resources and promote higher instructional quality across classrooms (Bryk et al., 2010).

We know little about whether these school-level characteristics are associated with differences in observational ratings, though some descriptive evidence from New York City suggests that observation scores are indeed higher and less variable in schools broadly characterized "more functional" (Cohen & Brown, 2016; Cohen & Grossman, 2016). Blazar and colleagues (2016) also find distinct differences in instructional profiles of teachers in different districts. Here too, it is unclear whether these differences are function of school or district supports or whether higher quality teachers select into specific types of schools or districts.

VAMs attempt to control for contextual variables that likely impact student test performance, including ethnic and linguistic backgrounds and prior achievement levels of students, along with readily measurable characteristics of schools or districts (McCaffrey, Lockwood, Louis, & Hamilton, 2004; Rivkin et al., 2005). In contrast, most observation instruments do not by design account for these contextual factors, tacitly suggesting that good teaching looks consistent across classroom, school, and district contexts. It is not clear whether these measures isolate the contributions of teacher-specific features of classroom or whether teaching is so inherently situated that any teaching practice is contingent on a teacher's students and the school or district in which one teaches (Ball & Forzani, 2009; Grossman & McDonald, 2008).

For the purposes of consequential evaluations, Whitehurst and colleagues (2014) suggest controlling for such characteristics similar to VAMs. We know little, however, about the benefits or drawbacks of this deceptively straightforward suggestion. This might prove essential for the purposes of fairness in cross-teacher comparisons in summative evaluation systems, particularly in regards to awards and sanctions. For example, if middle school students really are systematically more difficult to teach, then it is sensible to compare scores on observations only to other middle school teachers in the context of high-stakes evaluations. That said, one potential downside is that this could mask the fact that teachers with different practices are in fact assigned to different kinds of students. In addition, classroom observations also serve the potentially invaluable purpose of providing formative feedback for teachers. If we adjust observation scores for student demographics including grade-level or school characteristics, teachers may miss opportunities to get unstandardized feedback about the quality of their practice vis-à-vis descriptors at a particular score point on a scale (Papay, 2012). Moreover, a less accurate measure of instructional quality could lead to better performance outcomes, depending on how teachers respond to the measure. Statistical adjustments can decrease the face validity of a measure (the extent to which the measure is seen by teachers to reflect truth about performance) and therefore decrease the likelihood that teachers change their practices based on observational ratings. Moreover, as Steinberg and Garrett (2016) note, the problem of biased ratings would not be fully addressed by adjusting observation scores for student demographics and prior achievement because of the nonrandom assignment of teachers to classes on typically unobservable characteristics.

Numerous questions remain, including: How stable is instructional quality across content areas for elementary teachers and within subjects for subject specialists (Grossman et al., 2014; Qi, Bell, & Gitomer, 2014)? How can districts best address student and school characteristics in classroom observations, and to what extent should raters and observational protocols attend to classroom demographics? These issues are related to broader questions about comparison groups and who is a sensible counterfactual in assessing teaching quality in consequential evaluations. VAMs make clear assumptions about comparison groups in terms of impact on student achievement, but we have not figured those same assumptions out for observation instruments.

Raters

One source of error in classroom observations that has no analog in the statistical methods employed in VAMs is the rater. Research has demonstrated that raters struggle to keep multiple dimensions of quality in mind during observations and that content-specific aspects of instruction are especially cognitively demanding and subject to rater biases (Bell et al., 2014; Park, Chen, & Holtzman, 2014). Some evidence suggests that raters are the largest source of error in observational instruments in the context of no-stakes

research studies and that rater variance has accounted for 25% to as much as 70% of the variance in scores on different observation instruments (Casabianca, Lockwood, & McCaffrey, 2013; Curby et al., 2011; Hill et al., 2012; Ho & Kane, 2013).

Casabianca and colleagues (2015) analyzed data from ETS's Understanding Teaching Quality (UTQ) research study and found that despite extensive efforts to train raters, check for scoring quality over time, and provide ongoing feedback on scoring (often termed *calibration*), there was still substantial "drift" or movement away from master scored lessons throughout the two-year study. In fact, they found that variability among raters increased over the duration of the research (Casabianca et al., 2015). Think-aloud interviews with raters from the MET project suggest that even experienced raters struggled to score many lessons and used scoring processes that diverged from those used by "master raters" or instrument developers (Bell et al., 2014).

It is worth noting that these studies of raters were conducted in the context of research in which the observation scores were not consequential for teachers. In many cases, including the MET and the UTQ studies, lessons were videotaped for scoring purposes, allowing for rater calibration and multiple scores from multiple raters. When observations are used for consequential teacher evaluations, live observations are the norm (Casabianca et al., 2015). It is not clear whether these findings about raters would be consistent in high-stakes contexts or when raters such as principals know the teachers they are scoring.

Minimizing "rater effects" requires minimizing the more subjective biases that different observers bring to observations. A good first step is a rigorous system for training raters and ensuring that new raters' scores correspond with those given by "expert" or master raters, often called *certification*. It is equally vital to develop systems for periodic calibration to ensure that raters' scores continue to be consistent with expert scores (Gitomer et al., 2014; Park et al., 2014). Park and colleagues (2014) found that systems developed to minimize rater bias based on the characteristics of the rater, teacher, and classroom were effective—but in the context of video observations when raters were not personally connected to the teachers.

Even rigorous training and calibration systems may not be sufficient in reducing variability in scores, especially when stakes are attached (Casabianca et al., 2013, 2015). In consequential evaluations of teachers, North Carolina observers tended to rate a disproportionate numbers of teachers just above proficiency thresholds (Barrett Crittenden-Fuller & Guthrie, 2015). In Tennessee, where the evaluation system expects high correlations between observation ratings and VAMs, principals tend to lower observational scores to more closely align with VAMs (Poon & Schwartz, 2015). Conversely, in Miami, Grissom and Loeb (2014) demonstrate principals' tendency to "prop up" observation scores of teachers they worry may be subject to sanctions based on low overall evaluations.

These few studies of administrator scoring behavior suggest that the concerns with drift and rater variability established in low-stakes research studies may be exacerbated in real-world evaluations. In addition, these studies suggest there are additional potential constraints on accurate scoring based on fidelity to a rubric and consistency across multiple raters. For example, because administrators have existing relationships with the people they are observing and also multiple, competing demands on their time, they may make different strategic decisions about rating teachers that result in less accurate scores.

New research supports the notion that school culture may impede principals accurately assessing the teachers in their schools. Kraft and Gilmour (2016) find that few teachers are rated below "proficient" (a level where there are consequences associated with the rating), even in states that have recently implemented major changes to their performance evaluation systems to make them more rigorous and accurate. If ratings suggest that all teachers are proficient or better, they either reflect reality about the teacher workforce or the ratings cannot be accurate. Survey evidence from the Kraft and Gilmour study suggests it is the latter, as principals privately suggest that substantially more teachers are "below proficient" than are rated as such.

Bell and colleagues (2016) have done extensive mixed methods research on how principals in Los Angeles develop and use observational ratings. They suggest real-world raters, usually school administrators, do not conceptualize or use observation instruments as typical measurement tools with an emphasis on accurate scoring. Instead, principals factor in various organizational demands when scoring teachers. These include the need for trust and transparency within a school and individualized information about teachers and their students. Principals also tend to emphasize observations as formative rather than summative assessment tools used primarily to set goals for teachers. This more situated view of teacher evaluation emphasizes flexibility and professional discretion rather than standardization and consistency across settings. Depending on the aims of an evaluation system, we might interpret these findings differently. If "fairness" and uniformity are key goals, then principals may not be the best raters of consequential observations as they are always working within organizational contexts with specific constraints and competing goals. If we desire personalized formative feedback and contextualized understanding of teaching aimed at ongoing development, then we need much more research like the work in Los Angeles to determine the strategies that principals use to effectively leverage observations as tools for instructional improvement.

The question of who should be conducting observations has received much media attention given New York's recent mandate that external raters conduct at least one consequential observation (Ravitch, 2015). While principals provide more consistent ratings of teachers, outside observers' scores are more closely related to VAMs (Whitehurst et al., 2014). Indeed, a number of studies suggest multiple raters are needed for reliable estimates of practice (Casabianca et al., 2013; Hill et al., 2012; Kane & Staiger, 2012), and while training multiple raters or hiring outside raters may be sensible from a measurement perspective, it represents a costly and logistically challenging proposition for districts.

Next Steps for Research

Despite the ubiquity of classroom observations, we know relatively little about them, particularly as they are used in current evaluation systems. Observations have not been held under the microscope to the same degree as VAMs, perhaps because of their long history as the status quo measure of quality or because they

do not create a forced distribution of performance. Regardless, a great deal more work is needed to assess how to reliably measure classroom practices and how such measures can translate into meaningful improvements in students' classroom experiences. We need to know more about how we measure quality teaching in service of supporting the development of rigorous and fair teacher evaluation systems (for both improvement and personnel decisions) using classroom observations. Byproducts of such research would likely include a deeper understanding of teaching and, importantly, the improvement of teaching.

First, and most importantly, classroom observations need to be validated based on a range of student measures we care about. There is some evidence on the degree to which they predict student test achievement (e.g., Kane, McCaffrey, Miller, & Staiger, 2013), but the connection between observations and long-term student outcomes has not been established as it has with VAMs (Chetty et al., 2014).

Second, we need more robust evidence on the affordances and constraints to different measures of instruction. Are there particular practices that teachers can more readily improve upon and that are also valuable to student learning, broadly defined? How does the composition of practices in observation instruments influence implementation and effects of teacher evaluation? Some groups such as the New Teacher Project (2012) suggest evaluating teachers on a limited number of "essential components" of a successful lesson. The challenge with implementing such a recommendation is there is neither clarity about those "essential components" nor consensus about the process by which we would determine a practice's essentialness[9] (Cohen, 2015). We need to develop greater conceptual clarity around the features of instructional quality that we want to capture in consequential evaluation systems and make clear for teachers why and how we think such practices are essential in supporting a range of more and less readily assessable student outcomes.

Third, we need to know more about how to sample instruction in consequential evaluations to maximize the likelihood of observing a representative slice of teacher practice. How do different sampling plans translate into differential personnel decisions and teacher development over time? Do observers need to see a particular type of lesson or a range of lesson formats/content? How many observations are needed to get reliable estimates of teacher performance? We know a little about some of these issues,[10] but most of the evidence is new and based on relatively few studies, and much research in this area has been conducted in research studies where evaluations are low or no stakes for participating teachers. We need to learn more about observational evaluation as it is used in practice. Teachers' responses to consequential observations will ultimately dictate the degree to which these measures support improvement in instructional quality.

Fourth, we need to know more about who ought to be observing teachers and how to best train and support them. It is not clear whether school leaders or outside observers with content expertise would better provide teachers with useful feedback. Some districts, such as Washington, D.C. (Dee & Wyckoff, 2015) and Chicago (Sartain, Stoelinga, & Krone 2010; Steinberg & Sartain, 2015), have invested heavily in observer quality and developing cadres of "master raters" and supporting multiple observations of a set of a teacher's lessons. Evidence from Chicago suggests benefits in terms of improvements in teacher performance as a result of observations from these highly trained and supported raters (Steinberg & Sartain, 2015). The field would benefit a great deal from understanding more about the returns on such investments in terms of teacher performance over time as well as challenges faced and lessons learned.

One thing we do know is that it is generally difficult to implement observational performance evaluation systems that actually differentiate among teachers. This raises questions about the theory of action associated with classroom observations as performance measures. If administrators use observations primarily as a formative assessment to help teachers get better, then they might not serve well as summative performance measures. On the other hand, if the differentiation of teacher ratings is necessary to drive improvement and observational measures are a necessary component of a summative evaluation system, then policymakers should guard against the widget effect tendencies inherent in observation-based evaluations of teachers. Developing principal capacity is essential to such efforts.

Little is known about the degree to which, and ways in which, teaching practices evolve in response to observations systems.[11] Understanding the mechanisms through which teachers learn to improve instruction is a crucial step for research on observational measures as evaluation tools. Are there practices that evolve more readily and others that necessitate more intensive support? What evaluation structures best promote development? The extant research tends to separate the observational processes from their potential feedback function. To capitalize on the formative capacity of observational tools, we need not just accurate raters but also skilled feedback providers.

We are optimistic about the potential value of observational instruments as levers for improving instruction in American classrooms, but many important questions remain about how to isolate, measure, and promote quality teaching through observations (Goldring et al., 2015). We believe it necessary to answer the questions outlined previously to have classroom observations lead to meaningful improvements in practice at scale.

NOTES

[1]We use the term *value added* loosely in referring to all student test growth measures of teacher performance.

[2]As early as the 1920s, districts evaluated teachers for differential compensation based on personal traits such as "moral standing in the community" (Toch & Rothman, 2008). By the 1940s, researchers began developing frameworks for data-based, objective teacher evaluation (Ellett & Teddlie, 2003), and the federal government actively encouraged a focus on classroom practice in 1968 with the funding of competency-based teacher education programs (Semmel, 1976).

[3]A notable exception to this is IMPACT in Washington, DC (Dee & Wyckoff, 2015).

[4]It is not clear from the nature of the question that principals were asked how they use achievement measures; we speculate that the percentage of teachers who are being formally evaluated based on value-added measures of student test growth is far smaller.

[5]Observational measures consistently have low to moderate correlations with measures of student achievement (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Hill, Kapitula, & Umland, 2011; Kane, McCaffrey, Miller, & Staiger, 2013).

[6]An observational measure may capture more about instructional quality than was intended by instrument developers if omitted practices are in fact correlated with featured practices.

[7]Polikoff (2015) finds greater stability in the tails of the distribution. Garrett and Steinberg (2015) find somewhat higher year-to-year correlations in language arts instruction than in mathematics instruction.

[8]Administrators also rate their own teachers higher than outside raters, and their scores are more prone to personal biases and prior impressions (Whitehurst, Chingos, & Lindquist, 2014). Gitomer and colleagues (2014) suggest that high and undifferentiated principal ratings of teachers may not simply be a matter of administrative will but a result of administrators' knowledge (or lack thereof) of specific instructional content and/or grade levels.

[9]Districts vary substantially in the number of scales they use from a single observation instrument, Danielson's (2007) Framework for Teaching (FFT), reflecting variation in the aspects of instruction they most prioritize for consequential evaluations. New York City's system, Advance, highlights only 8 of the FFT practices, while Cincinnati's Teacher Evaluation System draws on 15 of the FFT practices.

[10]Especially, for instance, on the reliability benefits on increasing the number of observations, see Kane, McCaffrey, Miller, and Staiger (2013).

[11]Though there is some evidence that formal evaluations and feedback about teaching practices can increase teacher performance in ways that are detectable from student test scores (e.g., Taylor & Tyler, 2012).

REFERENCES

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, *74*(1), 1–28.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, *333*(6045), 1034–1037.

Anderson, J. (2013, March 30). Curious grade for teachers: Nearly all pass. *The New York Times*, A1.

Atteberry, A., Loeb, S., & Wyckoff, J. (2015). Do first impressions matter? Predicting early career teacher effectiveness. *AERA Open*, *1*(4), 1–23.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., . . . Shepard, L.A. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: Economic Policy Institute.

Ball, D. L., & Forzani, F. (2009). The work of teaching and the challenge for teacher education. *The Journal of Teacher Education*, *60*(5), 497–511.

Barret, N., Crittenden-Fuller, S., & Guthrie, J. E. (2015). *Subjective ratings of teachers: Implications for strategic and high-stakes decisions*. Paper presented at the annual meeting of the Association of Education Finance and Policy, Washington, DC.

Bell, C. A., Gitomer, D. A., McCaffrey, D., Hamre, B., Pianta, R., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2–3), 62–87.

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., Qi, Y., Lewis, J., Witherspoon, M., Redash, A., & Kirui, D. (2016). *Administrators' roles in "valid" observation scores: Moving beyond a narrow measurement perspective*. Paper presented at the annual meeting of the Association of Education Finance and Policy, Denver, CO.

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

Blazar, D., & Kraft, M. A. (2015). *Teacher and teaching effects on students' academic behaviors and mindsets* (Working Paper 41). Cambridge, MA: Mathematica Policy Research. Retrieved from http://www.mathematica-mpr.com/our-publications-and-findings/publications/teacher-and-teaching-effects-on-students-academic-behaviors-and-mindsets

Blazar, D., Litke, E., & Barmore, J. (2016). What does it mean to be ranked a "high" or "low" value-added teacher? Observing differences in instructional quality across districts. *American Educational Research Journal*, *53*(2), 324–359.

Brophy, J. E., Coulter, C. L., Crawford, W. J., Evertson, C. M., & King, C. E. (1975). Classroom observation scales: Stability across time and context and relationships with student learning gains. *Journal of Educational Psychology*, *67*(6), 873–881.

Brophy, J. E., & Good, T. L. (1986). Teacher behavior and student achievement. In M. E. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York, NY: Macmillan.

Bryk, A. S., Sebring, P. B., Allensworth, E., Easton, J. Q., & Luppescu, S. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago, IL: University of Chicago Press.

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. 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.

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh public schools*. Rockville, MD: Regional Educational Laboratory Mid-Atlantic.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–2679.

Cohen, J. (2015). The challenge of identifying high-leverage practices. *Teachers College Record*, *117*(8), 1–41.

Cohen, J., & Brown, M. (2016). Teaching quality across school settings. *The New Educator*, *12*(2), 1–30.

Cohen, J., & Grossman, P. (2016). Respecting complexity in measures of teaching: Keeping schools and students in focus, *Teaching and Teacher Education*, *55*, 308–317.

Compass overview. (n.d.). Retrieved from http://www.nctq.org/docs/2015-2016-Compass-Overview-in-Jefferson-Parish.pdf

Cor, K. (2011). *The measurement properties of the PLATO rubric*. Paper presented at the American Educational Research Association annual meeting, New Orleans, LA.

Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J., . . . Pianta, R. C. (2011). Within-day variability in the quality of classroom interactions during third and fifth grade. *The Elementary School Journal*, *112*(1), 16–37.

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

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, *93*(6), 8–15.

Darling-Hammond, L., 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.

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

Dorety, K. M., & Jacobs, S. (2015). *State of the states 2015: Evaluating teaching, leading and learning*. Washington, DC: National Council on Teacher Quality.

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

Gardner, D. P. (1983). *A nation at risk*. Washington, DC: The National Commission on Excellence in Education, U.S. Department of Education.

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.

Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., 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), 873–881.

Glazerman, S., Loeb, S., Goldhaber, D. D., Raudenbush, S., & Whitehurst, G. J. (2010). *Evaluating teachers: The important role of value-added* (Vol. 201). Washington, DC: Brown Center on Education Policy at Brookings.

Goldenberg, C. (2008). Teaching English language learners: What the research does and does not say. *American Educator*, 8–44.

Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher*, *44*(2), 87–95.

Goldhaber, D., & Brown, N. (in press) Teacher policy under the ESEA and the HEA: A convergent trajectory with an unclear future. In C. P. Loss & P. J. McGuinn (Eds.), *The convergence of K–12 and higher education: Policies and programs in a changing era*. Cambridge, MA: Harvard Education Press.

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica*, *80*(319), 589–612.

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.

Grissom, J. A., & Loeb, S. (2014). *Assessing principals' assessments: Subjective evaluations of teacher effectiveness in low- and high-stakes environments*. Paper presented at Association for Education Finance and Policy annual meeting, San Antonio, TX.

Grossman, P., Cohen, J., & Brown, L. (2014). Understanding instructional quality in English Language Arts: Variations in the relationship between PLATO and value-added by content and context. In K. Kerr, R. Pianta, & T. Kane (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 303–331). San Francisco, CA: Jossey-Bass.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education*, *119*(3), 445–470.

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.

Hazi, H. M., & Rucinski, D. (2009). Teacher evaluation as a policy target for improved student learning: A fifty-state review of statute and regulatory action since NCLB. *Education Policy Analysis Archives*, *17*(5), 1–22.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education*, *39*, 372–400.

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.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794–831.

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

Jacob, B.A., & Lefgren, L. 2007. What do parents value in education? An empirical investigation of parents' revealed preferences for teachers. *Quarterly Journal of Economics*, *122*, 1603–1637.

Joe, J., McClellan, C., & Holtzman, S. (2014). Scoring design decisions: Reliability and the length and focus of classroom observations. In K. Kerr, R. Pianta, & T. Kane (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 415–443). San Francisco, CA: Jossey-Bass.

Johnson, S.M. (2015). Will VAMs reinforce the walls of the egg-crate school? *Educational Researcher*, *44*(2), 117–126.

Johnson, S. M., Kraft, M. A., & Papay, J. P. (2012). How context matters in high-need schools: The effects of teachers' working conditions on their professional satisfaction and their students' achievement. *Teachers College Record*, *114*(10), 1–39.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment* (Measures of Effective Teaching report). Seattle, WA: Bill & Melinda Gates Foundation.

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

Kraft, M. A., & Gilmour, A. F. (2016). *Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness* (Working paper). Providence, RI: Brown University.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation And Policy Analysis*, *25*(3), 287–298.

Ladson-Billings, G. (1995). But that's just good teaching! The case for culturally relevant pedagogy. *Theory Into Practice*, *34*(3), 159–165.

Ladson-Billings, G. (2009). *The dreamkeepers: Successful teachers of African American children*. San Francisco, CA: John Wiley & Sons.

Lazarev, V., & Newman, D. (2015). *How teacher evaluation is affected by class characteristics: Are observations biased?* Paper presented at Association for Education Finance and Policy annual meeting, San Antonio, TX.

Little, J. W. (2001). Professional development in pursuit of school reform. In A. Lieberman & L. Miller (Eds.), *Teachers caught in the action: Professional development that matters* (pp. 28–44). New York, NY: Teachers College Press.

Loeb, S., Soland, J., & Fox, L. (2014). Is a good teacher a good teacher for all? Comparing value-added of teachers with their English learners and non-English learners. *Educational Evaluation and Policy Analysis*, *36*(4), 399–416.

Master, B., Loeb, S., Whitney, C., & Wyckoff, J. (2012). *Different skills: Identifying differentially effective teachers of English language learners* (Working Paper No. 68). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

McCaffrey, D. F., Lockwood, J. R., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67–101.

McCaffrey, D., Sass, T., Lockwood, J., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, *4*(4), 572–606.

McLaughlin, M. W., & Talbert, J. E. (2006). *Building school-based teacher learning communities: Professional strategies to improve student achievement*. New York, NY: Teachers College Press.

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) San Francisco, CA: Jossey-Bass.

The New Teacher Project. (2012). *Teacher evaluation systems comparative overview*. Retrieved from http://tntp.org/assets/tools/TNTP_Teacher+Evaluation+System+Comparative+Overview_TSLT+3.12.pdf

Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123–141.

Paris, D., & Alim, H. S. (2014). What are we seeking to sustain through culturally sustaining pedagogy? A loving critique forward. *Harvard Educational Review*, *84*(1), 85–100.

Park, Y. S., Chen, J., & Holtzman, S. (2014). Evaluating efforts to minimize rater bias in scoring classroom observations. In K. Kerr, R. Pianta, & T. Kane (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 383–414). San Francisco, CA: Jossey-Bass.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119.

Pianta, R. C., Hamre, B. K., Haynes, N. J, Mintz, S., & La Paro, K. M. (2006). *CLASS Classroom Assessment Scoring System: Manual Middle Secondary Version Pilot, June 2006*. Charlottesville, VA: Teachstone.

Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher-child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly*, *23*(4), 431–451.

Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education*, *121*(2), 183–212.

Poon, A., & Schwartz, N. (2015). *Improving feedback in teacher evaluations: An evaluation of Tennessee's TEAM coach initiative*. Paper presented at Association for Education Finance and Policy annual meeting, Washington DC.

Qi, Y., Bell, C., & Gitomer, D. (2014). *The role of topic and activity structure in teacher observation scores*. Paper presented at the annual conference of the American Educational Research Association, Philadelphia, PA.

Ravitch, D. (2015, April 1). *Here is the New York State teacher evaluation bill*. Retrieved from http://dianeravitch.net/2015/04/01/here-is-the-new-york-state-teacher-evaluation-bill/

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247–252.

Rosenshine, B. (1970). Evaluation of classroom instruction. *Review of Educational Research*, *40*(2), 279–300.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537–571.

Sarason, S. B. (1996). *Revisiting "The culture of school and the problem of change."* New York, NY: Teachers College Press.

Sartain, L., Stoelinga, S. R., & Krone, E. (2010). *Rethinking teacher evaluation: Findings from the first year of the Excellence in Teaching Project in Chicago Public Schools*. Chicago, IL: University of Chicago Consortium on Chicago School Research Brief.

Sawchuk, S. (2016). ESEA loosens reins on teacher evaluations, qualifications. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2016/01/06/essa-loosens-reins-on-teacher-evaluations-qualifications.html?cmp=eml-enl-eu-news1.

Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics*, *38*(2), 142–171.

Semmel, M. I. (1976). *Competency-based teacher education in special education: A review of research and training programs*. Bloomington, IN: Center for Innovation in Teaching the Handicapped.

Smolkowski, K., & Gunn, B. (2012). Reliability and validity of the Classroom Observations of Student-Teacher Interactions (COSTI) for kindergarten reading instruction. *Early Childhood Research Quarterly*, *27*(2), 316–328.

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.

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy*,*10*(4), 535–572.

Suggested observation pacing. (n.d.). Retrieved from http://team-tn.org/wp-content/uploads/2013/08/Suggested-Observation-Pacing.pdf

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *The American Economic Review*, *102*, 3628–3651.

Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education*. Washington, DC: Education Sector.

Watson, J. G., Kraemer, S. B., & Thorn, C. A. (2009). *The other 69 percent*. Washington, DC: Center for Educator Compensation Reform at the U.S. Department of Education, Office of Elementary and Secondary Education.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Chicago, IL: The New Teacher Project.

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts*. Washington, DC: Brown Center on Education Policy and Brookings Institute.

AUTHORS

JULIE COHEN, PhD, is an assistant professor of curriculum, instruction, and special education at the Curry School of Education at the University of Virginia, 405 Emmet St. S, Charlottesville, VA 22904; *jjc7f@virginia.edu*. Her research interests include the conceptualization and measurement of teaching quality, the ways in which accountability and teacher evaluation systems shape teaching practice, and the development of teachers' use of effective instructional practices in preservice teacher education and professional development.

DAN GOLDHABER, PhD, is a vice president at the American Institutes for Research, 1000 Thomas Jefferson Street, NW, Washington, DC 20007; *dgoldhaber@air.org*. His research focuses on teacher quality and teacher labor markets.

Manuscript received July 27, 2015
Revisions received January 29, 2016, and May 12, 2016
Accepted June 20, 2016