Research Article

Educational Evaluation and Policy Analysis
December 2021, Vol. 43, No. 4, pp. 587–614
DOI: 10.3102/01623737211009267
Article reuse guidelines: sagepub.com/journals-permissions

Measuring Teaching Practices at Scale: A Novel Application of Text-as-Data Methods

Jing Liu
University of Maryland

Julie Cohen
University of Virginia

Valid and reliable measurements of teaching quality facilitate school-level decision-making and policies pertaining to teachers. Using nearly 1,000 word-to-word transcriptions of fourth- and fifth-grade English language arts classes, we apply novel text-as-data methods to develop automated measures of teaching to complement classroom observations traditionally done by human raters. This approach is free of rater bias and enables the detection of three instructional factors that are well aligned with commonly used observation protocols: classroom management, interactive instruction, and teacher-centered instruction. The teacher-centered instruction factor is a consistent negative predictor of value-added scores, even after controlling for teachers' average classroom observation scores. The interactive instruction factor predicts positive value-added scores. Our results suggest that the text-as-data approach has the potential to enhance existing classroom observation systems through collecting far more data on teaching with a lower cost, higher speed, and the detection of multifaceted classroom practices.

Keywords: *classroom research, educational policy, instructional practices, teacher assessment, technology, validity/reliability, econometric analysis, factor analysis, measurements, regression analyses, textual analysis*

Classroom observations play a central role in education evaluation and improvement efforts (Cohen & Goldhaber, 2016; Hill & Grossman, 2013). Policies focused on both teacher professional development and evaluation are predicated on the ability to measure "good teaching" in consistent and fair ways. Gauging teachers' effectiveness through classroom observations also has the added benefits of focusing on classroom processes and providing intuitive and actionable feedback to teachers on areas for improvement, unlike teacher "value-added" or other measures based on student test scores. In practice, structured observation tools are used in districts across the country to evaluate whether and to what extent teachers, from early childhood to high school levels, are demonstrating certain teaching practices known to support student engagement and learning (Kane & Staiger, 2012; Pianta & Hamre, 2009).

Despite their ubiquity, such measures are prone to a range of measurement issues, including unreliable and biased raters (Bell et al., 2018; Ho & Kane, 2013; Kraft & Gilmour, 2017), variability by lesson, time within a lesson, time of year, and content of instruction, and the shifting and evolving nature of student–teacher interactions (Cohen et al., 2018; Hill et al., 2012; Joe et al., 2015; Meyer et al., 2011; Pianta & Hamre, 2009). The issue of temporal fluctuation is particularly noteworthy, given that many teachers are observed no more than 1 or 2 times per year (Cohen & Goldhaber, 2016). Inferences made about teaching quality, and associated supports

provided via coaching or other forms of professional development, may result from data that misrepresent a teacher's general instructional patterns and/or needs. Even if the data are representative, the feedback afforded by a summary of multiple classroom observations is often neither nuanced nor specific enough for teachers to improve their instruction in individual lessons, particularly as instructional needs may also fluctuate over time (Muijs et al., 2018). Moreover, traditional observational methods rely on the fundamental assumption that "good teaching" is characterized by visible features of instruction (Cohen et al., 2020; Kane & Staiger, 2012). There may well be less readily observable aspects of teaching that could provide valuable formative information for teachers, coaches, and school leaders about patterns of interaction or dialogue in classrooms. This creates inherent limitations to observation-based systems for formative and summative assessments of teachers.

Recent advances in computational linguistic methods offer a potentially transformative complement to these issues with traditional classroom observations. It is challenging for human raters to keep track of multiple teacher practices simultaneously, especially those requiring high cognitive loads, and score them without idiosyncratic and biased judgments based on their own experiences with a teacher (Kraft & Gilmour, 2017). Computers, instead of raters, can quickly process transcripts from a large volume of classroom videos and provide automated assessments of specific features of high-quality teaching, including those that are less readily visible. Instead of solely relying on expensive "masters raters," such as those used for a time in Washington D.C.'s IMPACT evaluation system (Dee & Wyckoff, 2015), or principals who are notoriously biased (Bell et al., 2018) and time-strapped (Grissom et al., 2013), text-based metrics could provide distinct and multifaceted insights into classroom practices and processes.

From a policy perspective, this could pay dividends down the road, despite a substantial upfront investment in infrastructure for a text-based system, including classroom microphones, recording devices, software for transcription, and data analysis. Once classrooms were set up to capture audio data, and schools and districts had systems for transcribing and processing such data, these systems could support the collection and analysis of a voluminous amount of information about teaching. Although such data might be helpful in enhancing teacher evaluation systems by providing insight into less readily visible aspects of teaching, it could be far more helpful for costly improvement systems like professional development and coaching.[1] Information about teaching is central to many professional development models (Allen et al., 2011; Kraft et al., 2018). Teachers' practice can shift from day to day in ways that are sometimes hard to detect by human raters, and text-as-data methods have the potential to provide a more comprehensive portrait of instruction, less limited by human observers' capacity. Such methods may also provide teachers an opportunity to receive immediate and more specific feedback that is perceived as less-threatening and more objective than that provided by a principal (Kraft & Gilmour, 2017).

A few recent studies have demonstrated the potential utility of this approach. For example, Kelly and colleagues (2018) used both automatic speech recognition and machine learning to detect teachers' use of authentic questions, an important dimension of classroom discourse (Mehan, 1979; Tharp & Gallimore, 1991). Relatedly, Wang et al. (2013) used an automated speech recognition tool to classify the interaction patterns between teachers and students and provide timely feedback to teachers that could help them monitor students' active participation in classroom discussion. While both studies demonstrate the potential of computational techniques in measuring teaching practices in some ways, they only focus on a single aspect of teaching, and neither of them corroborates the relationship between computer-generated measures and classroom observation scores and student outcomes (Kane & Staiger, 2012). Such convergent, discriminant, and predictive validity evidence is instrumental in illuminating the affordances and constraints of new ways of measuring teaching.

As a proof of concept, we explored the use of novel text-as-data methods to develop automated and objective measures of teaching practices. Drawing on research from instructional science and related disciplines such as social psychology, we leveraged computational power to analyze detailed transcripts of classroom conversations and generate measures of classroom dynamics that include aspects of teaching practices that

may not be easily detected by human raters. Specifically, we answered the following three research questions:

Research Question 1 (RQ1): What measures of teaching can we generate by applying text-as-data methods to transcripts of classroom videos?

Research Question 2 (RQ2): What are the psychometric properties of the computer-generated measures of teaching practices?

Research Question 3 (RQ3): How do the computer-generated teaching practice measures associate with classroom observations scores and value-added scores?

To answer these research questions, we analyzed word-to-word transcriptions of videos of fourth- and fifth-grade English language arts (ELA) classrooms collected as part of the Measures of Effective Teaching (MET) project. Our metrics were based on the premise that teacher–student interactions in classrooms are key for student learning and development, an idea well supported by developmental theory and research (e.g., Brophy, 1986) and serves as the foundation for the Classroom Assessment Scoring System (CLASS) observation protocol, used in districts all over the country and in scores of research studies (Pianta & Hamre, 2009). We created several measures focused on interaction and discourse patterns that captured more discrete teacher behaviors (e.g., the proportion of class time teachers spend talking), as well as higher inference measures (e.g., the level of coordination between teachers' and students' language by matching each other's use of function words), which would be difficult, if not impossible, to detect by human raters.

The automated measures we developed achieve reliabilities similar to those of conventional classroom observations when using *multiple* raters and course sections. Based on these measures, we created three instructional factors that are highly aligned with the dimensions identified by CLASS, the Framework for Teaching (FFT; Danielson, 2013), and Protocol for Language Arts Teaching Observations (PLATO; Grossman et al., 2014)—a classroom management factor, an interactive instruction factor, and a teacher-centered instruction factor. Teacher-centered instruction is consistently negatively associated with value-added scores computed using the Stanford Achievement Test, Ninth Edition (SAT-9), which assesses cognitively complex learning outcomes. This result is robust even after controlling for teachers' average CLASS, FFT, and PLATO scores, suggesting that this instructional factor is not fully captured by these classroom observation tools and our automated approach can detect teaching practices that are not readily visible to human raters. Our analyses also provide some evidence that interactive instruction is positively associated with student outcomes. A back-of-the-envelope cost-effectiveness analysis shows that the potential cost-saving from a text-as-data approach to measure teaching practices is 54% compared with a human-rater approach. While we do not propose that our approach can replace classroom observations done by human raters, it is possible to complement the existing systems with lower costs using a text-based system, especially for districts with fewer resources.

Background

Measures of Teaching and Teacher Quality

Teachers vary substantially in terms of their impact, making them one of the, if not *the* biggest, within-school determinants of student outcomes (Rivkin et al., 2005). As a result, the last three decades have witnessed a great increase in the study of different tools for evaluating teacher performance. A great deal of this research has focused on what teachers do in classrooms that signals high-quality, research-aligned instructional practice. The search for "high-leverage" practices has taken on an increased urgency in the accountability-focused policy climate during the past decade, with its emphasis not only on formal teacher evaluation (Ball & Forzani, 2009; Cohen, 2015) but also on identifying observed needs for targeted professional development efforts (Allen et al., 2011; Kennedy, 2016).

Contemporary research on observation protocols has yielded some insights about the relationship between measured practice and student outcomes. A few studies have found weak to modest correlations between *overall* teacher observation scores and value-added scores. Using MET data, Kane and Staiger (2012) found correlations between scores from observation protocols and value-added ranging from .12 to

.34. Using data from New York City, Grossman et al. (2013) found that teachers in the top quartile of value-added scores had higher scores on some PLATO and CLASS elements than their bottom-quartile peers. Such correlations are subject to change when different tests are used to compute value-added scores (Grossman et al., 2014; Papay, 2011).

Some researchers have gone beyond using summary observation scores and have tried to identify the impact of specific, potentially "high-leverage" teacher practices (Cohen, 2015). Depending on the specific test outcomes, identification strategies, and observation protocols used, these studies have found inconsistent outcomes. For example, Blazar (2015) exploited within-school, between-grade, and cross-cohort variation to overcome bias arising from student sorting within schools and found that "inquiry-oriented instruction" improved elementary math scores by about 0.1 standard deviation. In contrast, Kane et al. (2011) found evidence that classroom management practices contributed to math score growth more than other measured dimensions of teacher practices. However, teacher use of "thought-provoking" questions was most correlated with increasing reading scores. A recent study synthesized similar dimensions across the five most popular protocols used in the MET project (CLASS, FFT, PLATO, the UTeach Observation Protocol, and the Mathematical Quality of Instruction [MQI]) and found classroom management to be the most consistent dimension correlated with student test score growth (Gill et al., 2016). The field still lacks clarity about what aspects of instruction "matter" for which outcomes, and why.

Although the tools used to measure instructional quality have become more refined over time, there are persistent measurement and conceptual issues that complicate our ability to use observation protocols to identify "high-leverage" teaching practices that support student achievement (Kane & Staiger, 2012). Observation tools are designed around particular instructional theories, which limit the set of teaching practices they measure. Some instruments, such as PLATO and the MQI rubric, are domain-specific, with items focused on the clarity and accuracy of teachers' instructional explanations in specific academic subjects (e.g., mathematics; Hill et al., 2008). Others, such as CLASS and FFT, emphasize aspects of "good teaching" that cut across subjects. CLASS draws heavily from developmental theory, emphasizing the warmth and positivity of teacher–student interactions (Hamre & Pianta, 2001). FFT, by contrast, builds upon constructivist learning theory, privileging teachers' questioning techniques and students' intellectual engagement (Danielson, 2011). Despite these conceptual differences, all observation rubrics are predicated on the notion that "good teaching" is something one can see in a classroom. This focus on visible features of classroom interactions, readily observable by trained viewers, inherently limits the scope of understanding (Cohen et al., 2020). Classrooms are busy, social places, rich with discourse between students, as well as between a teacher and students. Observers are inherently limited in how much of this discourse they can notice and process, particularly in the moment.

Second, instrument developers and the school leaders who use observation tools struggle to clearly and consistently define *high-level demonstration* of practice. For example, "regard for student perspectives" (Pianta et al., 2012) might be a practice supported by dozens of empirical studies (Allen et al., 2013; Hamre & Pianta, 2001; Mashburn et al., 2008), but is there a clear threshold between classrooms that rate "high" and "mid" on this construct? Without greater conceptual precision in defining the empirical link between practice and demonstration of practice, it remains difficult to make inferences about teachers based on observations.

Third, reliable classroom observation ratings require skilled and well-trained observers who orient their assessments of quality to the specifics of a rubric rather than their personal interactions with a teacher (Kraft & Gilmour, 2017). In research and in practice, this has proven to be an extremely high hurdle (Cash et al., 2012; Hill et al., 2012; Park et al., 2015; Weisberg et al., 2009). Trained outside raters can generate more reliable ratings than school-based personnel like principals, but even the most practiced raters struggle to keep track of multiple teaching behaviors at the same time (Bell et al., 2014). The cognitively demanding process of conducting classroom observations can therefore limit the set of teaching practices featured in a tool, and

ultimately privilege readily observable aspects of teaching that might not be substantively the most important at supporting student outcomes (Kane & Staiger, 2012). In practice, raters tend to use only a small range of scores, shying away from rating teachers as performing poorly, particularly when raters are school administrators without extensive training (Kraft & Gilmour, 2017; Weisberg et al., 2009). The research has not surfaced clear strategies for mitigating rater bias. Rater effects also vary depending on the specific classroom observation mode as live or video based (Casabianca et al., 2013). Most of the research on classroom observations has been conducted via video-recorded observations, whereas most practitioners engage in live observations, complicating attempts to generalize findings from research studies into classroom observations in practice. For schools and districts wanting to assess and ultimately support high-quality teaching across millions of teachers, these issues present conceptual, logistical, and financial challenges.

Finally, a great deal of research has surfaced the temporal instability of traditional classroom observation instruments (Gitomer et al., 2014; Polikoff, 2015; Smolkowski & Gunn, 2012). Temporal instability can come from multiple sources. For example, student–teacher interactions are inherently dynamic and can evolve over time (Meyer et al., 2011). Many features of "good teaching" might also naturally fluctuate across lessons. Those temporal factors can generate measurement errors in observed scores and complicate inferences made from a handful of observations. Unfortunately, policies at the state and district level are limited by resources, and human raters are costly. As such, many evaluation systems require only one or two observations (Cohen & Goldhaber, 2016). Similarly, many professional development and coaching programs rely on a limited number of observations, which are unlikely to accurately capture a comprehensive portrait of a teacher's practice (Allen et al., 2011). Moreover, traditional classroom observation protocols are designed to capture "typical" or "average" classroom teaching, and they are often not designed to capture fine-grained teaching practices that may naturally vary from lesson to lesson over the course of a school year. What a teacher needs to support students in November may well be distinct from that same teacher's need in May. As such, teachers might find the feedback generated from traditional observational systems to be too general to be useful for improving their evolving instructional needs (Muijs et al., 2018).

Text-as-Data Methods

Text-as-data methods may help address many of the aforementioned issues with human raters. Analyzing large quantities of textual information, often records of verbal and written communications, to gain insights into human interaction and behavior is an increasingly popular approach in political science, economics, communication, and other disciplines (for overviews, see Gentzkow et al., 2017; Grimmer, 2015). Advances in computational methods make it possible to process, quantify, and analyze those data to address historically hard-to-tackle social science and policy questions through building measures that were infeasible before. For example, due to their high precision in transcribing human language, automated speech recognition systems have been used in some medical fields such as psychotherapy to improve patients' mental health (Miner et al., 2020). Moreover, researchers, especially computer scientists, have used computational methods to study conversation features, such as those that can lead to more constructive online discussions (Niculae & Danescu-Niculescu-Mizil, 2016), improve the success rate of job interviews (Naim et al., 2015), change someone's opinion by forming persuasive arguments (Tan et al., 2016), or productively address issues related to mental illness (Althoff et al., 2016).

Education policy researchers have recently started to use text-as-data methods to study a wide ranges of topics, including features of productive online learning environments (Bettinger et al., 2016), teacher applicants' perceptions on student achievement gaps (Penner et al., 2019), and strategies schools adopt in school reform efforts (Sun et al., 2019). But the use of such methods in natural classroom settings remains rare. One exception is Wang et al.'s (2013) study, which featured an automatic feedback system using a speech recognition recorder for teachers with respect to teacher and student talk, silence, overlap talk, and episodes of quality discussion. The authors find that

when teachers received timely feedback, they significantly reduced their talking and discussion time significantly increased. Another exception is Kelly et al. (2018), which applies automatic speech recognition and machine learning to automatically identify whether a question a teacher asks in her classroom is authentic, in that the question does not have a specific answer predetermined by the teacher. Their computer-coded measure of authenticity achieves reasonably high correlations with coding from human raters.

Some emerging work suggests that computer-generated measures can not only assess but also support teaching. For example, Lugini and colleagues (2020; Lugini, 2020) have started to track student discussion patterns using automated tools and demonstrate how such information supports teachers' instructional reflection and future planning. Such evidence is key, as many have argued that the promise of classroom observations lies in their formative—rather than summative or evaluative—potential (Cohen & Goldhaber, 2016; Hill & Grossman, 2013; Pianta & Hamre, 2009). For districts to invest in new systems or tools like text-as-data approaches, they will likely need to feel confident that such advances would not only measure teaching consistently but also provide timely and actionable information to teachers.

The development of text-as-data methods provides an unprecedented opportunity to extend prior classroom-based research, enabling researchers to address some of the inherent limitations of observation protocols and shed new light on how teaching practices affect student outcomes. If text-as-data methods could supplement—or even replace—traditional classroom observations in research studies, researchers could save precious time and resources by minimizing the expensive trainings, ongoing calibration, and inherent biases associated with humans scoring multifaceted and complex classroom instruction.

More importantly, the fast development of natural language processing and related techniques can facilitate detection of latent but not readily observable features of classrooms that are associated with student outcomes. Once districts cross the admittedly high hurdle of developing robust systems for audio capture, transcription, and data processing, they could rely on such systems to provide far more consistent and ongoing information to teachers about their practice than a human-rater-based system ever could. Districts also invest heavily in evaluation and instructional support systems, like teacher professional development. Text-based measures could enhance such support efforts by providing coaches, teachers, and other support providers with far more data and less readily visible insights about a teacher's instruction than could be gleaned from a handful of classroom observations.

Successful application of text-as-data methods in measuring teaching practices also has the potential to improve teaching quality and student achievement more generally. Compared with classroom observations from human raters, teachers might benefit from computerized metrics that provide faster feedback loops, which they may also perceive as more objective. Finally, if principals can apply text-as-data methods to collect similar information as walk-throughs, with better precision, it may also free up some time to focus on giving feedback to teachers. We are not suggesting that such methods would supplant human-based feedback and evaluation systems, but are instead arguing they may provide a powerful complement.

Prior work has demonstrated that it is possible to automatically measure one aspect of teaching reasonably well using text-as-data methods (e.g., Kelly et al., 2018). However, there are a myriad of other aspects of discourse that could be detected with automated measures. A central goal of this work is to use a range of computational techniques to demonstrate the potential of text-as-data methods in capturing a variety of classroom interactions.

Data and Sample

The MET Project

The data for this study come from the Bill and Melinda Gates Foundation–funded MET project, which is, to date, the largest research project in the United States on K–12 teacher effectiveness. More than 2,500 fourth- through ninth-grade teachers in 317 schools across six districts participated in the MET project over a 2-year span (academic years 2009–2010 and 2010–2011). The MET project's sample is composed mainly of students from high-poverty, urban school districts.[2]

Many of the MET project's features make it an ideal data source for this study. First, the project collected video recordings for each participating teacher, which enables repeated measurement of classroom processes using different tools. Second, the MET project aimed to advance the use of multiple measures of teacher effectiveness and provides both value-added scores and classroom observation scores based on five major observational protocols. This rich set of measures allows us to compare the measures we created to conventional ones.[3]

This study focuses on fourth- and fifth-grade ELA classrooms and teachers. We chose to focus on ELA as a starting point because a long line of research has demonstrated that classroom discourse and instructional formats matter for student learning in language arts classes (e.g., Beck & McKeown, 2001; Chinn et al., 2001; Nystrand & Gamoran, 1991). In addition, focusing our resources on transcribing more videos per teacher in one subject would provide us more data to evaluate the psychometrics properties of our measures and reduce measurement errors. While more research is needed to verify whether the proposed measures and methods would demonstrate similar properties across subjects, we purposefully focus on cross-domain teacher practices that are more likely to be generalizable.

Table 1 describes the sample, which includes a racially and socioeconomically student body: 43% of students are eligible for free or reduced-price lunch, 42% are African American, and 23% are Hispanic. The average standardized ELA scores on state exams in 2009 and 2010 are both around 0.1, which suggests that teachers in our sample taught a slightly higher than average set of students, based on achievement test performance. The large majority of teachers are women (92%), and most are White (63%); 32% are African American. This teacher sample reflects the characteristics of the students they teach, though it represents a much bigger share of African American teachers than an average district in the United States. On average, teachers have worked in their current district for 6 years, and 46% have a master's degree or higher.

TABLE 1

Descriptive Statistics

Characteristic	*M*	*SD*
Students (n = 13,370)		
Age	9.73	(0.87)
Male	0.50	
Gifted	0.09	
Special education	0.09	
English language learner	0.13	
Free or reduced-price lunch	0.43	
Race/Ethnicity		
White	0.25	
African American	0.42	
Hispanic	0.23	
Asian	0.06	
Other	0.03	
ELA score		
2009	0.12	(0.96)
2010	0.09	(0.97)
Teachers (n = 258)		
Male	0.08	
Race/Ethnicity		
White	0.63	
African American	0.32	
Hispanic	0.03	
Other	0.01	
Years in district	5.98	(5.43)
Master's or higher	0.46	

Note. Data are restricted to teachers who participated in the MET project's second-year randomization process. Four teachers are missing from the sample because the quality of their classroom audios is not sufficient for precise transcription. Student- and classroom-level statistics are calculated using both 2009 to 2010 and 2010 to 2011 data. Different variables may have different numbers of observations. ELA = English language arts; MET = Measures of Effective Teaching.

Classroom Videos and Transcriptions

The MET project collected classroom videos for all participating teachers. For subject-matter generalists, primarily elementary school teachers, the MET project collected videos of both math and ELA classes on four different days, producing four videos for each subject.[4] The recording days were spread out during February and June 2010 in the first year of the study and between October 2010 and June 2011 in the second, with the aim of making the videos more representative of teachers' practices. Teachers were required to teach half of the video-recorded classes on focal topics chosen by MET project researchers; the other half of the video-recorded classes could cover topics of teachers' choice.

Focal topics for fourth- and fifth-grade ELA classrooms included expository writing, making inferences and questioning, personal narratives, revision of writing, summarizing main ideas, and identifying theme and point of view.

For the teachers in our sample, each had four video recordings in the first year of the MET project and four recordings in the second year. Our main analytic sample consists of transcriptions of the first 30 minutes of all four videos for each teacher from the first year.[5] As a result of some videos having low-quality sound (or no sound) and errors by the transcription company, not all teachers in the sample have four complete videos transcribed. Approximately 75% of teachers in the sample have four videos transcribed, 20% have three, and 5% have two. In total, 976 videos were transcribed, amounting to 29,436 minutes of language arts teaching.

A professional transcription company transcribed each video at a word-to-word level, with time stamps attached to the beginning and end of each speaker's turn. Because the data usage agreement restricted us from sharing the videos with our transcribers, only audio was used in the transcription process. One drawback of this approach is that while classroom observers could note both verbal and nonverbal markers, our data are restricted to verbal information and do not include other cues, such as facial expressions, gestures, and classroom artifacts.

To the extent possible, the transcribers identified students by voice and labeled them as Student A, Student B, and so forth, or students (plural) if multiple students were talking simultaneously. The MET project used a specially designed rig to record the classes, and two microphones captured teachers' and students' voices. Throughout the recordings, teachers' voices have much better audio quality than students' because there were multiple students using the same microphone, and they were often not loud enough to be picked up clearly. When a voice was not identifiable, transcribers labeled it "inaudible." In some instances, it was difficult to precisely identify which student was talking. Supplementary Appendix A in the online version of the journal provides an example of the data. These detailed data allow us to identify several features of students' and teachers' language and interaction patterns in a classroom, such as the quantity (e.g., number of words spoken), style (e.g., usage of open-ended questions), and content (e.g., linguistic coordination).

Value-Added Scores and Classroom Observation Scores

The MET project data contain multiple measures of teaching quality. First, they include two measures of teacher effectiveness based on student performance for each subject—one derived from state achievement test scores and one based on the SAT-9 Open-Ended Tests in ELA, which are cognitively more demanding and lower stakes than the state tests. Previous research has shown that value-added scores calculated using state tests and the SAT-9 are weakly correlated (Papay, 2011). Several studies using MET project data have also found inconsistent relationships between classroom observation scores and value-added scores calculated using different tests, often with stronger correlations for low-stakes supplemental tests (Cohen, 2015; Grossman et al., 2014). For this study, we used value-added scores from the MET project database.[6]

The MET project data also contain multiple observational measures of teacher practices. We focused on measures generated from CLASS, FFT, and PLATO. Both CLASS and FFT measure aspects of classrooms that are hypothesized to be generalizable across subjects. CLASS is based on developmental theory.[7] Its primary focus is various aspects of teacher–student interactions, which closely align with the measures we propose to create using text-as-data methods.[8] FFT is the most widely used observational tool in the United States and grounded in a "constructivist" view of student learning, with emphasis on intellectual engagement such as high-quality questioning. We also chose PLATO because it focuses on effective teacher practices specific to literacy instruction (Grossman et al., 2014).

The CLASS protocol is designed to measure how teachers support children's social and academic development, with a focus on daily teacher–student interactions (Hamre et al., 2007). CLASS comprises three broad domains of teacher behaviors: emotional support (measured dimensions include positive climate, negative climate, teacher sensitivity, and regard for student perspectives), classroom organization (measured dimensions include behavior management,

productivity, and instructional learning formats), and instructional support (measured dimensions include content understanding, analysis and problem-solving, instructional dialogue, and quality of feedback; see White & Rowan, 2012, for a more detailed description of the CLASS dimensions and domains). Several studies have found that teachers rated higher on CLASS are associated with higher student test scores (e.g., Araujo et al., 2016) and that teachers supported with the CLASS framework have improved interactions with students and correspondingly substantial gains in student achievement at the secondary school level (Allen et al., 2011).

Teachers in nearly 30 states are both evaluated and coached using Danielson's FFT, making it the most commonly used protocol in the country (Goe et al., 2012). Although FFT has evolved over the years, the version used in the MET study focused on the broad domains of classroom environment (e.g., creating an environment of respect and rapport, managing classroom procedures) and instruction (e.g., using questioning and discussion techniques, demonstrating flexibility and responsiveness; for a detailed discussion of the measurement properties of FFT, see Liu et al., 2019; Mantzicopoulos et al., 2018). FFT is undergirded by a constructivist model of teaching, in which teachers facilitate—rather than lead—learning by engaging students in activities and discussions that promote critical thinking and encourage intellectual argumentation (Danielson, 2013). There is mixed evidence on the relationship between FFT scores and student outcomes (Gallagher, 2004; Holtzapple, 2003; Kane & Staiger, 2012; Liu et al., 2019), with recent work suggesting FFT scores explain a low percentage of the variance in student outcomes (Patrick et al., 2020).

The version of PLATO used in the MET project includes six teaching practices: modeling, strategy instruction, intellectual challenge, classroom discourse, time management, and behavior management. As a subject-specific protocol, PLATO complements CLASS by focusing on aspects of language arts teaching highlighted in the research literature. These include teacher scaffolding of literacy tasks (Beck & McKeown, 2001; Graham & Harris, 1993; Greenleaf et al., 2001; Palincsar & Brown, 1987) and providing challenging disciplinary tasks in which students are expected to engage in the majority of the intellectual work (Newmann et al., 1998). Several studies have found that ELA teachers with higher value-added scores perform better on multiple dimensions in PLATO (Grossman et al., 2013, 2014).

Findings

RQ1: What measures of teaching can we generate by applying text-as-data methods to transcripts of classroom videos?

We built two types of measures to capture teacher practices that are more nuanced and fine-grained than those captured in current observation protocols. The first focuses on patterns of discourse, primarily using information about language sources (e.g., teachers or students), time stamps, and words and punctuation marks (e.g., functional words, question marks) associated with a specific linguistic category. To identify words and punctuation marks that are meaningful for classroom conversation, we used the Linguistic Inquiry and Word Count (LIWC) program (Pennebaker et al., 2014), which counts the percentage of words that reflect different emotions, thinking styles, and social concerns of a given utterance.[9] LIWC is widely used in computational linguistic analyses and covers 93 linguistic dimensions,[10] from which we selected ones that are relevant for classroom teaching. The three measures we proposed include turn-taking, targeting, and the use of analytical and social language.

The second type of measure we developed captures more substantive aspects of teaching using more sophisticated text-as-data methods, including both partially automated (i.e., supervised) and fully automated (i.e., unsupervised) models.[11] We anchored the measures in theoretically identified indices of effective teaching and took advantage of the highly detailed data and the capacity of text-as-data methods to uncover latent language features. This set of measures includes language coordination, questioning, and the allocation of time between academic content and routine, of which we detail the definition below.

These micro-level measures can be captured with precision by using computers, and they have the potential to inform the macro-level norms of classroom participation, teacher–student dynamics, and teachers' pedagogical styles. For example, a classroom with extensive turn-taking between the teacher and students may suggest the teacher is more attentive and responsive to students' ideas. We also aggregated these measures using factor analyses to consider the interdependence between them, identify latent instructional factors, and provide more parsimonious models to facilitate our analysis in RQ3. While these measures cannot capture every aspect of teaching that might support student outcomes, they represent a step toward understanding the micro-processes of teaching, which are difficult to detect with human raters, that could better support teachers in providing more uniformly high-quality instruction.

Turn-Taking

Research on classroom discourse, especially in ELA classes, has identified different "instructional frames," with a recitation format at one extreme and a collaborative reasoning format at the other (Applebee et al., 2003; Chinn et al., 2001; Michaels et al., 2008; B. M. Taylor et al., 2002). Cazden (1988/2001) described a recitation format, or "Initiate, Respond, Evaluate" (IRE), as one in which "teachers give directions and children nonverbally carry them out; teachers ask questions and children answer them, frequently with only a word or a phrase." (pp.134) In contrast, a collaborative reasoning format promotes consistent and authentic student talk. Teachers serve the role of facilitator, rather than evaluator, and encourage students to directly engage with each other's ideas (Nystrand et al., 1997; Tharp & Gallimore, 1991). A sociocognitive view supports the idea that students need to be active agents in classroom interactions to construct meaning and form their own interpretations. Several studies have shown that high-quality discussion and exploration of ideas—not just the presentation of high-quality content by the teacher—can enhance students' literacy achievement and reading comprehension (e.g., Applebee et al., 2003; Grossman et al., 2014; Nystrand & Gamoran, 1991).

In characterizing instructional frames, turn-taking represents a key parameter that reflects who controls the classroom conversation. Following Chinn et al. (2001), we extracted multiple features of turn-taking, including the number of turns taken per minute, to measure the frequency of exchange of ideas between teachers and students; the percentage of time teachers talk in a given class, to measure teacher control in classroom discourse; and the average time spent per turn for both teachers and students and the number of words spoken per minute, to further gauge the allocation of control over discussion between teachers and students.

Targeting

Linguists suggest that in a classroom, teachers might *target* individual students or themselves as the focus of classroom discourse for different purposes. For example, teachers might refer to themselves more often when demonstrating a problem-solving process or modeling a certain skill (e.g., "If *I* can get your eyes here with *me* and *I*'ll go through *our* PowerPoint"). Teachers may also refer to students frequently to take control over the conversation or manage classroom order (e.g., when *you* get to page 178, I want *you* to stand up and hold *your* book up). Previous research has shown that in a conversation, the use of personal pronouns points to the targets of communication (McFarland et al., 2013). The sociolinguistic literature suggests personal pronouns also strongly relate to social status, with higher social status individuals using "you" more frequently and lower social status individuals using "I" more frequently (Pennebaker, 2011; e.g., *I*'m going to ask *you* again only if *I* call on *you. You* are going to have to be still right here because that is going to cause a problem.). Thus, these words can also reflect the power structure of a classroom. We extracted referential words (i.e., personal pronouns) to serve as proxies for how a teacher allocates her attention in her language and the power structure of a classroom.

Analytical and Social Language

Teachers provide models of language use for their students. A large body of research suggests

children's language development is shaped by the adults with whom they interact, both teachers and parents (Cabell et al., 2015; Goldin-Meadow et al., 2014; Hassinger-Das et al., 2017; Huttenlocher, 1998; Song et al., 2014). While there are many different types of language styles, we focused on two features of teachers' language use to put in contrast of analytical, logical, and consistent thinking versus more intuitive, narrative speaking, or social language. Analytical thinking is a category of words that capture formal, logical, and hierarchical thinking,[12] such as prepositions (e.g., to, with, above), cognitive mechanisms (e.g., cause, know, hence), and exclusive words (e.g., but, without, exclude), while social language concerns about human interaction, such as non-first-person-singular personal pronouns (e.g., we, us, you all) and verbs related to human interaction (e.g., sharing, talking).

Language Coordination

Social psychologists suggest that communicative behavior is patterened and coordinated, like a dance in the form of human talk (Niederhoffer & Pennebaker, 2002). In a classroom setting, an effective teacher might build on the previous student contribution by revoicing it and using it to set the direction of subsequent conversation, so that students' ideas are fully incorporated in classroom discourse in ways that also engage their deeper cognitive processes (Herbel-Eisenmann et al., 2009; Nystrand et al., 1997). Literature focused on teachers' "up-take" suggests that effective teaching is likely to exhibit a higher level of "language coordination" (Howe et al., 2019), in that there are more synchrony in teachers' and students' language. Although the exact definition of up-take might vary in the literature, we can observe the cues of *coordination*, such as the similarity of the type of words used between teachers' and students' language to serve as a proxy for one aspect of this construct.

Language style matching is an index that measures whether two people in a natural conversation match each other's speaking behavior or style using functional words. A high score indicates a better coordination process. Language style matching is designed for a dyadic conversation and can be computed at both a turn-to-turn level and a whole-conversation level. In this study, we treated all students as one party and computed language style matching by aggregating all words spoken by the teacher and students in a classroom. For details of this method and examples of functional words, see Supplementary Appendix B in the online version of the journal.

Questioning

Questioning plays a key role in eliciting rich discussion and engaging students, and both the quantity and quality of questions play an integral role. Chinn et al. (2001) argued that an overall decrease in the number of teacher questions is a primary indicator of decreased teacher control in a collaborative reasoning format. At the same time, the *nature* of those questions—whether they stimulate students' reasoning, have multiple correct answers, and are nonformulaic (i.e., whether they are open-ended or authentic questions)—determines whether they characterize a dynamic and dialogic conversation (Nystrand, 2006). Other types of questions may expect one-word, yes-or-no answers (i.e., are closed-ended questions) or be related to procedure, rhetoric, or discourse management. Applebee et al. (2003, pp. 699–700) described these latter question types as procedural questions (e.g., "How many pages do we need to read?"), rhetorical questions, "discourse-management questions" (e.g., "What?" "Excuse me?" "Did we talk about that?" "Where are we [in the text]?"), and finally questions that initiated discourse topics (e.g.,"Do you remember our discussion from yesterday?").

To measure the number of questions a teacher asks, we used *regular expressions*, a programming procedure to automatically identify clearly defined textual patterns, to extract question marks and the corresponding questions a teacher asks in a class. To distinguish the nature of the questions being asked, we need to "teach" our computer algorithm the features that differentiate open-ended questions from those that are not so that we can make reasonable predictions. To do this, two raters with extensive K–12 classroom experience and who are currently education researchers firsthand-labeled 600 randomly selected questions from the set of questions in the data, which serve as a "training" data set. As there are many features that can predict whether a question is open-ended or not, conventional

regression-based prediction methods are infeasible because there are likely more variables (i.e., words) than observations. Lasso is a feature reduction regression method that is designed to deal with this scenario.[13] We used a fivefold Lasso procedure to "learn" from this training sample and then make predictions for the rest of the questions. Supplementary Appendix Figure C1 in the online version of the journal shows the most predictive words for open-ended questions and non-open-ended questions. Supplementary Appendix C in the online version of the journal provides additional details on this method.

Allocation of Time Between Academic Content and Routine

Early research on the process–product model of teaching effectiveness focused on time on task, or the amount of time for which students are exposed to academic content (Brophy & Good, 1986). Modern classroom observation protocols also emphasize teachers' ability to minimize time spent on disruptions and classroom management and to provide ongoing learning opportunities to students (Gill et al., 2016; Pianta & Hamre, 2009). We measured the proportions of a teacher's language dedicated to academic content and classroom management routines as a proxy for the productivity of classroom time. We hypothesized that teachers who spend *less* time talking about routines are more likely to have a productive classroom. To test this hypothesis, we used *topic modeling*, a Bayesian generative model, to differentiate task-related and classroom management–related topics (Blei, 2012). A topic model automatically estimates the proportion of language devoted to different topics based on the co-occurrence of words across documents (e.g., classroom transcripts). Employing this approach enabled us to label the themes of those topics and classify them as related to academic content or routines. Supplementary Appendix D in the online version of the journal describes the details of this method. As Supplementary Appendix Figures D1 and D2 in the online version of the journal show, the most prevalent two topics in both 15- and 20-topic models have representative words that point to classroom management (e.g., group, partner, minute), with the rest of the topics about specific academic content (e.g., idea, predict, subject).

RQ2: What are the psychometric properties of the computer-generated measures of teaching practices?

Table 2 presents descriptive information on the measures described above. Consistent with findings from prior literature, teachers in the MET project sample usually occupied the central role in classroom talk. On average, they spent 85% of class time talking to their students. Classrooms varied considerably in prevalence of back-and-forth conversation, with an average of 4.5 turns per minute and a standard deviation of 2.1 turns per minute. Teachers used "you" more frequently than "I," suggesting that teachers address students much more often than they refer to themselves. In general, teachers' language use was more analytical than social (i.e., more formal, logical, and hierarchical instead of narrative or interpersonal), although the proportion of analytical language varied substantially across classrooms. Teachers and students exhibited high language coordination, with little variability across classrooms. On average, teachers asked about 0.22 open-ended questions per minute and spent 10% of their language on classroom management routines (e.g., putting students into groups or managing classroom disruptions) instead of instruction. Both these measures show a significant amount of variation, suggesting that classroom discourse patterns vary substantially across the classrooms in the MET study.

Psychometric Properties

For evaluation purposes, we want measures of teaching that differ systematically across teachers and are not influenced by idiosyncratic sources of variation, such as a specific lesson or the rater's mood on the observation day. Given that we have multiple class sessions transcribed for each teacher, one source of variation that we can identify comes from the sessions themselves (i.e., lesson "error"). As MET researchers prescribed a list of topics for teachers, we assume such topics would have a distinct effect on teacher practices. Unlike conventional observation protocols, the measures we created are extracted from the same computer program and are not subject to different raters' judgments; thus, they are free

TABLE 2

Computer-Generated Metrics on Teacher Practices

Variable	*M*	*SD*
Turn-taking		
Turns per minute	4.50	(2.08)
Proportion of time teacher talks	85.22	(10.90)
Average words per minute	115.45	(24.70)
Targeting (teacher)		
"You" (%)	4.76	(1.55)
"I" (%)	2.51	(1.17)
Analytic and social language (teacher)		
Analytic thinking	38.20	(12.39)
Social words (%)	13.71	(2.22)
Language coordination		
Language style matching (0–1)	0.80	(0.10)
Questioning (teacher)		
Open-ended questions per minute	0.22	(0.12)
Allocation of time between academic content and routine (teacher)		
Routine language (%)	10.63	(5.91)

Note. All statistics are calculated at the level of teacher video. Analytic thinking is a composite score that is converted to percentiles.

from individual raters' biases. Student composition is another possible source of bias, in that observations may be sensitive to the characteristics of students in a classroom (Campbell & Ronfeldt, 2018); however, elementary school teachers often teach only one course section, so the data do not allow for comparison of individual teachers teaching different students. MET project researchers found that course section (i.e., the student body) played little role in shaping observation scores in their sample, so this inability to compare results across course sections may not be constraining (Kane & Staiger, 2012). Future research could use transcriptions of classroom videos collected in the second year of the MET project to separate out the effects of teaching different cohorts of students.

We calculated the reliability of each measure y_{ij} for teacher *i* and lesson-topic *j* by running a cross-classified multilevel model that decomposes the variance of each teacher practice as a teacher part (γ_i), a lesson-topic part (δ_j), and a random error part (ε_{ij}). The model used is as follows:

$$y_{ij} = \alpha + \gamma_i + \delta_j + \varepsilon_{ij}. \quad (1)$$

The proportion of variance attributed to the teacher is the estimated reliability. Table 3 shows that the majority of the measures have reliability scores ranging from .15 to .35, with social words and language style matching having reliability scores below .2. As a benchmark, for the *domains* captured by the five instruments used in the MET project (e.g., emotional support under CLASS), reliability scores range from .14 to .37 when using more than one rater. Thus, each *individual measure* achieves a reliability score similar to those for these broader domains included in observation protocols, without relying on multiple raters. For the rest of our analyses, we use the averages of each measure at the teacher section level to reduce measurement error.

Factor Analysis

The metrics described above may reflect only a few latent instructional factors, so following prior work (Grossman et al., 2014), we conducted a factor analysis using all these metrics to identify such constructs. Three factors were retained based on the Kaiser criterion (eigenvalue greater than or equal to 1). Table 4 shows

TABLE 3

Variance Components and Reliability of Teacher Practices

Variable	Teacher	Lesson	Error	Reliability
Turn-taking				
Turns per minute	29.21	0.02	70.78	.29
Proportion of time teacher talks	21.73	0.56	77.71	.22
Average words per minute	31.34	0.00	68.66	.31
Targeting (teacher)				
"You" (%)	24.12	7.24	68.64	.24
"I" (%)	26.03	10.58	63.39	.26
Analytic and social language (teacher)				
Analytic words (%)	34.06	6.53	59.42	.34
Social words (%)	18.31	4.19	77.5	.18
Language coordination				
Language style matching (0–1)	14.84	0.55	84.61	.15
Questioning (teacher)				
Open-ended questions per minute	32.23	0.06	67.71	.32
Allocation of time between academic content and routine (teacher)				
Routine language (%)	35.34	0.48	64.18	.35

Note. Analysis is conducted at the class level. The variance components are based on a cross-classified multilevel model that decomposes each variable into a teacher component, a lesson-topic component, and an error component.

TABLE 4

Factor Analysis

	Rotated			Nonrotated		
Variable	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
Turns per minute	0.113	**0.696**	−0.341	**0.518**	**0.582**	−0.086
Proportion of time teacher talks	−0.102	−0.112	**0.651**	−0.285	−0.198	**0.571**
Average words per minute	0.030	0.373	0.329	0.139	0.200	**0.435**
"You"	**0.638**	0.003	−0.088	**0.555**	−0.323	−0.050
"I"	0.211	−0.065	**0.442**	0.046	−0.278	**0.406**
Analytic words	**−0.718**	−0.232	−0.097	**−0.695**	0.229	−0.205
Social words	**0.509**	0.052	−0.104	**0.475**	−0.209	−0.054
Language style matching	0.103	0.069	−0.265	0.180	0.066	−0.222
Open-ended questions per minute	0.064	**0.738**	0.108	0.398	**0.529**	0.349
Routine language	**0.582**	−0.045	−0.145	**0.497**	−0.316	−0.122

Note. Analysis is conducted at the teacher section level. Factors are extracted using the principal factor method. Rotation is orthogonal. Bold values indicate that the size of the loading is bigger than 0.4 for the corresponding factor.

the factor loadings.[14] The rotated factor structure is not substantially different from the nonrotated factor structure, so we focus on the rotated factor loadings for ease of interpretation.

The first factor is heavily loaded on using "you," social language, and routine language and negatively loaded on analytic language. This factor points to a dimension of teaching related to classroom management and establishing routines. Teachers with high scores on this factor spend more of their language managing student disruptions, putting students into groups, and performing other noninstructional activities. They are also more likely to attentionally target

students (i.e., using "you") and engage in social talk that is more narrative and intuitive in nature in contrast to analytical language.[15] Thus, this factor appears to carrying less desirable teaching practices. The second factor highlights the use of open-ended questions, more back-and-forth conversation between the teacher and students, and more words spoken per minute. This dimension indicates a more interactive instructional format and students taking a more active role in classroom discourse. The third factor is primarily loaded on more teacher talk, more self-reference by the teacher (i.e., using "I"), and more words spoken per minute, suggesting a classroom dominated by the teacher's speech, leaving students little time to participate. Although both the second and third factors feature more talk during a class, the distinction between them lies in whether students have abundant opportunities to express their opinions and whether the discourse is interactive. Overall, the three factors represent a classroom management dimension, an interactive instruction format, and a teacher-centered instruction format. In the rest of the analyses, we examine how these factors relate to classroom observation scores and value-added scores to corroborate our interpretation.

RQ3: How do the computer-generated teaching practice measures associate with classroom observations scores and value-added scores?

Classroom Observation Scores

The purpose of correlating the instructional factors with classroom observation scores is twofold. First, this shows whether computer-generated measures and observation protocol measures capture similar constructs so that they might be used interchangeably on some aspects of teaching to provide teachers and/or researchers similar information. If we observe strong evidence on convergent validity, it is possible to improve the measurement quality of these constructs through automatically scoring far more lessons than what can be done with human raters, alleviating concerns related to temporal instability mentioned above. Second, the size and direction of these correlations can facilitate and corroborate the interpretation of the newly created instructional factors. For example, we would expect the factor "interactive instruction" to be positively correlated with instructional dialogue[16] in CLASS because they should capture similar constructs. For the classroom management factor and the teacher-centered instruction factor, we would expect to see negative correlations with similar dimensions in classroom observations. Higher scores on text-as-data measures reflect a greater focus on management and teacher-controlled talk, whereas higher scores on related measures of traditional observations reflect less focus on management (i.e., in well-managed classrooms, teachers do not talk about management). Using CLASS, FFT, and PLATO can generate a deeper understanding of the instructional factors we developed and their relationships with other widely used tools to measure and support instructional quality. We correlated each of the three new factors with each of the domains of all three observational protocols.

Overall, as shown in Table 5, the correlations are small but still meaningful, given that both the computer-generated measures and the observational measures contain error. Factor 1, the classroom management factor, has the strongest correlations with student behavior management under CLASS ($r = -.280$, $p < .01$), FFT ($r = -.233$, $p < .01$), and PLATO ($r = -.183$, $p < .01$). These consistent correlations provide support for the hypothesis that Factor 1 captures teacher time spent on managing disruptions. At the same time, Factor 1 is also significantly correlated with other, less obviously connected teaching practices, such as the domains of emotional support and instructional support under CLASS and classroom discourse under PLATO. These unexpected correlations highlight the need to improve the text-based measures for better convergent validity.

Factor 2, the interactive instruction factor which features abundant back-and-forth interaction between teacher and students, is primarily related to the CLASS domain of instructional support, which emphasizes teachers' use of consistent feedback and their focus on higher order thinking skills to enhance student learning. The strongest correlations are with the finer-grained CLASS dimension of instructional dialogue ($r = .259$, $p < .01$) and PLATO scale for classroom discourse ($r = .220$, $p < .01$) under

TABLE 5

Correlations With CLASS, FFT, and PLATO

Dimension	Factor 1: classroom management	Factor 2: interactive instruction	Factor 3: teacher-centered instruction
CLASS			
Domain 1: emotional support			
Positive climate	.124*	.168**	−.07
Negative climate	.217**	−.006	.04
Teacher sensitivity	.198**	.169**	−.069
Domain 2: classroom management			
Behavior management	−.280**	−.002	.082
Productivity	−.186**	.043	.078
Instructional learning formats	.100†	.166**	.04
Domain 3: instructional support			
Content understanding	.049	.233**	.029
Analysis and problem-solving	.090	.120*	.011
Quality of feedback	.202**	.239**	−.024
Instructional dialogue	.191**	.259**	−.117*
Student engagement	.036	.143*	.028
FFT			
Creating an environment of respect and rapport	−.115†	−.003	−.034
Communicating with students	−.048	.104†	−.067
Establishing a culture for learning	.016	.011	−.118*
Engaging students in learning	.052	−.012	−.129*
Managing classroom procedures	−.120*	.098†	−.053
Managing student behavior	−.233**	.032	.008
Using assessments in instruction	.041	.058	−.067
Using questioning and discussion techniques	.088	.004	−.170**
PLATO			
Intellectual challenge	.083	.137*	−.172**
Classroom discourse	.143*	.220**	−.098
Behavior management	−.183**	.055	.094
Modeling	−.140*	−.025	.325**
Strategy use and instruction	−.068	.043	.310**
Time management	−.167**	.112†	.042
Representation of content	.002	−.049	−.106†

Note. CLASS = Classroom Assessment Scoring System; FFT = Framework for Teaching; PLATO = Protocol for Language Arts Teaching Observations.
†$p < .10$. *$p < .05$. **$p < .01$.

PLATO, both of which capture a similar construct and are scored similarly (i.e., high scores reflect more discourse). Instructional dialogue (CLASS) focuses on teachers' use of questioning and discussion to guide and prompt students' understanding, and classroom discourse (PLATO) emphasizes opportunities for student talk and teachers' uptake of students' ideas. It is thus reasonable to argue that the frequency of open-ended questions and the frequency of turn-taking, which are Factor 2's driving components, together capture a more dialogic form of instruction. We do not observe stronger correlations between Factor 2 and the domains in FFT, despite the fact that FFT is conceptually designed to assess and support more dialogic instruction.

Factor 3, the teacher-centered instruction factor also has some conceptually intuitive relationships with different observational scales. In particular, this factor has negative and statistically significant correlations with instructional dialogue (CLASS), establishing a culture for learning, engaging students in learning, using questions and discussion (FFT), and intellectual challenge (PLATO). The traditional observational measure prioritizes student autonomy, discussion, and academic rigor, while the text-based measures capture an emphasis on teacher lecture and minimal student participation in academic discourse. Factor 3 also has positive correlations with the PLATO domains of modeling and strategy use and instruction, which privilege explicit, teacher-led instruction (Cohen, 2015; Cohen et al., 2018; Grossman et al., 2013). This, too, makes intuitive sense, as the PLATO metrics assess the degree to which teachers provide students with detailed instruction about an academic process or skill, perhaps contributing to higher levels of "teacher talk."

Overall, we observe strong and consistent correlations between the three instructional factors identified with text-as-data methods and the most theoretically aligned constructs across CLASS, FFT, and PLATO, though we also observe some unexpected correlations. The comparison between these factors and the three observation protocols largely confirms that these text-based factors capture a classroom management dimension, an interactive instruction format, and a teacher-centered instruction format, respectively.

Value-Added Scores

To test whether the identified instructional factors are associated with teachers' contributions to student achievement gains, we conducted regression analyses using value-added scores. The use of both state ELA and supplemental SAT-9 tests in the creation of the value-added measures (VAMs) allows us to examine whether the nature of the assessment changes the relationship between the novel computational measures of teaching detailed here and VAMs (Grossman et al., 2014). We ran the regression model below to consider multiple factors simultaneously:

$$VAM_{tdg} = \beta Practice_{tdg} + \theta_d + \tau_g + \varepsilon_{tdg}. \quad (2)$$

In this model, VAM_{tdg} is the value-added score for teacher t in district d and grade g. We controlled for district fixed effects because each district administered a different test. We controlled for grade fixed effects to compare teachers who teach in the same grade. For $Practice_{tdg}$, we used the factors derived from the factor analyses described above (e.g., interactive instruction). We also ran a separate model controlling for student characteristics. We further controlled for teachers' average CLASS, FFT, and PLATO scores to test whether the new factors have predictive power beyond that of the classroom observation scores. The results are presented in Table 6.

Across specifications, teacher-centered instruction (Factor 3) negatively predicts value-added scores calculated using SAT-9. After controlling for average CLASS, FFT, and PLATO scores, the coefficients become even larger, suggesting that Factor 3 has extra predictive power and might capture teaching practices beyond those assessed in traditional classroom observations. Specifically, an increase of 1 standard deviation in Factor 3 is associated with a reduction of 0.041 on standardized value-added scores.

The results for the classroom management factor (Factor 1) and the interactive instruction factor (Factor 2) are less clear. Although the classroom management factor—which captures more time and talk dedicated to management concerns—has negative coefficients across all the specifications and both tests, none is significant, possibly due to a lack of power from the comparatively small sample size. In contrast, the interactive instruction factor positively predicts the state ELA value-added scores with a marginal significance, and the results are robust both with and without classroom controls. After controlling for classroom observation scores, these effects are no longer statistically significant, but the point estimates are similar. The small sample size does not allow us to tease out whether these results are due to a power issue or whether the classroom observations already capture most of the variation identified by Factor 2. Quite similarly, interactive instruction does not have significant associations with value-added scores using SAT-9 across specifications.

To show whether specific text-as-data metrics are driving the results, we ran a similar analysis

TABLE 6

Regression of VAMs on Predicted Factors

Factor	(1)	(2)	(3)	(4)	(5)	(6)
			State ELA			
Factor 1: classroom management	−0.009	−0.007	−0.007	−0.006	−0.005	−0.006
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Factor 2: interactive instruction	0.021†	0.020†	0.019	0.018	0.016	0.018
	(0.011)	(0.011)	(0.012)	(0.011)	(0.012)	(0.011)
Factor 3: teacher-centered instruction	−0.007	−0.006	−0.006	−0.004	−0.009	−0.004
	(0.010)	(0.009)	(0.009)	(0.009)	(0.010)	(0.009)
Observations	281	281	281	281	281	281
R^2	.029	.091	.091	.100	.101	.100
			SAT-9			
Factor 1: classroom management	−0.024	−0.025	−0.027	−0.024	−0.022	−0.022
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
Factor 2: interactive instruction	0.004	0.002	−0.004	−0.000	−0.007	−0.007
	(0.026)	(0.026)	(0.026)	(0.027)	(0.027)	(0.026)
Factor 3: teacher-centered instruction	−0.033†	−0.036*	−0.035†	−0.033†	−0.042*	−0.041*
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.019)
Observations	279	279	279	279	279	279
R^2	.024	.052	.057	.057	.069	.069
Class characteristics		X	X	X	X	X
CLASS average score			X			X
FFT average score				X		X
PLATO average score					X	X
District fixed effects	X	X	X	X	X	X
Grade fixed effects	X	X	X	X	X	X

Note. Robust standard errors are reported in parentheses. All factors are in standardized values. Class characteristics include percentage of students who are male, in special education, English language learners, Asian, Hispanic, and African American; average age; and average prior test scores in ELA and math. ELA = English language arts; SAT-9 = Stanford Achievement Test, Ninth Edition; CLASS = Classroom Assessment Scoring System; FFT = Framework for Teaching; PLATO = Protocol for Language Arts Teaching Observations; VAM = value-added measure.

†$p < .10$. *$p < .05$.

using each individual micro-measure of teaching, the results of which are reported in Supplementary Appendix Table E1 in the online version of the journal. Both the percentage of time teachers talk, as well as the use of "I," which together comprise the teacher-centered instruction factor (Factor 3), have significant and negative coefficients. Notably, although the teacher-centered instruction factor (Factor 3) does not show a significant association with state value-added scores, a higher proportion of teacher talk itself negatively predicts both SAT-9 value-added scores and state ELA value-added scores. These results suggest that a high proportion of teacher talk is negatively associated with student gains across different achievement measures. In addition, while none of the three factors identified through the factor analysis has a high loading on "language style matching," this measure independently shows positive and significant correlations with SAT-9 value-added scores.

Taken together, the analyses using the instructional factors and the individual teacher practice measures provide evidence of a negative association between teacher-centered classroom discourse and student achievement gains, particularly for SAT-9 value-added scores, which are supposed to measure higher order skills. The results also provide suggestive evidence that a more dialogic format may support student outcomes.

Again, given the small sample size, the estimates are not precise enough to provide a more definite conclusion. Moreover, this analysis is correlational instead of causal. There may be other omitted, unmeasured classroom practices or teacher behaviors driving the results, which prevents us from making strong cause-and-effect claims. However, the results are promising in terms of the alignment between the text-driven instructional factors and observation scores, as well as some consistent associations with value-added scores. Moreover, the results demonstrate the potential of this approach, which can be used at scale to capture a great deal of classroom data with a relatively low cost compared with human raters. If districts and schools invest in systems for data capture and analysis, these types of data could provide a helpful complement to the more sporadic observations typically conducted by principals, coaches, and other instructional support providers.

Cost-Effectiveness

So far, our analyses have demonstrated that text-as-data methods are a promising approach to measure teaching practices. These computer-generated measures align well with both measures from classroom observational protocols and are, in some cases, predictive of value-added scores. They also have reasonable reliabilities similar to measures from observation protocols. Given these promising results, a cost-effectiveness analysis would further help us to understand the advantage of an automated approach compared with using human raters. The cost-effectiveness might be especially relevant for school districts with fewer resources and a desire for an economical option for supplemental measures of teaching. Such information could prove instrumental for costly professional development efforts and other policies and systems designed to give teachers formative feedback on their instruction.

The biggest cost-saving aspect of a text-as-data method is that it does not rely on human labor to rate teaching practices. Most school districts rely on principals to do classroom observations or walk-throughs, a process that is time-consuming and burdensome and still results, on average, in only a few observations each school year, leading to broad conclusions about the characteristics and quality of someone's teaching based on one or two classroom visits (Cohen & Goldhaber, 2016; Grissom et al., 2013). Having exponentially more data from text-based systems could provide teachers with far more insight into their practice, and afford principals and coaches with a more comprehensive and multifaceted window into instructional quality in a classroom. Although we do not propose that text-based data systems replace human observers, we suggest that they could be an invaluable complement. Our analyses are premised on the fact that more observations—and more metrics—would improve both formative and summative uses of teaching performance data. To get more performance information from such systems, a district would either need *more* raters to conduct *more* observations or utilize a text-based data system, which provide additional information.

Because it is not straightforward to directly quantify such cost-saving using monetary terms, we focus our cost–benefit analysis on a scenario where districts hire expert raters to conduct classroom observation, such as those used for a time in Washington D.C.'s IMPACT evaluation system (Dee & Wyckoff, 2015). For simplicity, we focus on costs due to human resource expenses (i.e., raters and trainers) and do not account for infrastructure costs, such as purchasing observational protocols in conventional classroom observations or developing computer algorithms and installing cameras and microphones in a text-based system.

Based on our interview with a researcher and a staff member at the National Center for Teacher Effectiveness at Harvard University (H. Hill & S. Booth, email communication, March 6, 2020), the costs of using expert raters mainly include two parts: rater training and the rating process. The average fixed cost of training a rater is US$750, which includes 30 hours of training that costs US$25 per hour. To ensure reliable ratings, assuming we need to double rate 15% of the 976 videos in this study, we would need 10 raters (i.e., close to 112 videos per rater), and training alone would cost US$7,500. An hour-long video requires roughly 2 hours of a rater's time, and typically costs US$25 per hour for a rater. Thus, it would cost US$56,120 to rate all videos included in our study. The total cost of rater training and

ongoing rater support would total US$63,620. In contrast, once an algorithm is developed, the only cost of a text-as-data approach is from transcribing videos. In this study, we spent US$30 to transcribe each video and in total it costed US$29,280 to transcribe all the videos. As the technology of automated transcription continues to improve in accuracy, these costs will continue to drop. Thus, based on this simple back-of-envelope analysis, the minimum cost-saving from a text-as-data approach is 54% compared with a human-rater approach.

Discussion and Implications

Measuring and supporting teaching quality is a perennial topic in education policy research. For decades, classroom observations have contributed to our understanding of what "good teaching" looks like, and yet researchers and practitioners would benefit from new tools that could identify a broader and more expansive set of classroom features, improve measurement precision by collecting data from more lessons, reduce cost, and help teachers better align their practices with those associated with achievement gains for students. As a proof of concept, this study took a novel approach to measuring teaching quality, exploring the potential of text-as-data methods for creating automated and objective measures of classroom interactions and discourse. Using nearly 1,000 transcriptions of fourth- and fifth-grade ELA classes, we created six distinct measures, which can be reduced to a classroom management factor, an interactive instruction factor, and a teacher-centered instruction factor.

These three instructional factors are aligned in conceptually coherent ways with many of the domains and dimensions identified by the popular observation protocols CLASS, FFT, and PLATO, meaning that the text-as-data approach can detect classroom instructional practices that are consistent with professional assessments of teaching quality conducted by human raters. The findings from the factor analysis also provide new evidence on the teaching practices associated with student learning gains. Notably, the teacher-centered instruction factor negatively predicts teachers' value-added scores computed using SAT-9, suggesting the importance of students' active participation in classroom discourse for their development of higher order thinking skills. Moreover, this association is robust even after controlling for teachers' average CLASS, FFT, and PLATO scores, demonstrating that text-as-data methods have the potential to identify teaching practices that may be overlooked by current protocols.

To our knowledge, this study is the first to apply text-as-data methods to measuring multiple teacher practices and corroborate such measures by using both classroom observation scores and student learning outcomes. It certainly does not represent the last word on the subject, though, as the measures we created are far from sufficient to capture all aspects of effective teaching. There is also plenty of room to refine the methods we used and improve these measures. Nonetheless, this study demonstrates the potential of text-as-data methods to measure some aspects of teaching and suggests promising avenues for future research. In particular, we have only begun to explore the content of language; new dictionaries and methods such as neural network analysis can create far richer measures that are more closely linked to classroom content. Moreover, due to its small sample size, this study may not have enough power to identify important relationships between the constructs we developed and other measures of teaching. The specific grade levels, subject, and student population we examine also preclude us from generalizing the findings because, for example, classroom discourse may well look different in mathematics or in the primary grades. As one of the few studies to apply computational tools to education policy research, however, this study serves as a demonstration that the use of rich textual information and technology can inform critical education policy discussions. It is worth noting that all the MET data are based on video-based observations, while in practice, observations are generally done live, with in-person raters. Studies have suggested that the modality of observation can prove consequential for assessments of teachers, and as such, future research should also look at correlations between text-based metrics and those scored in live classroom observations (Casabianca et al., 2013).

Classroom observation is time-consuming for principals, instructional coaches, and other

school leaders (Grissom et al., 2013), and as a result, we often make consequential inferences about teachers and provide corresponding supports based on only one or two classroom observations (Cohen & Goldhaber, 2016). Given the resources districts allocate to professional development and supports for teachers, having systems to collect far more, and multifaceted, data could dramatically improve the precision and impacts of such improvement efforts. Formative assessment policies and programs like coaching have been shown to improve a range of student outcomes (e.g., Kraft et al., 2018), but the success of such policies rests on coach capacity and skills at detecting key features of instructional quality. Moreover, existing systems are predicated on the assumption that the features of instruction that "matter" for students are readily visible to trained observers. Given the volume of discourse and interactions that occur in busy classrooms, there may well be aspects of instruction that are difficult, if not impossible, to detect in this way. Our findings, which suggest additional utility of text-based data above and beyond information collected by three distinct classroom observation systems, indicate the potential for text-based data to enhance coach capacity in important ways that ultimately improve teacher and student outcomes. Of course, determining the realization of such benefits is outside the scope of this study, but they are important directions for future research.

An additional limitation of these analyses is our focus on language arts teaching. It is unclear from these data the degree to which or ways in which the psychometric properties detailed here would extend to other classrooms subjects. That said, discourse (Hess, 2002; O'Connor & Michaels, 2007; Walshaw & Anthony, 2008) and classroom management are key issues raised in teaching across content areas (Allen et al., 2011; Brophy, 2006). There is no conceptual reason to believe that text-as-data metrics would work more or less well in other subjects, though this, too, is an important empirical question.

Classroom observations are also time-consuming and resource intensive for researchers studying effective teaching (Kane & Staiger, 2012). Automated metrics, such as the ones we discuss here, could help mitigate these issues for both practice and research in quick, cost-effective ways. Researchers have spent the last two decades trying to identify the highest leverage practices associated with a range of student outcomes (e.g., Ball & Forzani, 2009; Pianta & Hamre, 2009). Unpredictable raters and relatively modest associations with achievement measures have marked these efforts. Using text-as-data methods to generate metrics could free up time and resources to conceptualizing the measures themselves and empirically testing the impacts of teaching practices on student outcomes. This renewed process of measurement building might speed up the process of searching for high-leverage practices and shed new light on the relationship between aspects of teaching and student outcomes.

Admittedly, districts and schools need to invest in the infrastructure that allows them to record, transcribe, and analyze classroom data before they can benefit from the proposed methods. However, despite the initial costs, computer algorithms that measure certain teaching practices like those in this study, once developed and validated by researchers, do not need additional efforts to ensure measurement quality, and can be applied to any classroom transcripts across settings with minimal to no additional requirements (Demszky et al., Under review).[17]

As such research matures, new tools that are based on computational techniques can be applied in practice to complement conventional classroom observations and provide teachers timely and informative feedback. Although districts and schools need to invest in the upfront infrastructure before they can benefit from the proposed methods, many of the automated metrics we present here should theoretically be readily interpretable by teachers. The "use of open-ended questions" and "percent of instructional time dedicated to classroom management" are constructs with which all teachers would be familiar. For school leaders, there is also enormous potential upside to having such automated metrics about instruction. Allocating time to observing and scoring can also limit the time for providing teachers feedback on their instruction, which might be most instrumental in driving improvement (Cohen et al., 2020; E. S. Taylor & Tyler, 2012). As such, school leaders would benefit from tools that provide accessible information about instruction, so principals and coaches could then focus their efforts on helping teachers

make sense of the information provided and identifying strategies for improvement.

Finally, we see potential for text-based metrics to provide teachers and school leaders with invaluable information about the distribution of talk across different students and groups of students in a classroom. Participatory equity is a huge concern, as many scholars have noted disparities in who participates in classroom discourse, as well as how they participate (Boaler, 2008; Langer-Osuna, 2011). Recent work has demonstrated the potential for observations to capture some more nuanced discourse patterns, including how students from different racial and ethnic groups are engaged by their teachers (Shah & Crespo, 2018). A related ethical issue is about the privacy of language data, a commonly shared concern in research and practice that involves the use of human language. Fortunately, as technology advances and automated speech recognition systems more readily detect different speakers but preserve privacy (Silva et al., 2020), we are optimistic that such data could provide teachers with helpful insight into participation patterns in their classrooms.

Of course, an important step in this would be the understanding of how principals and teachers perceive automated measures and respond to the information they provide. A plus of classroom observations—versus computer-generated VAMs—is their face validity among educators (Cohen & Goldhaber, 2016). We need to understand more about the degree to which teachers and principals will see these automated measures as valid in the same way they do traditional observation-based metrics. Relatedly, we also need work focused on how interpretable automated metrics are for teachers and whether they are able to leverage the information provided by such measures to improve their teaching. We hypothesize that many of these measures would be accessible for teachers, but it remains an empirical question. These are all key directions for future research, but we see the evidence presented here as an important first step, a proof of concept that it is feasible to generate automated and objective measures of teaching practices that align with student outcomes and conventional classroom observations using text-as-data methods.

Acknowledgments

The authors are grateful to Allison Atteberry, Eric Bettinger, Ben Domingue, Greg Duncan, Pam Grossman, Heather Hill, Helen Ladd, Susanna Loeb, Richard Murnane, Ann Porteus, and Sam Wineburg, for their helpful suggestions and comments. All opinions expressed are solely those of the authors.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The study received generous financial support from the National Academy of Education/Spencer Dissertation Fellowship, the Shultz Graduate Student Fellowship in Economic Policy from the Stanford Institute for Economic Policy Research, the Stanford Graduate School of Education Doctoral Student Award from the Technology for Equity in Learning Opportunities initiative, the Dissertation Support Grant from Stanford Graduate School of Education, and a dissertation grant from the Stanford Freeman Spogli Institute.

ORCID iD

Jing Liu https://orcid.org/0000-0002-9918-8642

Notes

1. Based on one estimate, districts spend on average US$18,000 per teacher every year on professional development (A. Jacob & McGovern, 2015).

2. Participating districts include the Charlotte-Mecklenburg Schools, the Dallas Independent Schools, the Denver Public Schools, the Hillsborough County Public Schools, the Memphis Public Schools, and the New York City Schools.

3. The Measures of Effective Teaching (MET) project sets out to improve researchers' abilities to make causal inferences regarding teacher effectiveness. To avoid issues arising from students' and teachers' sorting into classes, MET project researchers randomly assigned participating teachers to classrooms in each school, grade, and subject in the second year of the study. In constructing the sample for this study, we only included teachers who participated in the second-year randomization process so that we can leverage the

randomization in follow-up studies and reduce costs for transcription.

4. For subject-matter specialists (mainly sixth- to ninth-grade teachers and a few fourth- to fifth-grade teachers), the MET project recorded four videos for each teacher. In the first year of the project, two class sections taught by each subject-matter specialist were recorded on two different days; in the second year, one class section was recorded on four different days.

5. For the approximately 25% of videos lasting less than 30 minutes, we transcribed the entire class time.

6. Due to data availability, the value-added scores are based on the performance of only those students who participated in the MET project rather than all students in the district.

7. Compared with Danielson's Framework for Teaching (FFT), another widely used protocol that can be applied across subject domains, Classroom Assessment Scoring System (CLASS) also has more empirical evidence to support its use for evaluation (e.g., Pianta et al., 2002) and improvement efforts (e.g., Allen et al., 2011).

8. For example, using the time stamps and number of words from the transcripts, we were able to measure the back-and-forth exchanges between the teacher and students, which are also assessed by the "quality of feedback" dimension of CLASS.

9. Linguistic Inquiry and Word Count (LIWC) identifies words associated with certain psychological dimensions, by using human judgment and extensive validation. For example, based on LIWC's official documentation, "LIWC measures the degree to which texts reveal interests in power, status, and dominance using its Power dictionary. By definition, someone who is concerned with power is more likely to be sizing other people up in terms of their relative status. Such a person will be more likely to use words such as boss, underling, president, Dr., strong, and poor when compared with someone who simply doesn't care about power and status." https://liwc.wpengine.com/how-it-works/

10. These dimensions comprised four broad categories, including summary language variables (e.g., analytical thinking, authentic, emotional tone), linguistic dimensions (e.g., articles, prepositions, auxiliary verbs), other grammar (e.g., numbers, quantifiers, comparisons), and psychological processes (e.g., affective processes, social processes, cognitive processes).

11. Supervised methods require a training data set with human labels that can "teach" the computer algorithm features of the construct researchers try to measure. In contrast, unsupervised methods are fully automated and do not require human input other than the programming process.

12. These two categories are originally developed to categorize writing styles (Pennebaker et al., 2014).

13. Lasso serves as a method of feature selection because it adds a penalty to small coefficients so that many coefficients are reduced to zero. For detailed explanation on how Lasso works, see Hastie et al. (2009).

14. To test whether accounting for lesson-topic variances would change our factor structures, we follow McCaffrey et al. (2015) by fitting a separate lesson-topic fixed effect model for each measure, and then subtract the estimated lesson-topic fixed effects from the raw scores. We then use the adjusted scores to fit an exploratory factor analysis (EFA) model and correlate these factors with those without adjusting for lesson-topic fixed effects. The correlations for the three factors are .995, .998, and .992.

15. One important indicator for social talk is using non-first-person-singular personal pronouns.

16. "Instructional Dialogue captures the purposeful use of content-focused discussion among teachers and students that is cumulative, with the teacher supporting students to chain ideas together in ways that lead to deeper understanding of content. Students take an active role in these dialogues and both the teacher and students" (Pianta et al., 2012).

17. The minimum requirements for using a fully developed computer algorithm that measures teaching practice should be precisely transcribed classroom talks, with clear distinction on the speakers (i.e., teachers and students). With text-as-data methods, more language data would yield more precise inferences. Thus, we would envision a lesson should be at least 20 to 30 minutes long to produce useful measures.

References

Allen, J. P., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher–student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system-secondary. *School Psychology Review*, *42*(1), 76–98.

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.

Althoff, T., Clark, K., & Leskovec, J. (2016). *Large-scale analysis of counseling conversations: An application of natural language processing to mental health*. http://arxiv.org/abs/1605.04462

Applebee, A. N., Langer, J. A., Nystrand, M., & Gamoran, A. (2003). Discussion-based approaches

to developing understanding: Classroom instruction and student performance in middle and high school English. *American Educational Research Journal*, *40*(3), 685–730.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *Quarterly Journal of Economics*, *131*(3), 1415–1453.

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

Beck, I. L., & McKeown, M. G. (2001). Text talk: Capturing the benefits of read-aloud experiences for young children. *The Reading Teacher*, *55*(1), 10–20.

Bell, C. A., Jones, N. D., Qi, Y., & Lewis, J. M. (2018). Strategies for assessing classroom teaching: Examining administrator thinking as validity evidence. *Educational Assessment*, *23*(4), 229–249.

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., Mccaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (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). Jossey-Bass.

Bettinger, E., Liu, J., & Loeb, S. (2016). Connections matter: How interactive peers affect students in online college courses. *Journal of Policy Analysis and Management*, *35*(4), 932–954.

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review*, *48*, 16–29.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Boaler, J. (2008). Promoting "relational equity" and high mathematics achievement through an innovative mixed-ability approach. *British Educational Research Journal*, *34*(2), 167–194.

Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist*, *41*(10), 1069–1077.

Brophy, J. (2006). History of research on classroom management. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 17–43). Lawrence Erlbaum.

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

Cabell, S. Q., Justice, L. M., McGinty, A. S., DeCoster, J., & Forston, L. D. (2015). Teacher–child conversations in preschool classrooms: Contributions to children's vocabulary development. *Early Childhood Research Quarterly*, *30*, 80–92.

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal*, *55*(6), 1233–1267.

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.

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly*, *27*(3), 529–542.

Cazden, C. B. (2001). *Classroom discourse: The language of teaching and learning* (2nd ed.). Heinemann. (Original work published 1988)

Chinn, C. A., Anderson, R. C., & Waggoner, M. A. (2001). Patterns of discourse in two kinds of literature discussion. *Reading Research Quarterly*, *36*(4), 378–411.

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

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*(6), 378–387.

Cohen, J., Hutt, E., Berlin, R., & Wiseman, E. (2020). The change we cannot see: Instructional quality and classroom observation in the era of common core. *Educational Policy*, *1*, 27. https://doi.org/10.1177/0895904820951114

Cohen, J., Ruzek, E., & Sandilos, L. (2018). Does teaching quality cross subjects? Understanding consistency in elementary teacher practice across subjects. *AERA Open*, *4*(3), https://doi.org/10.1177/2332858418794492

Danielson, C. (2011). Evaluations that help teachers learn. *Educational Leadership*, *68*(4), 35–39.

Danielson, C. (2013). *The framework for teacher evaluation instrument* (2013 ed.). The Danielson Group.

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

Demszky, D., Liu, J., Cohen, J., Hill, H., Mancenido, Z., Jurafsky, D., & Hashimoto, T. (Under review). *Measuring conversational uptake: A case study on student-teacher interactions* [Working paper].

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

Gentzkow, M., Kelly, B., & Taddy, M. (2017). *Text as data* (NBER Working Paper No. 23276). National Bureau of Economic Research.

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

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), 1–32.

Goe, L., Biggers, K., & Croft, A. (2012). *Linking teacher evaluation to professional development: Focusing on improving teaching and learning* [Research & policy brief]. National Comprehensive Center for Teacher Quality.

Goldin-Meadow, S., Levine, S. C., Hedges, L. V., Huttenlocher, J., Raudenbush, S. W., & Small, S. L. (2014). New evidence about language and cognitive development based on a longitudinal study: Hypotheses for intervention. *American Psychologist*, *69*(6), 588–599.

Graham, S., & Harris, K. R. (1993). Self-regulated strategy development: Helping students with learning problems develop as writers. *The Elementary School Journal*, *94*(2), 169–181.

Greenleaf, C., Schoenbach, R., Cziko, C., & Mueller, F. (2001). Apprenticing adolescent readers to academic literacy. *Harvard Educational Review*, *71*(1), 79–130.

Grimmer, J. (2015). We are all social scientists now: How big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, *48*(1), 80–83.

Grissom, J. A., Loeb, S., & Master, B. (2013). Effective instructional time use for school leaders: Longitudinal evidence from observations of principals. *Educational Researcher*, *42*(8), 433–444.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value-added on multiple types of assessment. *Educational Researcher*, *43*(6), 293–303.

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.

Hamre, B. K., & Pianta, R. C. (2001). Early teacher–child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development*, *72*(2), 625–638.

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). *Building a science of classrooms: Application of the CLASS framework in over 4,000 U.S. early childhood and elementary classrooms*. Foundation for Child Development.

Hassinger-Das, B., Toub, T. S., Hirsh-Pasek, K., & Golinkoff, R. M. (2017). A matter of principle: Applying language science to the classroom and beyond. *Translational Issues in Psychological Science*, *3*(1), 5–18.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Herbel-Eisenmann, B., Drake, C., & Cirillo, M. (2009). "Muddying the clear waters": Teachers' take-up of the linguistic idea of revoicing. *Teaching and Teacher Education*, *25*(2), 268–277.

Hess, D. E. (2002). Discussing controversial public issues in secondary social studies classrooms: Learning from skilled teachers. *Theory & Research in Social Education*, *30*(1), 10–41.

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., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, *83*(2), 371–384.

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

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education*, *17*(3), 207–219.

Howe, C., Hennessy, S., Mercer, N., Vrikki, M., & Wheatley, L. (2019). Teacher–student dialogue during classroom teaching: Does it really impact on student outcomes? *Journal of the Learning Sciences*, *28*(4–5), 462–512.

Huttenlocher, J. (1998). Language input and language growth. *Preventive Medicine*, *27*(2), 195–199.

Jacob, A., & McGovern, K. (2015). *The mirage: Confronting the hard truth about our quest for teacher development*. TNTP.

Joe, J. N., McClellan, C. A., & Holtzman, S. L. (2015). 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). John Wiley.

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

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*(3), 587–613.

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, *47*, 451–464.

Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research*, *86*(4), 945–980.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547–588.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, *46*(5), 234–249.

Langer-Osuna, J. M. (2011). How Brianna became bossy and Kofi came out smart: Understanding the trajectories of identity and engagement for two group leaders in a project-based mathematics classroom. *Canadian Journal of Science, Mathematics and Technology Education*, *11*(3), 207–225.

Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, *31*(1), 61–95.

Lugini, L. (2020). *Analysis of collaborative argumentation in text-based classroom discussions* [Doctoral dissertation]. Department of Computer Science, University of Pittsburgh.

Lugini, L., Olshefski, C., Singh, R., Litman, D., & Godley, A. (2020). Discussion tracker: Supporting teacher learning about students' collaborative argumentation in high school classrooms. In *Proceedings of the 28th international conference on computational linguistics*. https://www.aclweb.org/anthology/2020.coling-demos.10.pdf

Mantzicopoulos, P., French, B. F., Patrick, H., Watson, J. S., & Ahn, I. (2018). The stability of kindergarten teachers' effectiveness: A generalizability study comparing the framework for teaching and the classroom assessment scoring system. *Educational Assessment*, *23*(1), 24–46.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., & Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*, *79*(3), 732–749.

McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2015). Uncovering multivariate structure in classroom observations in the presence of rater errors. *Educational Measurement: Issues and Practice*, *34*(2), 34–46.

McFarland, D. A., Jurafsky, D., & Rawlings, C. (2013). Making the connection: Social bonding in courtship situations. *American Journal of Sociology*, *118*(6), 1596–1649.

Mehan, H. (1979). "What time is it, Denise?" Asking known information questions in classroom discourse. *Theory Into Practice*, *18*(4), 285–294.

Meyer, J. P., Cash, A. H., & Mashburn, A. (2011). Occasions and the reliability of classroom observations: Alternative conceptualizations and methods of analysis. *Educational Assessment*, *16*(4), 227–243.

Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education*, *27*(4), 283–297.

Miner, A. S., Haque, A., Fries, J. A., Fleming, S. L., Wilfley, D. E., Wilson, G. T., & Fei-Fei, L. (2020). Assessing the accuracy of automatic speech recognition for psychotherapy. *Npj Digital Medicine*, *3*(1), Article 82.

Muijs, D., Reynolds, D., Sammons, P., Kyriakides, L., Creemers, B. P., & Teddlie, C. (2018). Assessing individual lessons using a generic teacher observation instrument: How useful is the International System for Teacher Observation and Feedback (ISTOF)? *ZDM*, *50*(3), 395–406.

Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M. E. (2015, May). Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition* (Vol. 1, pp. 1–6).

Institute of Electrical and Electronics Engineers. https://ieeexplore.ieee.org/document/7163127

Newmann, F. M., Lopez, G., & Bryk, A. S. (1998). *The quality of intellectual work in Chicago schools: A baseline report*. Consortium on Chicago School Research.

Niculae, V., & Danescu-Niculescu-Mizil, C. (2016). *Conversational markers of constructive discussions*. http://arxiv.org/abs/1604.07407

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, *21*(4), 337–360.

Nystrand, M. (2006). Research on the role of classroom discourse as it affects reading comprehension. *Research in the Teaching of English*, *40*(4), 392–412.

Nystrand, M., & Gamoran, A. (1991). Instructional discourse, student engagement, and literature achievement. *Research in the Teaching of English*, *25*, 261–290.

Nystrand, M., Gamoran, A., Kachur, R., & Prendergast, C. (1997). *Opening dialogue*. Teachers College Press.

O'Connor, C., & Michaels, S. (2007). When is dialogue "dialogic"? *Human Development*, *50*(5), 275–285.

Palincsar, A. S., & Brown, D. A. (1987). Enhancing instructional time through attention to metacognition. *Journal of Learning Disabilities*, *20*(2), 66–75.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, *48*(1), 163–193.

Park, Y. S., Chen, J., & Holtzman, S. L. (2015). Evaluating efforts to minimize rater bias in scoring classroom observations. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 381–414). Jossey-Bass.

Patrick, H., Mantzicopoulos, P., & French, B. F. (2020). The predictive validity of classroom observations: Do teachers' framework for teaching scores predict kindergarteners' achievement and motivation? *American Educational Research Journal*, *57*(5), 2021–2058.

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. Bloomsbury.

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PLOS ONE*, *9*(12), e115844.

Penner, E. K., Rochmes, J., Liu, J., Solanki, S. M., & Loeb, S. (2019). Differing views of equity: How prospective educators perceive their role in closing achievement gaps. *RSF: The Russell SAGE Foundation Journal of the Social Sciences*, *5*(3), 103–127.

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., & Mintz, S. (2012). *Classroom assessment scoring system (CLASS) manual, secondary*. Teachstone.

Pianta, R. C., La Paro, K. M., Payne, C., Cox, M. J., & Bradley, R. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *The Elementary School Journal*, *102*(3), 225–238.

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

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

Shah, N., & Crespo, S. (2018). Cultural narratives and status hierarchies: Tools for identifying and disrupting inequity in mathematics classroom interaction. In R. Hunter, M. Civil, B. Herbel-Eisenmann, N. Planas, & D. Wagner (Eds.), *Mathematical discourse that breaks barriers and creates space for marginalized learners* (pp. 23–37). Brill Sense.

Silva, P., Gonçalves, C., Godinho, C., Antunes, N., & Curado, M. (2020, July). Using NLP and machine learning to detect data privacy violations. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 972–977). Institute of Electrical and Electronics Engineers. https://ieeexplore.ieee.org/document/9162683

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.

Song, L., Spier, E. T., & Tamis-Lemonda, C. S. (2014). Reciprocal influences between maternal language and children's language and cognitive development in low-income families. *Journal of Child Language*, *41*(2), 305–326.

Sun, M., Liu, J., Zhu, J., & LeClair, Z. (2019). Using a text-as-data approach to understand reform processes: A deep exploration of school improvement strategies. *Educational Evaluation and Policy Analysis*, *41*(4), 510–536.

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016). Winning arguments: Interaction

dynamics and persuasion strategies in good-faith online discussions. In J. Bourdeau, J. A. Hendler, R. N. Nkambou, I. Horrocks, & B. Y. Zhao (Eds.), *WWW '16: Proceedings of the 25th International Conference on the World Wide Web* (pp. 613–624). International World Wide Web Conferences Steering Committee.

Taylor, B. M., Peterson, D. S., Pearson, P. D., & Rodriguez, M. C. (2002). Looking inside classrooms: Reflecting on the "how" as well as the "what" in effective reading instruction. *The Reading Teacher*, *56*(3), 270–279.

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, *102*(7), 3628–3651.

Tharp, R. G., & Gallimore, R. (1991). *Rousing minds to life: Teaching, learning, and schooling in social context*. Cambridge University Press.

Walshaw, M., & Anthony, G. (2008). The teacher's role in classroom discourse: A review of recent research into mathematics classrooms. *Review of Educational Research*, *78*(3), 516–551.

Wang, Z., Miller, K., & Cortina, K. (2013). Using the LENA in teacher training: Promoting student involvement through automated feedback. *Unterrichtswissenschaft*, *41*(4), 290–305.

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

White, M., & Rowan, B. (2012). *A user guide to the "core study" data files available to MET early career grantees*. Inter-University Consortium for Political and Social Research, University of Michigan.

Authors

JING LIU is an assistant professor of education policy at the University of Maryland, College Park. He studies student absenteeism, exclusionary discipline, and educators' labor market, with a special interest in the intersection of data science and education policy.

JULIE COHEN is an associate professor of curriculum, instruction and special education in the School of Education and Human Development at the University of Virginia. She studies teachers and teaching, with a focus on policies that support the development of effective instructional practices.

Manuscript received June 7, 2020
First revision received March 4, 2021
Accepted March 16, 2021