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

Jing Liu
University of Maryland College Park

Julie Cohen
University of Virginia

Valid and reliable measurements of teaching quality facilitate school-level decision-making and policies pertaining to teachers, but conventional classroom observations are costly, prone to rater bias, and hard to implement at scale. Using nearly 1,000 word-to-word transcriptions of 4th- and 5th-grade English language arts classes, we apply novel text-as-data methods to develop automated, objective measures of teaching to complement classroom observations. 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.

VERSION: June 2020

Measuring Teaching Practices at Scale:

A Novel Application of Text-as-Data Methods[*]

Jing Liu[†]

Julie Cohen[‡]

Abstract

Valid and reliable measurements of teaching quality facilitate school-level decision-making and policies pertaining to teachers, but conventional classroom observations are costly, prone to rater bias, and hard to implement at scale. Using nearly 1,000 word-to-word transcriptions of 4th- and 5th-grade English language arts classes, we apply novel text-as-data methods to develop automated, objective measures of teaching to complement classroom observations. 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.

Keywords: teaching quality; K-12 education; classroom observation; text as data.

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, particularly unreliable and biased raters (Bell, et al., 2018; Ho & Kane, 2013; Kraft & Gilmour, 2017)

Recent advances in computational linguistic methods offer a potentially transformative solution to the issue of biased, unreliable, and expensive raters. It is challenging for human raters to keep track of multiple teacher practices simultaneously and score those practices 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 classroom videos and provide automated assessments of specific features of high-quality teaching. In addition to saving districts costs in employing a cadre of skilled and impartial "masters raters", such as those used for a time in Washington DC's IMPACT evaluation system (Dee & Wyckoff, 2015), or relying on principals who are notoriously biased (Bell et al., 2018) and time-strapped (Grissom, Loeb, & Master, 2013), such methods may also provide an opportunity for teachers to

receive immediate feedback on their work that is perceived as less-threatening and more objective than that provided by a colleague or school leader (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, Miller, and Cortina (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 student outcomes, generally seen as an essential feature for any tool used for teacher evaluation (Hill et al., 2008; Grossman et al., 2013; Kane & Staiger, 2012).

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 beneficial teaching practices that may not be easily detected by human raters.

Specifically, we answered the following three research questions.

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

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

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 4th- and 5th-grade English language arts classrooms collected as part of the Measures of Effective Teaching (MET) project. To obtain precise time stamps for each utterance and differentiate the identities of the speakers, we hired professional transcribers instead of using automatic speech recognition (cf. Kelly et al., 2018). 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 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).

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 and 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 (SAT-9), which assesses cognitively complex learning outcomes. This result is robust even after controlling for teachers' average CLASS and PLATO scores, suggesting that this instructional factor is not fully captured by these classroom observation tools. Our analyses also provide some evidence that interactive instruction is associated with student outcomes.

The rest of the paper proceeds as follows. First, we introduce the theoretical and empirical literature on the relationship between teacher classroom practices and student outcomes. We then provide an overview of the MET project from which we draw our data, as well the specifics of our video transcription process and our analytic sample. We present the measures we created from the transcriptions, discuss their reliability, and provide a factor analysis to show what latent instructional factors these metrics capture. Using the extracted factors, we show the correlations between those computer-generated measures and ratings from classroom observations and VAMs. Finally, we conduct a cost effectiveness analysis to compare our proposed approach and the approach of hiring expert raters, and discuss the significance of these findings and their implications.

Background

Measures of Teaching and Teacher Quality

Teachers vary a lot in terms of their impact, making them one of, if not *the* biggest within-school determinant of student outcomes (Rivkin, Hanushek, & Kain, 2005). As a result, the last three decades have witnessed a great increase in the study of different tools for evaluating teacher performance. Although controversial in both methodology and application (Papay, 2011; Rothstein, 2010), teacher value-added models (VAMs) have advanced our understanding of the importance of teachers in relation to student learning outcomes, as well as

longer term life outcomes (Chetty, Friedman, & Rockoff, 2014). Multiple studies have shown that for students in upper elementary schools, having a teacher with value-added scores one standard deviation above average results in student test scores that are about 0.1–0.2 standard deviations above average. (For a review, see Koedel, Mihaly, & Rockoff, 2015.) However, it has been difficult to find systematic predictors of teachers' value-added measures. Studies have found only weak links between value-added and observable teacher characteristics, such as education and years of experience（Clotfelter, Ladd, & Vigdor, 2007; Rivkin, Hanushek, & Kane, 2005）, though recent work using more nuanced measures of teachers' knowledge and dispositions is somewhat more promising (Jacob et al., 2018).

Another strand of 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 on formal teacher evaluation (Cohen, 2015; Ball & Forzani, 2009), but also on identifying observed needs for targeted professional development efforts (Allen et al., 2011; Kennedy, 2016).

The concept of measuring observable teaching practices, and in some cases linking those observation ratings to student outcomes, is not new. The process-product model—which posits that discrete teacher behaviors—classroom "processes"—can benefit student learning—the "product"—boomed in the 1960s and 1970s (Brophy & Good, 1986; Gage & Needels, 1989). For decades, scholars have examined many low-inference observations of teacher behaviors, including the "stating of a clear learning objective" or the provision of adequate "wait time" after asking a question (Brophy, 1986; Rowe, 1986). Principals correspondingly focused classroom "walk-throughs" on these discrete and readily observable behaviors (Cohen &

Goldhaber, 2016). Over time, those walk throughs with binary checklists have evolved into more rigorous and nuanced observations using instruments that often comprise several, broader domains of teaching practices (e.g., instructional support) and include a point scale (e.g., 5 or 7 points) referenced to low-, medium-, and high-performance levels (Danielson, 2013, Marzano et al.,2001).

The process-product model continues to be the conceptual basis for much contemporary research on the effects of teaching (e.g., Kane & Staiger, 2012), though it is not without its critics (Fenstermacher & Richardson, 2005). Current, more detailed observation protocols, have also yielded some new 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. For example, using MET project data, Kane and Staiger (2012) found correlations between scores from the observation protocols and value-added scores ranging from .12 to .34. Using data from New York City, Grossman and her colleagues (2013) found that teachers in the top quartile of value-added scores had higher scores on a few of the elements of PLATO and CLASS than their bottom-quartile peers did. 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 deviations. In contrast, Kane et al. (2010) 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 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.

Though 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 limits the set of teaching practices they measure. Some instruments,

such as PLATO and the Mathematical Quality of Instruction (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, Ball, & Schilling, 2008). Others, such as CLASS and

Danielson's (2013) Framework for Teaching (FFT), emphasize aspects of "good teaching" that

cut across subjects. However, CLASS draws heavily from developmental theory, emphasizing

the warmth and positivity teacher–student interactions (Hamre & Pianta, 2001). FFT, by contrast,

builds upon constructivist learning theory, privileging teachers' questioning techniques and

students' intellectual engagement (Danielson, 2011).

Second, instrument developers and the school leaders who use observation tools struggle

to clearly and consistently define what constitutes *high-level demonstration* of this practice. For

example, "regard for student perspectives" (Pianta, Hamre, and Mintz, 2012) might be a teaching 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-level" 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 of teaching.

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, Charalambous, & Kraft, 2012; Park, Chen, & Holtzman, 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 (Weisberg et al., 2009; Kraft & Gilmour, 2017). The research has not surfaced clear strategies for mitigating rater bias. 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.

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, Kelly, & Taddy, 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 questions. For example, 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), change someone's opinion by forming persuasive arguments (Tan. et al., 2016), or productively address issues related to mental illness (Althoff, Clark, & Leskovec, 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, Liu, and Loeb, 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 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 teacher does not have a specific answer in mind, or has an answer predetermined by the teacher. Their computer-coded authenticity achieves reasonably high correlations with coding from human raters.

The development of text-as-data methods provides an unprecedented opportunity to extend prior research on teaching and learning in the classroom, enabling researchers to address

some of the inherent weaknesses of observation protocols and shed new light on how teaching practices impact student outcomes. First, text-as-data methods can save researchers precious time and resources by doing away with expensive trainings, ongoing calibration, and inherent biases associated with humans scoring multi-faceted and complex classroom instruction. Moreover, 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.

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 of their time so that they can focus on giving feedback to teachers.

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 myriad 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 the current study come from the Bill and Melinda Gates Foundation funded Measures of Effective Teaching (MET) project, which is, to date, the largest research project in

the United States on K-12 teacher effectiveness. More than 2,500 4th- through 9th-grade teachers in 317 schools across six districts participated in the MET project over a two-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.[1]

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 its dataset 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.[2]

The current study focuses on 4th- and 5th-grade English language arts (ELA) classrooms and teachers. We chose to focus on ELA 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, Anderson, & Waggoner, 2001; Nystrand & Gamoran, 1991;).

[Table 1 here]

Table 1 describes the sample, which includes a racially and socio-economically student body: 43% of students are eligible for free or reduced-price lunch, 42% are African American,

---

[1] 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.

[2] The 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 the current 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.

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 US. On average, teachers have worked in their current district for six years, and 46% have a master's degree or higher.

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.[3] 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 4th- and 5th-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 (or, for the approximately 25% of videos lasting less than

---

[3] For subject-matter specialists (mainly 6th–9th grade teachers and a few 4th–5th 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.

30 minutes, the entire class time) of all four videos for each teacher from the first year. 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 complete 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, teacher–student proximity, 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. The following is an example of the transcription data:

Student A: [00:00:02] Tell you the definition of cause? [00:00:03]

Teacher: [00:00:03] Yeah, what's the definition of cause, what does it mean? [00:00:05]

Student A: [00:00:06] It's like something that happens and that the something that happens and the cause is what happens. [00:00:13]

Teacher: [00:00:16] Okay, you said the effect, the cause is why something happens. Is that what you're saying? And the effect is why something happens. Or the effect is what happens, the cause is why it happens. Okay, keep that in mind as we watch this short video. You need to listen, okay, because it's going to talk to you. [00:00:34]

[Silence]

Teacher: [00:00:41] Okay, so keep that in mind what she's saying. Cause is why something happens, the effect is what actually happens. [00:00:43]

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) of the interactions.

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 (Grossman et al., 2015; Grossman et al., 2014). For the current study, we used value-added scores from the MET project database. [4]

The MET project data also contain multiple observational measures of teacher practices. We focused on measures generated from CLASS and PLATO. CLASS measures aspects of classrooms that are hypothesized to be generalizable across subjects based on developmental psychology theory. 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. 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. Compared with Danielson's FFT, another widely used protocol that can be applied across subject domains, 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). We 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, Pianta, Mashburn, & Downer, 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

---

[4] 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.

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, Carneiro, Cruz-Aguayo, & Schady, 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).

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, 2002; Graham & Harris, 1994; Greenleaf et al., 2001; Palinscar & Brown, 1987) and providing challenging disciplinary tasks in which students are expected to engage in the majority of the intellectual work (Newmann, Lopez, & Bryk, 1998). Several studies have found that ELA teachers with higher value-added scores perform better on multiple dimensions in PLATO (Grossman et al., 2014; Grossman et al., 2013).

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 potentially beneficial teacher practices that are more nuanced and fine-grained than those captured in current observation protocols. The first focuses on patterns of classroom 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, etc.) 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., 2015), which counts the percentage of words that reflect different emotions, thinking styles, and social concerns of a given utterance.[5] LIWC is widely used in computational linguistic analyses and covers 93 linguistic dimensions,[6] 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.[7] 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 include language coordination, questioning, and the allocation of time between academic content and routine, of which we detail the definition below.

---

[5] LIWC identifies words associated with certain psychological dimensions, by using human judgement 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/

[6] These dimensions are comprised of 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 process, social processes, cognitive processes).

[7] Supervised methods require a training dataset 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.

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 RQ 3. 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 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, O'Connor, & Resnick, 2008; Taylor et al., 2002). Cazden (1988) 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 (Tharp & Gallimore, 1991; Nystrand 1997). A sociocognitive view supports the idea that students need to be active agents in classroom interactions in order 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; 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 in order 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, Jurafsky, & Rawlings, 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's 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 allocate 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; McGinty et al., 2012; 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,[8] 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, y'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 (Collins 1982; Nystrand 1997; Herbel-Eisenmann et al. 2009). 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

---

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

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. (For examples of these words, see Appendix A.) 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 the current 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.

$$\text{LSM} = 1 - [\frac{|preps_t - preps_s|}{preps_t - preps_s + 0.0001}]$$

Language style matching is computed using the formula above, where $preps_t$ represents the percentage of prepositions used by the teacher for a given class and $preps_s$ represents the percentage used by students; 0.0001 is a constant that helps to avoid a zero denominator.

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, Anderson, and Waggoner (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 either procedural questions

(e.g., "How many pages do we need to read?"), rhetorical question, "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 teaching experience and who

are currently education researchers first hand-labeled 600 randomly selected questions from the

set of questions in the data, which serve as a "training" dataset. As there are many features that

can predict whether a question is open-ended or not, conventional regression-based prediction

methods are infeasible, since 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.[9] We

used a five-fold Lasso procedure to "learn" from this training sample and then make predictions

for the rest of the questions. Figure 1 shows the most predictive words for open-ended questions

and non-open-ended questions. Appendix B provides additional details on this method.

[Figure 1 here]

---

[9] 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, Tibshirani, and Friedman (2009).

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 (Pianta & Hamre, 2009; Gill et al., 2016).

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. As Figures 2 and 3 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). Appendix C describes the details of this method.

[Figures 2 and 3 here]

RQ2: What are the psychometric properties of the computer-generated measures on 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 analytic 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 of these measures show a significant amount of variation, suggesting that classroom discourse patterns vary substantially across the classrooms in the MET study.

[Table 2 here]

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"). 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 from individual raters' biases. Student composition is another possible source of bias in that observations may be sensistive 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 in addition to the first to separate out the effects of teaching different cohorts of students.

[Table 3 here]

We calculated the reliability of each measure by running a cross-classified multilevel model that decomposes the variance of each teacher practice as a teacher part, a lesson-topic part, and an error part. 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* we created 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 of 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 the factor loadings. The rotated factor

structure is not substantially different from the non-rotated factor structure, so we focus on the rotated factor loadings for ease of interpretation.

[Table 4 here]

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 routine. Teachers with high scores on this factor spend more of their language managing student disruptions, putting students into groups, and performing other non-instructional activities. They are also more likely to attentionally target students (i.e., using "you") and engage in social talk that are more narrative and intuitive in nature in contrast to analytical language.[10] 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. By interpreting the loadings for each factor, we find that the first and third factors tend to suggest less desirable teaching

---

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

practices, while the second factor is more positive. 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. 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[11] 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 as these two factors signal less desirable teaching practices and higher values in observation protocols suggest teachers do better on a certain dimension. As discussed above, CLASS is conceptually the closest observation tool to the current study because of its focus on teacher-student interactions, and PLATO is an ELA-specific protocol focused more on the substantive content of ELA teaching. Thus, using both CLASS and PLATO can generate a

---

[11] "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, Hamre, and Mintz, 2012).

deeper understanding of the instructional factors we developed. We correlated each of the three new factors with each of the domains of CLASS and PLATO.

[Table 5 here]

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 classroom organization management under both CLASS ($r = -0.28, p < .01$) and PLATO ($r = -0.18, p < .01$). These 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 some positive teaching practices, such as the domains of emotional support and instructional support under CLASS and classroom discourse under PLATO. These correlations highlight how the new measures can shed light on the micro-behaviors that correspond with the more holistic measures in the observation protocols.

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 ($r = 0.259, p < .01$) and PLATO scale for classroom discourse ($r = -0.183, p < .01$) under PLATO, both of which capture a similar construct. Instructional dialogue focuses on teachers' use of questioning and discussion to guide and prompt students' understanding, and classroom discourse 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.

While Factor 3, the teacher-centered instruction factor, has only one significant correlation with a dimension called instructional dialogue under CLASS, it is correlated with most of PLATO's dimensions. Similar to Factor 1, Factor 3 arguably captures a mix of more and less desirable practices, as defined by the two observation protocols. Factor 3 has a relatively strong positive correlations with modeling ($r = 0.325$, $p<0.01$) and with strategy use and instruction ($r=0.310$, $p<0.01$), which measure the degree to which teachers' provide students with detailed instruction about an acdemic process or skill, perhaps contributing to higher levels of "teacher talk." However, the factor's negative correlations with instructional dialogue (CLASS) and intellectual challenge (PLATO) may also indicate an overemphasis on lecture and an insufficient emphasis on inviting students to be part of academic discourse.

Overall, the three instructional factors we have identified align at least to some extent with observation scores from CLASS and PLATO. The comparison between these factors and the two observation protocols largely confirms that the factors capture a classroom management dimension, an interactive instruction format, and a teacher-centered instruction format, respectively. Since the domains and dimensions under each protocol are also interrelated, each factor is correlated with multiple domains and dimensions, highlighting that the micro-behaviors that compose these domains and dimensions overlap.

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 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.

$$VAM_{tdg} = \beta Practice_{tdg} + \theta_d + \tau_g + \varepsilon_{tdg}$$

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 in order 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 and PLATO scores in order to test whether the new factors have predictive power beyond that of the classroom observation scores. The results are presented in Table 6.

[Table 6 here]

Across specifications, teacher-centered instruction (Factor 3) negatively predicts value-added scores calculated using SAT-9. After controlling for average CLASS and PLATO scores, the coefficients become even larger, suggesting that Factor 3 has extra predictive power and might capture teaching practices beyond that of classroom observations. Specifically, an increase of one 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 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.

[Table 7 here]

In order to show whether specific computer-generated teacher practices are driving the results, we ran a similar analysis using each individual micro-measure of teaching, the results of which are reported in Table 7. 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 does not show a significant association with state ELA 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 achievement gains. Although the individual measure "language style matching" independently shows positive and significant correlations with SAT-9 value-added scores, this relationship is not reflected in the factor analyses, mainly because none of the three factors has a high loading on this behavior.

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 with relatively low cost.

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 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 measuring teacher practices and providing them feedback.

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 walkthroughs, a process that is time consuming and burdensome (Grissom, Loeb, and Master, 2013). A text-as-data approach can release principals' precious time to focus on more productive aspects of school management. Since 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 DC's IMPACT evaluation system (Dee & Wyckoff, 2015).

Based on our interview with a researcher and a staff member at the National Center for

Teacher Effectiveness at Harvard University (H. Hill and 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 $750, which includes 30 hours of

training that costs $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),

training alone would cost $7,500. An hour-long video requires roughly 2 hours of a rater's time,

and typically costs $25 per hour for a rater. Thus, it would cost $56,120 to rate all videos

included in our study. The total cost of rater training and ongoing rater support would total

$63,620. In contrast, once an algorithm is developed, the only cost of a text-as-data approach is

from transcribing videos. In the current study, we spent $30 to transcribe each video and in total

it cost $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, reduce cost, and help teachers better align their practices with those associated with achievement gains for students. As a proof of concept, the current 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 videos of 4th- and 5th-grade ELA classes collected by the MET project, 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 with many of the domains and dimensions identified by the popular observation protocols CLASS and PLATO, meaning that the text-as-data approach can detect classroom instructional practices that are consistent with professional assessments of teaching quality. 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 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, the current 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 Lasso can create far richer measures that are more closely linked to classroom content. Moreover, due to its small sample size, the current 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 since, 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.

Classroom observation is time-consuming for principals, instructional coaches and other school leaders (Grissom, Loeb, & Master, 2013). It is also time-consuming and resource intensive for researchers studying effective teaching (Kane & Staiger, 2012). Automated metrics, such as the ones we discuss here, could address 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 lights on the technology of teaching and learning.

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. For schools leaders, there is also enormous potential upside to having 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 the most instrumental in driving improvement (Cohen et al., 2020; Taylor & Tyler, 2012). As such, districts are in need of tools that faciliate quick feedback to teachers who are looking to improve. Principals and coaches could then focus their efforts on helping teachers make sense of the information provided and identifying strategies for improvement.

Of course, an important step in this would be the understanding how principals and teachers perceive automated measures and respond to the information they provide. A plus of classroom observatioons—versus computer-generated value-added measures—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. 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.

References

Allen, J., 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*. Retrieved from 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.

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

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. *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project*, 50-97.

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.

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

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). New York, NY: 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.

Cazden, C. B. (2001). *Classroom discourse: The language of teaching and learning* (2nd ed.). Portsmouth, NH: Heinemann.

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.

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).

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.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*(6), 673–682.

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

Fenstermacher, G. D., & Richardson, V. (2005). On making determinations of quality in teaching. *Teachers College Record*, *107*(1), 186–213.

Gage, N. L., & Needels, M. C. (1989). Process-product research on teaching: A review of criticisms. *The Elementary School Journal*, *89*(3), 253–300.

Gentzkow, M., Kelly, B., & Taddy, M. (2017). *Text as data* (NBER Working Paper No. 23276). Cambridge, MA: 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). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

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.

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.

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.

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. (2017). Assessing principals' assessments: Subjective evaluations of teacher effectiveness in low-and high-stakes environments. *Education Finance and Policy*, 12(3), 369-395.

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*. New York, NY: 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.

Heath, S. B. (1983). *Ways with words: Language, life and work in communities and classrooms*. Cambridge university Press.

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., 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*, 372-400.

Ho, A. D., & Kane, T. J. (2013). The Reliability of Classroom Observations by School Personnel. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

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, B., Rockoff, J. E., Taylor, E. S., Lindy, B., & Rosen, R. (2016). *Teacher applicant hiring and teacher performance: Evidence from DC public schools* (NBER Working Paper No. 22054). Cambridge, MA: National Bureau of Economic Research.

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). Seattle, WA: 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*. Advance online publication.

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

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling : A review. *Economics of Education Review*, *47*, 180–195.

Lee, C. D., & Smagorinsky, P. (2000). *Vygotskian Perspectives on Literacy Research: Constructing Meaning through Collaborative Inquiry. Learning in Doing: Social, Cognitive, and Computational Perspectives*. Cambridge University Press

Marzano, R. J., Pickering, D., & Pollock, J. E. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement*. Ascd.

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.

Mehan, H. (1979). 'What time is it, Denise?": Asking known information questions in classroom discourse. *Theory into practice*, *18*(4), 285-294.

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

McGinty, A. S., Justice, L. M., Zucker, T. A., Gosse, C., & Skibbe, L. E. (2012). Shared-reading dynamics: Mothers' question use and the verbal participation of children with specific language impairment. *Journal of Speech, Language, and Hearing Research*.

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.

Niculae, V., & Danescu-Niculescu-Mizil, C. (2016). *Conversational markers of constructive discussions*. Retrieved from 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*, 261-290.

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

Park, Y. S., Chen, J., & Holtzman, S. L. (2015). Evaluating efforts to minimize rater bias in scoring classroom observations. *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project*, 381-414.

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.

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us.* New York, NY: 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), 1–10.

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., 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.

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

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(1), 175-214.

Rowe, M. B. (1986). Wait time: slowing down may be a way of speeding up!. *Journal of teacher education*, *37*(1), 43-50.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, *77*(4), 454–499.

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*, 613–624. Geneva, Switzerland: 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), 3628-51.

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

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

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. *Ann Arbor: Inter-University Consortium for Political and Social Research, The University of Michigan*

Figure 1: Features of Open-Ended and Non-Open-Ended Questions

Open-Ended   Non-Open-Ended

Open-Ended	Non-Open-Ended
saying	together
feel	totally
important	thank
think	madison
what's	ever
happened	cat
infer	colors
might	wants
place	love
listen	let
passage	box
gold	jordan
mean	may
things	last
else	answer
home	everyone
water	supposed
somebody	read
means	will
number	please

Note. Results are from Lasso regression. Word size is proportionate to coefficient size.

Figure 2: 15-Topic Model

Topic 11: yeah, realli, mayb, think, guy, right, bit, kind, might, lot
Topic 9: group, four, student, five, partner, name, three, seat, six, tabl
Topic 2: main, idea, support, detail, summari, summar, antarctica, passag, most, paragraph
Topic 6: revis, draft, elabor, edit, expositori, plan, narrat, descript, writer, outlin
Topic 15: subject, predic, word, synonym, sequenc, transit, card, compound, context, antonym
Topic 10: quotat, answer, question, mark, letter, comma, capit, elimin, address, strategi
Topic 3: caption, contrast, sourc, articl, diagram, non, venn, nonfict, featur, comet
Topic 14: father, bruno, barbi, crow, bed, gretel, dad, papa, mama, solomon
Topic 4: infer, conclus, knowledg, erandi, inferenc, prior, clue, background, titan, draw
Topic 7: persuad, opinion, persuas, entertain, fact, negat, proven, convinc, purpos, argument
Topic 1: mysteri, tone, fabl, mood, charact, mous, trait, solut, isabel, mule
Topic 12: immigr, war, govern, railroad, slave, slaveri, wong, american, america, york
Topic 13: poem, simil, stanza, rhyme, metaphor, poetri, poet, personif, alliter, willow
Topic 8: effect, caus, walter, plant, seed, tree, earth, relationship, turtl, climb
Topic 5: pig, wolf, view, gilli, theme, perspect, navajo, kenni, straw, stellaluna

0.0   0.1   0.2   0.3   0.4

Expected Topic Proportions

Figure 3: 20-Topic Model

Topic 4: student, group, four, three, name, five, partner, two, okay, minut
Topic 13: yeah, pictur, mayb, dog, feel, kind, realli, bit, guy, mind
Topic 7: answer, question, strategi, select, text, choic, number, test, elimin, evid
Topic 9: topic, write, expositori, paper, writer, revis, narrat, plan, paragraph, step
Topic 1: main, support, idea, detail, passag, most, antarctica, topic, paragraph, franklin
Topic 14: infer, effect, caus, conclus, knowledg, inferenc, draw, prior, result, clue
Topic 19: dad, mom, summer, famili, negat, kid, game, parent, ball, play
Topic 17: earth, ocean, featur, comet, caption, river, sea, explor, planet, non
Topic 18: theme, charact, plot, erandi, mysteri, trait, fabl, solv, conflict, problem
Topic 15: synonym, context, vocabulari, word, prefix, antonym, sequenc, suffix, homophon, opposit
Topic 12: simil, stanza, rhyme, poem, metaphor, poetri, alliter, verb, personif, adverb
Topic 6: immigr, tax, govern, coloni, elli, war, civil, uniform, island, barbi
Topic 10: view, persuad, entertain, wolf, pig, purpos, tone, persuas, perspect, convinc
Topic 16: opinion, fact, proven, biographi, prove, statement, thorp, general, websit, valid
Topic 11: summar, summari, articl, whale, highlight, oil, transit, import, friction, trout
Topic 2: bruno, walter, gretel, jessi, mama, libbi, runt, edward, solomon, philip
Topic 8: plant, seed, tree, beaver, navajo, mous, cell, stellaluna, lorax, nest
Topic 20: milk, ride, bake, butter, tsar, sugar, cream, hood, bread, bowl
Topic 5: quotat, capit, alright, comma, letter, punctuat, quot, mark, address, wong
Topic 3: predic, myth, legend, subject, compound, henri, titan, mule, simpl, lightn

0.0   0.1   0.2   0.3   0.4

Expected Topic Proportions

Table 1: Descriptive Statistics

Characteristic	Mean	SD
Students (*n* = 13,370)		
Age	9.73	(0.87)
Male	.50	
Gifted	.09	
Special education	.09	
English language learner	.13	
Free or reduced-price lunch	.43	
Race/ethnicity		
White	.25	
African American	.42	
Hispanic	.23	
Asian	.06	
Other	.03	
ELA score		
2009	0.12	(0.96)
2010	0.09	(0.97)
Teachers (*n* = 258)		
Male	.08	
Race/ethnicity		
White	.63	
African American	.32	
Hispanic	.03	
Other	.01	
Years in district	5.98	(5.43)
Master's or higher	.46	
Classrooms (*n* = 95)		
Class size	24.58	(5.58)
Average age	9.74	(.75)
Male	.50	(.10)
Gifted	.09	(.16)
Special education	.10	(.09)
English language learner	.14	(.16)
Free or reduced-price lunch	.42	(.29)
Race/ethnicity		
White	.26	(.29)
African American	.42	(.36)
Hispanic	.23	(.22)
Asian	.06	(.11)
Other	.03	(.04)
Average ELA score		
2009	0.04	(0.58)

2010	0.07	(0.51)

Note. Data are restricted to teachers who participated in the MET project's second-year randomization process. 4 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 2009–2010 and 2010–2011 data. Different variables may have different numbers of observations.

Table 2: Computer-Generated Metrics on Teacher Practices

Variable	*Mean*	*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.

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

Variable	Rotated			Non-rotated		
	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.

Table 5: Correlations with CLASS and PLATO

Dimension	Factor 1: Classroom Management	Factor 2: Interactive Instruction	Factor 3: Teacher-Centered Instruction
CLASS			
Domain 1: Emotional support	0.165**	0.158**	-0.069
Positive climate	0.124*	0.168**	-0.070
Negative climate	0.217**	-0.006	0.040
Teacher sensitivity	0.198**	0.169**	-0.069
Regard for student perspectives	0.253**	0.106+	-0.063
Domain 2: Classroom management	-0.147*	0.082	0.080
Behavior management	-0.280**	-0.002	0.082
Productivity	-0.186**	0.043	0.078
Instructional learning formats	0.100+	0.166**	0.040
Domain 3: Instructional support	0.157**	0.246**	-0.033
Content understanding	0.049	0.233**	0.029
Analysis and problem solving	0.090	0.120*	0.011
Quality of feedback	0.202**	0.239**	-0.024
Instructional dialogue	0.191**	0.259**	-0.117*
Student engagement	0.036	0.143*	0.028
PLATO			
Intellectual challenge	0.083	0.137*	-0.172**
Classroom discourse	0.143*	0.220**	-0.098
Behavior management	-0.183**	0.055	0.094
Modeling	-0.140*	-0.025	0.325**
Strategy use and instruction	-0.068	0.043	0.310**
Time management	-0.167**	0.112+	0.042
Representation of content	0.002	-0.049	-0.106+

** $p < .01$. * $p < .05$. + $p < .10$.

Table 6: Regression of VAMs on Predicted Factors

Factor	State ELA					SAT-9				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Factor 1: Classroom management	-0.009 (0.011)	-0.007 (0.011)	-0.007 (0.011)	-0.005 (0.011)	-0.004 (0.011)	-0.023 (0.019)	-0.024 (0.019)	-0.023 (0.019)	-0.022 (0.019)	-0.022 (0.019)
Factor 2: Interactive instruction	0.021+ (0.011)	0.020+ (0.011)	0.018 (0.012)	0.016 (0.011)	0.017 (0.012)	0.006 (0.026)	0.004 (0.026)	-0.006 (0.026)	-0.005 (0.027)	-0.006 (0.026)
Factor 3: Teacher-centereded instruction	-0.007 (0.010)	-0.007 (0.010)	-0.006 (0.009)	-0.008 (0.010)	-0.009 (0.010)	-0.033+ (0.018)	-0.036* (0.018)	-0.033+ (0.018)	-0.041* (0.018)	-0.041* (0.018)
Observations	285	285	283	281	281	282	282	281	279	279
R-squared	0.029	0.091	0.091	0.100	0.103	0.020	0.049	0.054	0.068	0.068
Class characteristics		X	X	X	X		X	X	X	X
CLASS average score			X		X			X		X
PLATO average score				X	X				X	X
District fixed effects	X	X	X	X	X	X	X	X	X	X
Grade fixed effects	X	X	X	X	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.
** $p < .01$. * $p < .05$. + $p < .10$.

Table 7: Regression of VAMs on Individual Measures

Measure	State ELA		SAT-9	
	(1)	(2)	(3)	(4)
Turns per minute	0.012	0.011	0.007	0.007
	(0.012)	(0.012)	(0.025)	(0.025)
Proportion of time teacher talks	-0.016	-0.016+	-0.037*	-0.043*
	(0.010)	(0.010)	(0.018)	(0.017)
Average words per minute	0.028*	0.028*	0.020	0.024
	(0.012)	(0.011)	(0.019)	(0.019)
"You"	-0.009	-0.007	-0.012	-0.012
	(0.011)	(0.012)	(0.018)	(0.019)
"I"	-0.013	-0.014	-0.042*	-0.040*
	(0.009)	(0.009)	(0.017)	(0.019)
Language style matching	0.014	0.014	0.042*	0.048*
	(0.010)	(0.009)	(0.019)	(0.019)
Analytic words	-0.001	-0.004	0.022	0.024
	(0.011)	(0.011)	(0.020)	(0.020)
Social words	0.000	0.002	-0.004	-0.007
	(0.010)	(0.011)	(0.017)	(0.017)
Open-ended questions per minute	0.012	0.012	0.004	-0.001
	(0.011)	(0.011)	(0.022)	(0.023)
Routine language	-0.015	-0.014	-0.015	-0.017
	(0.012)	(0.011)	(0.018)	(0.018)
Observations	285	285	282	282
Class characteristics		X		X
District fixed effects	X	X	X	X
Grade fixed effects	X	X	X	X

Note. Robust standard errors are reported in parentheses. Each cell is from a separate regression. All independent variables are in standardized values. All regressions control for district and grade fixed effects. Robust standard errors are in parentheses.
** $p < .01$. * $p < .05$. + $p < .10$.

Appendix A

Word Categories Used for Calculating Language Style Matching

Category	Examples
Personal pronouns	I, his, their
Impersonal pronouns	it, that, anything
Articles	a, an, the
Conjunctions	and, but, because
Prepositions	in, under, about
Auxiliary verbs	shall, be, was
High-frequency adverbs	very, rather, just
Negations	no, not, never

Appendix B

Suppose there are N questions that teachers ask in a classroom, with each question i having a label $y_i \in \{-1,1\}$, which is a binary indicator for whether the question is open ended. Each question i is represented as $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ij})$, where X_{ij} represents a specific word. To identify the features (i.e., words) that predict whether a question is open-ended, I need to estimate Function 1 below.

$$f(\boldsymbol{\beta}, \mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{N}(y_i - \beta_0 + \sum_{j=1}^{J}\beta_j X_{ij})^2 \qquad (1)$$

However, since J is likely to be larger than N, and many variables are correlated, an ordinary least squares approach is not feasible. Thus, a Lasso approach is used to add "penalty" to coefficients in the objective function. Specifically, we estimate Function 2 instead, which reduces small coefficients to zero.

$$f(\boldsymbol{\beta}, \mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{N}(y_i - \beta_0 + \sum_{j=1}^{J}\beta_j X_{ij})^2 + \lambda \sum_{j=1}^{J}|\beta_j| \qquad (2)$$

Appendix C

Before running the topic model, we conducted standard pre-processing to prepare the texts for analysis. We first removed all stop words and punctuation. Then, we transformed all the words to lowercase and reduced them to their root form in a procedure called *stemming*.

As an unsupervised method, topic modeling requires the number of topics to be set before the model is run. The optimal number of topics achieves a balance between model fit and substantive interpretation. We ran both 15-topic and 20-topic models, which yielded similar results, with the 20-topic model capturing some more nuanced themes. Essentially, we estimated Model 3, shown below, in which each classroom i is a combination of topics 1 through 15. I ultimately decided that 15 is the optimal number of topics for this study because the 15-topic model achieved better reliability than the 20-topic model. The results of this model were then used in other statistical analyses.

$$Class_i = (topic_{i1,} \ topic_{i1}, ..., topic_{i15,}) \qquad (3)$$