In 2009, The New Teacher Project published a report called The Widget Effect, which highlighted the lack of meaningful evaluation practices within public education by pointing out how few teachers ever received anything lower than the highest available rating and how many were not rated at all. In the wake of public outrage over the newly exposed lack of feedback and supervision in teacher evaluation nationwide, federal policies were introduced with the explicit goal of differentiating teacher effectiveness, rather than just supporting it (through legislation aimed at requiring certifications or investments in professional development) or incentivizing it (through teacher performance incentives). A shift in the orientation of questions for teacher evaluation avoids the foibles of illogical inference and may precipitate a shift in conceptions of teaching.

Beginning with Race to the Top competition criteria and continuing with criteria for No Child Left Behind waivers, federal policies began to require a new generation of teacher evaluation systems. Even states that did not seek these funds or waivers had to respond to the trends and assumptions they produced. These include teacher evaluation systems that can be used to differentiate teachers in order to inform human capital decisions (hiring, firing, promotion, reward, and tenure), and that are based on multiple measures of teaching effectiveness. Measures typically include a rubric-driven classroom observation system; a survey of students, parents, and/or peers; and some measure(s) of student growth and/or achievement. The fundamental question that drives current teacher evaluation policies is whether a teacher is effective or not (Gabriel & Lester, 2013). Though the answer to this question may fall along a spectrum of proficiency or development, it is primarily aimed at sorting and labeling teacher performance.

For preservice teachers, Wei and Pecheone (2010) argue that “recent changes in national and state accreditation processes have put program outcomes under the microscope, and the policy environment increasingly demands that teacher education programs provide evidence that their graduates have learned to teach” (p. 69). Thus, the fundamental question at the heart of performance assessments for preservice teachers is whether a teacher is well prepared. In an effort to provide more authentic evidence that may also have diagnostic and formative value, teacher preparation programs are increasingly using performance assessments that involve observation, as well as artifacts of planning and reflection. Some states and schools of education also follow their graduates using large longitudinal data systems in order to trace the teacher’s classroom performance back to their preparation program (e.g., Burns & Noell, 2006; Darling-Hammond, Newton, & Wei, 2012).

New generation teacher evaluation policies for both pre- and inservice teachers can be characterized by three important shifts in conceptions of teacher effectiveness as they relate to the questions listed in the first paragraph. First, rather than assuming teacher effectiveness is a personal, fixed trait that can be measured once or only early in a career, it is now measured almost annually, with the idea that proficiency may develop and/or change.
over time. This contrasts with evaluation policies involving preservice graduation requirements based on a passing score on a summative exam or model lesson rather than infrequent or only initial observation of inservice teachers.

Second, a small part of an overall rating now often involves a collective measure (e.g., performance on a teamwide or schoolwide goal) to acknowledge the likelihood and desirability of collaboration. Still, the largest percentage of teacher evaluations in pre- and inservice settings measure teacher effects on student achievement as a feature of an individual teacher. This persists in the face of strong evidence that teaching in US schools is a collective activity (Croninger & Valli, 2009; Yuan, 2015) in which a range of adults, both within and outside of school, are responsible for contributing to teacher-level effects on student performance as measured in various ways over time. Applying a teacher effect score (assigned to an individual teacher) rather than a classroom-level score (assigned to an entire class that is associated with a group of educators) is related to the desire to use the assessment of teaching performance in human capital decisions.

Third, new generation teacher evaluation systems take a multiple-measures approach to the assessment of teacher effectiveness, acknowledging the complexity and contingency of effectiveness across settings, but often failing to measure effectiveness in ways that guide or inspire development. Multiple measures include both multiple data points within a measure (e.g., 2–4 observations per year) and a variety of different types of measures, such as observation, student test scores, surveys, and artifacts of planning and reflection.

The multiple-measures approach is a response to growing awareness of the imperfect nature of any single measure of performance. Not only are student test scores and classroom observations susceptible to bias and measurement error, they are time consuming and expensive, especially when repeated with the frequency required to minimize such error (Measures of Effective Teaching Project, 2012; Sartain, Stoelinga, & Brown, 2011). The Pollyanna logic of educational economist Thomas Kane is that each measure has a “superpower” that must be combined with others (Kane, 2012), as if the combination of imperfect measures might accumulate into a more perfect measure. In fact, tens of millions of dollars have been invested in fine-tuning that combination of measures without regard for the fruitfulness of any of these as tools for teacher development (Gabriel & Allington, 2012). Fine-tuning sets of measures so that they align with value-added ratings or trends in student achievement on standardized test scores may do more to limit and disguise what counts as effective teaching because of the loose relationship between and the enactment of research-based instructional practices (Polikoff & Porter, 2014). Currently, the primary goal of measuring teacher effectiveness is to sort and label teachers, rather than nurture and develop teacher effectiveness.

Further evidence for the sort-and-label purpose of current evaluation efforts includes the use of student outcome data in generating an overall rating. Advances in statistical measurement, like hierarchical linear modeling, Structural Equation Modeling, and Value-Added Measurement (VAM), have allowed us to isolate classroom-level effects on student achievement with greater precision and certainty than ever before. However, media representations of the slick sophistication of such tools have awarded them more power in public policy than their unenviable record for reliability and validity in research settings has earned (Amrein-Beardsley, 2014; Gabriel & Lester, 2013; Shavelson, et al., 2010).

In practice, the increased sophistication and complexity of evaluation tools have not led to more accurate or fruitful evaluations. Roughly the same number of teachers are rated proficient under new evaluation systems as they were before 2009 (Heyburn & Goe, 2013; Morgan, Hodge, Trepinski, & Anderson, 2014; Rund & Wiedenbener, 2014;
As Polikoff and Porter (2014) demonstrated, there are weak associations between alignment with standards and measures of pedagogical quality with value-added ratings. Over the last four years, other researchers have demonstrated only small or inconsistent links between observation ratings and other measures of effectiveness (Kane et al., 2010; Strunk, Weinstein, & Makkonen, 2014; Whitehurst, Chingos, & Lindquist, 2011). Even though all state policies related to teacher evaluation require a link to professional support and development, the mechanisms for such efforts are unspecified and unfunded (Gabriel & Paulus, 2014). Polikoff and Porter (2014) note that there is limited evidence that evaluation practices will consistently or significantly drive improvement without a tight link to instruction.

In the section that follows, I describe one major effort to increase the reliability and validity of systems for evaluating teacher performance. I argue, however, that current efforts to increase reliability and validity fall short of the goal of improving instruction because they are aimed at sorting and labeling rather than nurturing and developing teacher effectiveness. Using indicators of classroom discourse as an example, I suggest that a nurture-and-develop approach to teacher evaluation is possible, but not likely, within current performance assessment systems.

The “Measures of Effective Teaching” Project

New generation evaluation policies often cite the Measures of Effective Teaching Project (MET Project) as the source for current assumptions about using multiple measures of teacher effectiveness for high-stakes, human capital decisions. The MET Project, funded by the Bill & Melinda Gates Foundation, was a two-year, 45-million-dollar investigation of a range of measures aimed at identifying effective teachers. Multiple streams of evidence for more than 3,000 individual teachers were collected from six geographically distant urban school districts across the US and correlated with each other in order to identify which were most strongly associated with a teacher’s VAMs—that is, Value-Added Measures the amount of value a teacher adds to a student’s averaged historical performance on standardized tests. It would be more accurate to advertise the MET Project as the “measures of teaching that predict value-added scores project,” as VAMs were the gold standard measure of effectiveness by which all other measures were validated. However, in the absence of a fully articulated, nationally agreed-upon definition of effectiveness in teaching, student test scores continue to serve as the imperfect proxy of effectiveness.

Findings from the MET study include suggestions about everything from the kinds of measures that should be included in high-stakes teacher evaluations (e.g., observations, student surveys, student test scores), to the number and length of observations that are optimal for ensuring reliability. Though the final reports of the MET study were released after the first wave of new state policies, its initial findings are frequently referenced in rationales for current systems. In addition, the database of videos, ratings, scores, and surveys for each of the 3,000 teachers in the MET study was made available to researchers as a longitudinal database (the METLDB) for the study of effectiveness in teaching.

As part of a larger study of contrasting tools for evaluating literacy instruction in the METLDB (Gabriel, Woulfin, & LeChasseur, 2014), our team of researchers selected two videos from each of the top ten highest- and lowest-scoring 4th-grade ELA teachers on both VAMs and the Framework for Teaching rubric (Danielson, 1996). Blind to the MET Project’s ratings of these videos, we coded each for features of effective instruction drawn from a review of research on effective ELA instruction and then transcribed them in order to code for features of discourse that have been associated with increased engagement and achievement in ELA classrooms.

Indicators of classroom discourse are especially important to examine because of their high correlation with measures of student growth and

Language Arts, Volume 93, Number 2, November 2015
achievement (Grossman, Loeb, Cohen, & Wychoff, 2013; Darling-Hammond, Newton, & Wei, 2012) and the frequency with which they are measured across varied observation systems. I could not identify a single commercially available rubric for instruction that did not include some indicator of classroom discourse (sometimes included as questioning, discussion, or academic language; see Table 1 for examples).

In the section that follows, I describe some of the challenges in assessments of teacher performance using examples from this reanalysis of MET Project videos, focusing specifically on the limitations of current tools and how they relate to reigning conceptions of effectiveness and purposes for evaluation.

Current Challenges in the Assessment of Teacher Performance

The major challenge of performance assessment via observation is that indicators are counted as if their presence or absence indicates quality. For example, one feature of classroom discourse that is often included in commercially available rubrics for observations is the use of open-ended and/or higher-order questions. Though the presence of higher-order questions, or more higher- than lower-order questions, has been associated with increased engagement and achievement (e.g., Nystrand, Wu, Gamoran, Zeiser, & Long, 2003), its absence does not indicate lack of quality. Moreover, there is no magic number or ratio of higher-order questions that should be asked. Just as it would be concerning if higher-order questions were never asked, it would also be concerning if higher-order questions were always asked. More is not necessarily better.

Rating instruction based on a set of indicators, even when rating along a rubric rather than on a checklist, turns indicators into individual proxies for effectiveness that allow a yes/no (or partially yes/partially no) designation for each indicator. This allows evaluators to decide whether a teacher is effective (or somewhat effective) or not based on the observed presence/absence (or relative presence/absence) of an indicator. The same either/or logic is present in discourses defending the need for VAMs and in reigning conceptions of effectiveness that aim to identify and remove “the ineffectives” (Gabriel & Lester, 2013). This either/or conception of effectiveness is the foundation for using teacher evaluations in human capital decisions, but it may actually be inappropriate for measuring what we value most about an individual teachers’ role in teaching and learning.

It could be that every observable feature or “best practice” involves a compromise and thus cannot be viewed in isolation as evidence of effectiveness or not. For example, calling on an equal number of boys and girls may extend the length of discussion and limit time for independent practice. Similarly, pursuing a back-and-forth discussion to support a student’s understanding might limit other students’ participation. A teacher could invest in one indicator of effectiveness at the expense of another. Thus, effective teaching may be about managing the dynamic balance of certain features of instruction rather than simply displaying such features. Furthermore,
discrete, observable indicators of effectiveness (e.g., student work is posted on the walls) may not accurately represent the value of ongoing instructional interactions.

When analyzing MET project videos, we found higher-order questions in low-performing classrooms on every measure of the MET study, and high-scoring classrooms that had no evidence of higher-order questions. Similarly, like Polikoff and Porter (2014), we found high-scoring classrooms with few features of effective instruction, and low-scoring classrooms with similar frequencies of features like student choice, opportunities to talk, and opportunities to read. When it comes to opportunities to develop literacy, it isn’t the fact of allotted time for independent reading or writing, but rather the nature and use of that time that determines its value as a practice.

For example, several videos of MET project classrooms included time spent writing for five minutes or more, but the writing tasks often involved filling in blanks of a formulaic paragraph structure or copying notes from the board into a graphic organizer. Neither of these tasks involves a robust opportunity to develop literacy because students are not generating original language, employing a writing strategy, writing for a purpose, or writing to an audience. However, in observation, especially brief observation, it may appear that students are all engaged in writing, and this instrumental engagement may be viewed as evidence of effectiveness because students are quietly complying with a writing-based activity.

Likewise, it is not the presence of a higher-order question that indicates effectiveness or not, just like it is not the fact of sustained silent reading time that indicates robust opportunities to develop literacy. When analyzing the discourse of classroom interactions, the question should not be whether higher-order questions appear in this lesson, but how they were used. In some cases, higher-order questions are asked, but go unanswered. In others, they are asked, but answered incorrectly without opportunities to build knowledge in the teacher’s response. A wrong, unclear, or incomplete answer followed by a responsive instructional move might be more of an indication of teacher effectiveness than right answers that show students already understand. Higher-order questions may also be answered by a single individual (or the teachers themselves) and thus do not indicate the student engagement their presence implies. Finally, they may be asked but irrelevant, inappropriately challenging, or badly timed.

Questioning is undoubtedly important, but focusing on certain kinds of questions, rather than the results of all kinds of questions, may unnecessarily limit conceptions of effectiveness and minimize other possible tools and approaches. The opposite is also true: eliminating undesirable practices, rather than asking what such practices are being used to do, may similarly compromise conceptions of effectiveness. For example, the presence of an Initiate–Respond–Evaluate (IRE) pattern has been associated with lower-order questioning and unbalanced power dynamics. It is thus sometimes associated with decreased engagement, rigor, and achievement (Nystrand, 1997). Yet, we found evidence of this pattern in both high-value-added and low-value-added classrooms, as well as classrooms with high and low scores on the Framework for Teaching.

It seems that some instances of the IRE pattern are fruitful, while others are not. Some high-scoring classrooms included IRE-patterned talk with low-level questions as a management and efficiency measure during a homework review that left room for longer and more focused independent practice. Again, it is not the fact of a low-level IRE pattern, but its use that indicates effectiveness.

At best, rubrics are filled with actions that are sometimes associated with effectiveness, not foolproof indicators of effectiveness. This leaves evaluators in the unenviable position of attempting to
come up with feedback on a teacher’s performance based on a set of indicators that may not indicate anything. Given the importance of some features, the assumption may be that more is better, thus teachers are encouraged to ask more open-ended questions, engage students in more meaningful conversations, or encourage more participation. The inclusion of such indicators to mark the highest levels of proficiency on a rubric may inspire instrumental compliance rather than thoughtful integration. Unfortunately, encouraging participation for participation’s sake may not deepen or extend learning opportunities. But, considering how participation could contribute to the goal of the lesson (how is this effective?) or how participation has been attempted (how does the teacher encourage participation?) is likely to generate useful feedback aimed at improving or expanding effective practices.

Considering how practices are enacted, rather than whether they exist, also limits the problem of sampling associated with observation. At most, beginning teachers are formally observed four times a year for an entire lesson. This represents less than 2% of the time they spend teaching, so it can never be a representative sample of what practices do/do not exist. Even when a handful of brief walkthrough observations are added, teachers might be doing something more than 95% of the time that is never observed by an evaluator. Therefore, examining lessons with a whether rather than how orientation to specific features of instruction is not likely to improve overall quality. Discussing how observed features work, however, might build both pedagogical and metacognitive knowledge about both successful and unsuccessful practices.

Purposes and Possibilities for Measuring ELA Teacher Effectiveness

The challenge of measuring teacher effectiveness goes beyond merely achieving some level of reliability and validity. Constructive evaluation must be able to produce insight that guides an individual teacher’s development. In other words, it must be fruitful, and fruitfulness requires an inquiry—rather than an economic—orientation to classroom observation. As Aukerman and Pandya (2013) have suggested, fruitfulness depends on “a nonneutral, political decision based on a set of beliefs about what education should be” (p. 44). That is, a fruitful assessment does not merely count what already exists and ask for more or less of certain indicators. A fruitful assessment investigates how various features of instruction are contributing to “what education should be” and, therefore, how they may be used to highlight and generate responses toward that shared goal.

In the current policy context, the goal of measuring a teacher’s performance is to identify a teacher’s value-as-measured-by-test-score, rather than to ask how that teacher is valuable. Since observation rubrics will always offer only a partial vision of effective teaching, the valuing of teachers is limited to the practices inscribed in each particular rubric. This limits what counts as effectiveness, but it also limits our ability to nurture and develop teachers at their personal point(s) of need. Just as teachers do when differentiating instruction for beginning readers, we might ask: What is the teacher using—but—confusing? What is she doing well that could be extended? What strategies does this teacher attempt without the desired result? This set of questions may be informed by indicators on a rubric or core practices in a set of standards, but the indicators themselves cannot be used to nurture or develop unless taken up with an orientation of how rather than whether. This means believing teachers and their practices are always already effective at something, even and especially when effects do not match their best intentions. It also means investing in the discussions of how practices work, rather than whether they are occurring.

A shift toward a how rather than a whether orientation when evaluating effectiveness might also precipitate a shift in conceptions of teaching itself—one that leaves room for the possibility that the awareness, analysis, and responsiveness
that led a teacher to enact a certain practice should be uncovered, nurtured, and developed. This conception of teaching as a dynamic coordination of resources and responses makes it complicated to label a teacher as effective or not, but leaves room to consider more aspects of teaching—learning interactions that are of value to a community, such as equity, inclusiveness, meaning, and even joy—all of which are unidentifiable in student test scores.

**Conclusion**

Twenty-five years ago, Darling-Hammond (1990) wrote, “[W]hether intentionally or not a teacher-evaluation system . . . communicates conceptions about teaching and expectations regarding performance priorities, norms for behavior, and the nature of the work itself” (p. 21). The language and content of observation rubrics is therefore consequential not only for evaluation, but for conceptions of teaching quality in general. As Danielson and Schutz (2014) have noted, whenever a set of practices are designated as “core,” some others are consigned to the periphery, and teaching is redefined as a collection of core features rather than the dynamic coordination of features, including those we have not yet specified. Despite the best intentions, rubrics based on a set of core practices limit what counts as effective instruction to that which can be understood in relation to these practices. Though this represents a marked improvement from more unidimensional conceptions of what counts as good teaching (e.g., checklists concerning classroom environment or curricular compliance), it is still a closed set.

We must leave room for the possibility that there are other means to desired ends, that no practice is always right all the time, and that more observations will not always equal better data. We must stop asking whether, and start asking how practices are contributing to enacting a set of beliefs about what education should be. Only then will the assessment of teaching in action contribute to the improvement of teaching and learning.

**References**


*Language Arts*, Volume 93, Number 2, November 2015


Rachael Gabriel is an assistant professor of reading education at the University of Connecticut and can be reached at rachael.gabriel@uconn.edu.