2020

Explorations of Classroom Talk and Links to Reading Achievement in Upper Elementary Classroo

Amanda P. Goodwin

Sun-Joo Cho

Daniel Reynolds

Stephanie Nunn

Rebecca Silverman

Explorations of Classroom Talk and Links to Reading Achievement in Upper Elementary Classrooms

Amanda P. Goodwin and Sun-Joo Cho
Vanderbilt University's Peabody College

Dan Reynolds
John Carroll University

Rebecca Silverman
Stanford University

Stephanie Nunn
SRI International, Education Division, Arlington, Virginia

The current study reports on a large-scale quantitative analysis of classroom talk practices and links to different measures of reading achievement within upper elementary classrooms. Data involving 745 fourth- and fifth-grade teachers and 18,844 students from the Measures of Effective Teaching (MET) study were used. Talk was quantified via various talk-related indicators from 2 observation protocols and a student survey. Dimensionality analyses suggest these indicators represent 4 factors consisting of teacher explaining, questioning, encouraging of student talk, and big-picture communicating. Links to 2 different standardized reading achievement measures were also modeled with improved ratings of teacher explanations and questioning predicting higher standardized reading scores. Relationships varied, though, by different measures of classroom talk (i.e., observational protocols vs. student surveys) and levels of analysis (i.e., the student, class period, or school level). Students' but not observers' ratings of talk practices linked to standardized reading at the class period level, whereas observers' ratings related to standardized reading performance at the school level. Interpretations, implications for future research, and connections to educational practice are conveyed.

Educational Impact and Implications Statement

The current study makes important contributions to the literature in terms of understanding the makeup of talk present in United States elementary Language Arts classrooms and links to reading performance. We consider a larger sample than previously possible (745 teachers and 18,844 students) as well as multiple ways that talk and reading are measured. Overall, we found that talk matters and different types of talk are more or less supportive of reading achievement. Specifically, teacher explanations and teacher questions seemed to improve reading performance, but student talk and big-picture communicating did not, although we emphasize quality may be more important than quantity. Also, students and observers have important lenses to consider as they noted different components of talk that linked differently to different measures of reading achievement. Student seemed to notice talk patterns that mattered to their own learning as well as the learning of those in their class. Observers noted talk patterns that linked to learning across the school. Overall, the study shows that talk matters to different types of reading achievement and also when considering large numbers of students, teachers, classrooms, districts, and even states.

Keywords: reading, teacher talk, student talk, explaining

Research highlights the relationship between language and reading comprehension (i.e., vocabulary, syntax, etc.; Florit & Cain, 2011; Snow & Kim, 2007), and one way of building language skills is to embed the child in a language-rich environment. As Gámez and Lesaux (2015) write, "Language is central to reading comprehension skills . . . and . . . the curriculum is largely mediated by language, [so] it is important to better understand . . . [classroom] language" (p. 448). For this study, we term classroom

Amanda P. Goodwin, Department of Teaching and Learning, Vanderbilt University's Peabody College; Sun-Joo Cho, Department of Psychology and Human Development, Vanderbilt University's Peabody College; Dan Reynolds, Department of Education and School Psychology, John Carroll University; Rebecca Silverman, Stanford Graduate School of Education, Stanford University; Stephanie Nunn, SRI International, Education Division, Arlington, Virginia.

Correspondence concerning this article should be addressed to Amanda P. Goodwin, Department of Teaching and Learning, Vanderbilt University's Peabody College, PMB, 230 Appleton Place, Nashville, TN 37203. E-mail: amanda.goodwin@vanderbilt.edu

language as classroom talk, which includes teacher talk and the larger discourse going on within the classroom.

In general, literature exploring classroom talk has used different definitions (i.e., discussions, exploratory talk, dialogic talk, teacher-talk, or student-talk) and frameworks (sociocognitive, sociocultural, ethnographic, and social-linguistic; see Applebee, Langer, Nystrand, & Gamoran, 2003 for detailed discussion). These varied lenses each provide fine-grained understandings of high quality classroom interactions that promote language skills and reading comprehension (i.e., meaning making/envisionment; see Alexander, 2008; Langer, 1995; Mercer, 2008; Mercer & Littleton, 2007; Nystrand, 1997). For example, when analyzing transcripts of 8th and 9th grade literature discussions, Nystrand (1997) used a particular lens, specifically Bakhtin's (1981) idea of dialogic interaction, ultimately recommending that teachers strive for deeper comprehension by using authentic questions, building on student comments, and integrating students' personal interpretations. Missing in the literature, though, are studies that investigate the relationship between classroom talk and student achievement using different ways of measuring these constructs and at a large scale that can inform on average, how these relations work across many different students, classrooms, schools, districts, and states.

Our goal is to explore the general classroom talk practices that are typical in current U.S. classrooms and their links to reading achievement. As Nystrand (2006) notes, "Until recently, there have been few large-scale quantitative studies of the effect of classroom discourse on reading comprehension, and these studies have mainly focused on middle and high schools" (pp. 403–404). The present study adds to the research base by investigating relationships between indicators of classroom talk and student achievement in a large number of upper elementary school classrooms. To do this, we use two common observation protocols and a student survey with talk-related indicators. By using multiple measures of classroom talk as well as multiple measures of reading achievement, we are able to explore how the different ways classroom talk and student outcomes are measured influence conclusions at a large scale. Specifically, our study identifies talk-related practices (e.g., explaining content a different way or encouraging students to share their thoughts), groups them into meaningful constructs present within the larger classroom talk literature (i.e., teacher questioning, teacher explaining, and student talk), links these practices to different measures of reading achievement (i.e., primarily multiple choice vs. open response), and explores differences in how talk is measured (i.e., observations vs. student ratings) that can reveal nuances in how classroom talk and students' achievement are related.

Theoretical Framework

Our study is grounded in two theoretical ideas. The first, supported by a plethora of reading theories is that reading is a complex process based in language (see Alvermann, Unrau, & Ruddell, 2013 for an overview) that involves different processes depending on the activity (RAND Reading Study Group, 2002). At its most basic level, the simple view of reading suggests that reading is the product of decoding and linguistic comprehension (Gough & Tunmer, 1986; Hoover & Gough, 1990). Here, linguistic comprehension, which could also be referred to as language skill, is defined as the ability to "understand language" (Hoover & Gough, 1990, p. 131) or "the process by which, given lexical (i.e., word) information, sentences and discourses are interpreted" (Gough & Tunmer, 1986, p. 7). By upper elementary school, language skills become especially important to reading (Florit & Cain, 2011) because fewer resources are needed for word reading, allowing students to focus on using understandings of language to put the meanings of the words they read together into larger idea units that are integrated across the larger text (Perfetti, 1988). The way in which students read differs depending on the activity involved in their reading (i.e., the purpose or task; RAND Reading Study Group, 2002), suggesting the need to consider potentially different roles of language in different reading comprehension activities such as those involving lower level demands versus more cognitively demanding tasks. As such, we consider the role of language in two different types of reading comprehension activities.

The second theoretical lens relates to how these language skills are developed. Language skill develops through interactions with more knowledgeable others (MKOs), such as parents, teachers, and even peers, who mediate learning through linguistic and instructional support (Vygotsky, 1978). Talk begets language development and, therefore, the more students interact with MKOs through language, the more their language skills develop. Applied to a school context, interactions with MKOs (Mercer, 1995, 2000) help children learn about the academic language of school (Bailey, 2007). Specifically, MKOs use language to transmit information to children and support their meaning-making, and students learn not just about that information but also about how language works. Here, teachers and peers facilitate language use through explanation and focused questioning, modeling the higher order thinking and processing skills involved in reading (Ninio & Bruner, 1978). Children internalize this language and use it to guide them in accomplishing new tasks on their own (Vygotsky, 1986). The context of the talk (i.e., the school and classroom culture), the histories of the speakers, and the collective thinking that is occurring all affect classroom talk practices and, thereby, reading successes (Nystrand, Wu, Gamoran, Zeiser, & Long, 2003).

Classroom Talk and Reading

Current research shows a consistent relationship between classroom talk and student achievement on reading-related outcomes (Applebee et al., 2003; Murphy, Wilkinson, Soter, Hennessey, & Alexander, 2009). The content of classroom talk matters to literacy learning in multiple ways (Carlisle, Kelcey, Berebitsky, & Phelps, 2011; Connor et al., 2014; Silverman et al., 2014). Teachers' use of sophisticated vocabulary and more vocabulary instruction link to reading comprehension performance (Dickinson & Porche, 2011; Gámez & Lesaux, 2015). Also, the frequency of certain instructional practices like teaching definitions, word relations, morphosyntax, and inferential comprehension predicts better standardized reading comprehension and vocabulary scores (Silverman et al., 2014).

Beyond content, types of classroom talk (i.e., explanations, questions, or student-talk) likely matter too, and these features tend to be embedded within other general instructional practices. For example, teachers may embed clear explanations within instruction focused on eliciting problem solving and critical thinking skills. Similarly, clear explanations may be embedded within lessons

designed to teach content knowledge. Either way, teachers choose different words and approaches to deliver identical lessons on the same skill (e.g., comprehension monitoring), and these differences lead to different learning outcomes (Duffy, Roehler, & Rackliffe, 1986). Observational studies highlight these differences, although many focus on older students. In a review of research on literacy classroom talk, Nystrand (2006) stated that, "Discussion practices vary widely among classrooms, from teacher elaborations during question-and-answer recitation, or what Wells (1993) calls IRF (Initiation-Response-Followup), to debates, to open-ended sharing of ideas, including multiple turns uninterrupted by teacher test questions" (p. 395). This is highlighted in a study of 64 middle and high school classrooms and 1,111 students where Applebee et al. (2003) found "broad and important differences in approaches to teaching and learning" visible in teachers' talk practices (p. 710). For example, while one teacher led students through a story using questions and explanations to convey her view of the text, a second used talk to direct conversation, focus on textual complexities, and develop students' own interpretations. Findings across the 64 classrooms showed that that discussion-based approaches (as measured by the Classroom Assessment Scoring System [CLASS], Hamre & Pianta, 2007) similar to the second teacher's approach linked positively to students' literacy performance. Similarly, Michener and colleagues (2018) studied 31 upper elementary teachers and 236 students, finding that teachers' explanations and their follow-up moves predicted increased student comprehension (Michener et al., 2018). Below, we review types of classroom talk that have been identified as supporting reading comprehension.

Teacher Questioning

Questioning can be used for different purposes, including assessing and building understanding as well as directing attention to learning procedures and objectives. Questions tend to be classified based on the amount of talk encouraged, the level of thinking involved, or the content being assessed. The most often used questioning format follows an Initiate-Response-Evaluate (IRE) structure (Cazden, 2001; Nystrand, 2006; Sinclair & Coulthard, 1975). Here, teachers ask closed-ended questions, students respond, and then teachers evaluate the students' response. To move away from this recitation-style teaching, teachers might prime the discussion with accessible open-ended questions, build on engaged student responses, and facilitate student-to-student talk (Nystrand et al., 2003). Additional types of questions include procedural, rhetorical, or discourse-management questions as well as open-discussion questions, authentic teacher questions with no specific prespecified answer, or questions with uptake that incorporate what an earlier speaker had said (Applebee et al., 2003). Although questioning varies, what is clear from the literature is that questioning helps teachers determine and expand students' understanding, suggesting a potential support for reading comprehension.

Teacher Explaining

Teacher explanations tend to be used to build understanding and are generally linked to direct instruction (Winograd & Chou, 1988), and have been linked with increased reading comprehension in upper elementary students (Michener et al., 2018). They can also be part of responsive instruction when students need explanations in the middle of a lesson (Roehler & Cantlon, 1997). As Roehler and Duffy (1984) write,

> Teacher explanations of processes are designed to be metacognitive, not mechanistic. They make students aware of the purpose of the skill and how successful readers use it to activate, monitor, regulate, and make sense out of text, creating in students an awareness and a conscious realization of the function and utility of reading skills and the linkages between these processes and the activities of reading (p. 266).

Explanations often differ in describing procedures of learning versus larger purposes for learning. For example, when teaching how to use context to determine the meaning of a new word, one teacher described the skill focusing on terminology (e.g., context clues), a specific set of steps to use, and related it to how it would help in school (e.g., being easier than a dictionary). In contrast, another teacher focused on the process, suggested steps as part of a flexible process, and discussed how the learning would be helpful even outside of school (Duffy et al., 1986). Research suggests several characteristics of explanations that support reading comprehension. These include explicitness, such as directly teaching comprehension strategies (Shanahan et al., 2010); presentation within the context of application in real texts (Duffy et al., 1986); structures that involve small but brisk steps with many examples (Rosenshine, 1983); and highlighting purpose via communicating what is being learned, why it is being learned, or how to use the learning (see Winograd & Chou, 1988, for examples).

Encouraging Student Talk

Students tend to use their talk to develop and communicate thinking, potentially an avenue for increasing language skill and, therefore, reading comprehension, but teachers merely increasing percentages of student talk during instructional time does not necessarily improve reading comprehension scores (Murphy et al., 2009). Efforts to improve the quality of student talk stem from Cazden's (2001) work exploring students' peer-to-peer talk, which has also been called exploratory talk (Mercer, 1995). These findings align with Soter et al.'s (2008) elements of productive discussion of text, which include setting up classroom routines where students have the floor for extended turns, using authentic and open-ended question prompts, and encouraging uptake of other students' ideas. This supports calls to establish productive norms for class discussion to ensure that student talk is of high quality and productive for learning (Hogan & Pressley, 1997; Rex & McEachen, 1999). Similarly, Wolf, Crosson, and Resnick (2005) showed that teachers who fostered student talk moves that presented evidence and explained thought processes predicted lesson rigor within elementary and middle school comprehension lessons. This links to the importance of uptake and conceptual press: pressing students to elaborate on their initial responses has positive effects on reading comprehension (Boyd & Rubin, 2006; McElhone, 2012; Michener, 2014; Nystrand & Gamoran, 1991). Student talk is rare, though: after observing 33 third to fifth grade reading or language arts classrooms three times across a year, student talk made up only 25% of classroom talk (Silverman et al., 2014). This may be because encouraging student talk is not easy. Daniel, Martin-Beltrán, Peercy, and Silverman (2015) illustrate that even within carefully designed curricula, activities and questions may

unintentionally constrict opportunities for student-to-student discussion and instead encourage students to enact rote IRE sequences with each other.

Assessing Classroom Talk

To understand features of classroom talk and potential links to reading, one must consider both the dynamic nature of talk as well as the different ways that talk might be measured within the classroom context. For example, Connor et al. (2014) analyzed 27 third grade classrooms and showed that "both global quality of the classroom learning environment and time individual students spent in specific types of literacy instruction covering specific content interacted to predict students' comprehension and vocabulary gains, whereas neither system alone did" (p. 762). This dynamic approach considers that good teachers "routinely provide timely and detailed feedback, but not necessarily in the same ways for all students" (Porter & Brophy, 1988, p. 82). It may be that a teacher explains an idea to one student but elicits student talk with another. Alternatively, students may interpret teacher talk differently depending on their histories and skills resulting in one student viewing an explanation as clear whereas another student may find that same explanation vague. Furthermore, research indicates that student perceptions of their teacher are key considerations in their interactions—and that these perceptions impact their working relationship and their achievement (Toste, Heath, & Dallaire, 2010). As such, a student with a perceived positive teacher-student relationship may interpret an explanation as more informative or helpful compared with a student with a perceived negative teacher-student relationship. Considering these individual student experiences related to classroom talk within the larger classroom experiences can deepen understanding of talk that supports reading.

Another important consideration is assessment. Both general (i.e., the CLASS; Hamre & Pianta, 2007) and content-specific (i.e., the Protocol for Language Arts Teaching Observation [PLATO]; Grossman et al., 2010) classroom observation systems and student survey measures (i.e., the Tripod Student Survey; Ferguson, 2008) include indicators of teacher and student talk. These provide different lenses through which to examine classroom talk. Whereas one measure looks at talk practices within instruction more generally, another focuses on talk practices related to language arts instruction more specifically. Furthermore, observational protocols tend to reflect instruction that goes on during a single visit as interpreted by an impartial visitor. In contrast, student surveys take into account perception of talk practices that go on across a school year as interpreted by a student who is a member of that classroom (see above points regarding student perception). Studying data from multiple measures may allow researchers to better understand how different measurement systems highlight certain aspects of classroom talk and relate differentially to reading comprehension.

Assessing Reading Comprehension

To understand the relationship between classroom talk and reading, the way in which reading is assessed must be taken into account as well. While a review of reading comprehension assessment is beyond the scope of the current article, we highlight this because research and theory emphasize the importance of considering multiple measures of comprehension—and thinking carefully about what those measures assess. From the perspective of content, research and theory has emphasized multiple levels of comprehension (i.e., literal, inferential, and evaluative). As Basaraba, Yovanoff, Alonzo, and Tindal (2013) write,

> Each of the tasks involved in understanding a text—whether it is simply to recall what is stated in the text (literal comprehension), to interpret the authors' meaning through connecting information that is implicit in the text (inferential comprehension), or to go beyond the text by relating what is being read to prior experiences and knowledge (evaluative comprehension)—requires a different level of cognitive processing by the reader.

Indeed, previous studies have shown that determining what predicts reading comprehension depends on which measure of reading comprehension is being used (Cutting & Scarborough, 2006). Additionally, the format of the assessment matters as researchers have shown that forced-choice formats induce responses and processes very different from other formats like open-response or cognitive interviews (Rupp, Ferne, & Choi, 2006). These different content and formats evoke different comprehension products that could relate differently to classroom talk. For example, emphasizing student talk might build higher level thinking that can be better showcased in evaluative questions presented within open-response environments compared with literal questions presented in forced-choice formats. While it is beyond the scope of our study to look at each type of question, we do consider potential differences in the relationship between classroom talk and different comprehension products. In particular, we consider both state standardized tests (i.e., primarily multiple choice) as well as a standardized test that purposely assesses higher-level thinking via open response format.

The Current Study

Our study determines features of classroom talk as embedded within larger instruction and links between those features and different measures of reading comprehension. We build on earlier work with smaller samples that have unraveled student, classroom, and school level variables that support classroom talk (see earlier review and Applebee et al., 2003 for larger discussion). Our study instead focuses on identifying the structure of classroom talk and its relationship to different measures of reading comprehension at a scale not previously considered. This scale allows us to model on average, how these relationships work across a large number of different students, classrooms, schools, districts, and states.

To do this, we capitalize on the Measures of Effective Teaching's (MET) data that includes multiple measures with talk-related indicators from a large number of classrooms across different school districts teaching varied curricula and also multiple measures of reading comprehension, each purporting to measure a different type of comprehension. While working with large data sets like the MET project dataset involves challenges such as being limited to the measures collected (e.g., state standardized tests that are not exactly identical across states) and the select details of measures and procedures provided via project documentation (e.g., of which we do not have intimate knowledge), the benefits outweigh the challenges because of the scale of the this data, which cost more than one hundred million dollars to collect and that includes a range of data related to upper elementary learning

environments and performance for thousands of students and hundreds of teachers. Hence, we examine (a) how do different items related to classroom talk from various instruments and surveys relate within- and across- classrooms and (b) how does talk as embedded in general instruction relate to different measures of standardized reading achievement? Answers to such questions have important theoretical and practical implications.

Method

Data

The data for this study are drawn from the Bill and Melinda Gates Foundation's Measures of Effective Teaching Study (Bill & Melinda Gates Foundation, 2012; Gates Foundation, 2010). The larger study, conducted over 2 years in six urban school districts, involved approximately 3,000 fourth through ninth grade teachers and their students. The data include practices and quality information at the classroom, school, and district levels. For the purposes of studying elementary school classroom talk, the five school districts providing data on fourth and fifth grade language arts sections were included in our analysis. To minimize differences in data collection procedures across years, Year 1 data were analyzed.

Sample

Our sample included 18,844 fourth and fifth graders ($N = 9,287$ fourth graders; 9,557 fifth graders) learning within 822 class periods of 745 language arts elementary teachers ($N = 371$ fourth grade teachers; 374 fifth grade teachers). These students and teachers learned in 129 schools within five school districts. Some teachers taught self-contained elementary classrooms ($N = 610$) while others taught only language arts ($N = 135$). Students were mostly from minority backgrounds (see background characteristics in Table 1); 43.7% were Black, 23.3% were Hispanic, and 13.8% were English language learners. Less than 25% were White. In contrast, of the 720 teachers reporting demographic data, most were white ($N = 423$, 56.8%) or Black ($N = 251$, 33.7%) with few identifying as Hispanic ($N = 39$, 5.4%) or of other racial or ethnic background ($N = 7$, 0.97%). Less than 10% of the teachers were male ($N = 67$).

Table 1
Descriptive Statistics of Students (N = 18,844) and Teacher Background Characteristics (N = 745)

Background characteristics	Percentage of students	Percentage of teachers
Gender		
Male	48.92	8.72
Female	49.80	87.92
Missing	1.28	3.36
Ethnicity		
White	23.59	56.51
Black	43.67	33.83
American Indian	0.26	—
Hispanic	23.34	5.37
Asian	5.17	—
Other	2.69	0.94
Missing	1.28	3.36
ELL		
Yes	13.77	—
No	84.96	—
Missing	1.28	—
Special Ed		
Yes	8.27	—
No	88.49	—
Missing	3.24	—

Note. ELL = English language learner. Em dash (—) indicates demographics not applicable to the teacher participants.

Measures

Table 2 presents descriptive information for all measures. Each member of our research team independently determined which items from the observational measures and the student survey were relevant to classroom talk. These ratings were then compared and discussed until a consensus was obtained on the final set of items included in this study.

Student Survey Ratings (Tripod; Ferguson, 2008). Students completed the Tripod survey at a single point in time, although reflecting on instructional practices across the academic year in general. The survey was completed in either paper or electronic form, as preferred by their classroom teacher. The Tripod consists of items within seven domains: Care, Control, Clarify, Challenge, Captivate, Confer, and Consolidate. As mentioned previously, from the full survey, the research team independently rated the individual items' relevancy to classroom talk, coming to consensus on 17 items that were relevant to classroom talk practices. These items tend to focus on features of classroom talk (i.e., explanations, questioning, etc.) embedded within different content areas and goals. As such, this data provide students' impressions on classroom talk as embedded within instruction. A list of these items can be found in Table 2. Students rated items on a scale of 1–5. Reliability was found to be $\alpha = .865$ for the 17 items.

Rater Classroom Observations (CLASS and PLATO). Ratings of classroom quality stemmed from an average of two trained raters' scoring of up to four videoed lessons submitted by participating teachers using two measures of classroom quality adapted for the study: the CLASS and PLATOPrime. Videos were recorded between February and June and the teachers were the ones who did the video-recording. Training and special cameras were used. Each setup included two cameras with one focused on the board and the other showing a 360 degree view of the classroom. A microphone for the teacher and another designed to capture student voices was also used.

These observation protocols were used to capture the general classroom experience because observers were trained to consider the experience of the average student, so they would link small groups or pairs and whole class instruction to the scoring rubrics. Adaptations included limiting observations to the first 30–35 min of the lesson and including fewer items from the original PLATO tool ($N = 6$) based on prior studies that showed reliability and strong associations between these elements and student outcomes (Grossman, Cohen, Ronfeldt, & Brown, 2014). Before coding, raters received 17–25 hr of self-directed training on the procedures, how to eliminate bias, and the various protocols. They had to score a 70% match (on the 4-point PLATO scale) with master-coded video segments and reliability was reestablished multiple

Table 2
Descriptive Information for Measures

System	Item/measure	*M*	*SD*	Missingness (%)
	Measures of classroom talk practices			
CLASS (average scores) ratings 1–7	C1: Analysis and problem solving	2.87	0.49	39.05
	C2: Content understanding	4.11	0.46	39.05
	C3: Instructional dialogue	3.63	0.56	39.05
	C4: Quality of feedback	3.84	0.55	39.05
	C5: Regard for students perspective	3.41	0.55	39.05
PLATO ratings 1–7	P1: Classroom discourse (seg 1)	2.31	0.42	31.16
	P2: Classroom discourse (seg 2)	2.41	0.48	31.16
Tripod ratings 1–5	S1: T nice when ask I ask q's	4.31	0.94	14.34
	S2: T explains another way	4.31	0.91	14.28
	S3: T several ways of explaining	4.20	0.92	16.06
	S4: T explains difficult things clearly	4.28	0.93	15.49
	S5: T explains in orderly way	4.13	1.00	16.25
	S6: T marks papers to show S how to improve	3.79	1.23	16.16
	S7: T wants S to explain answers	4.20	0.95	16.42
	S8: We learn to correct mistakes	4.47	0.80	15.37
	S9: T asks whether we understand	4.36	0.92	15.52
	S10: T tells us what we are learning and why	4.25	0.96	15.76
	S11: T asks q's to make sure S follows along	4.44	0.84	15.77
	S12: T checks to make sure S understands	4.43	0.85	15.9
	S13: T wants S to share ideas	3.88	1.12	15.94
	S14: T summarizes what is learned each day	3.75	1.20	16.68
	S15[a]: S don't share ideas, just listen to T	2.64	1.30	17.10
	S16: S speak up and share ideas	3.69	1.15	16.46
	S17: T gives S time to explain ideas	4.14	0.95	15.62
	Measures of reading achievement			
State test	ELA_{pre}: State standardized ELA test	0.10	0.95	10.25
State test	ELA_{post}: State standardized ELA test	0.09	0.96	4.59
Open-response	SAT: Stanford achievement ninth edition open-ended reading test	608.2	37.4	13.10

Note. CLASS = classroom assessment scoring system; PLATO = protocol for language arts teaching observation; EFA = exploratory factor analysis; ELA = English language arts assessment; SAT = Stanford 9 Open-Ended Reading Assessment.
[a] S15 was not used because of negative wording that resulted in a negative discrimination value with the recoded S15.

times during a rater's viewing of video segments. The project used multiple protocols to assure reliability including raters beginning each shift with the scoring of calibration videos, then scoring 5% of videos as validity videos where true scores were available to compare the performance of the rater with ideal performance (these were unknown to the rater), and then a scoring leader also scored one video per shift to confirm that raters were scoring accurately. Overall, while the MET project does not provide interrater reliability, less than 10% in the variance in scores was because of rater effects (MET project user guide). While designed to assess general quality, we identified a subset of talk-related items on these measures to use in our analyses. More information is below for each specific observational measure.

Classroom assessment scoring system (CLASS; Hamre & Pianta, 2007). Data for five of the 11 domains of the CLASS tool were identified as containing scoring criteria relating to classroom talk. As part of the MET project scoring, each domain was rated on a scale of 1–7. Scoring rubrics highlighted quantity with low scores of 1 or 2 tending to be awarded for no occasions of that behavior, mid scores of 3, 4, or 5 awarded for occasional or limited transfer opportunities, and high scores of 6 or 7 awarded for extended or consistent occasions of the coded behavior. Phrases like "no opportunities" and "not encouraged" link to scores of 1 or 2, words and phrases like "occasionally" or "opportunities in familiar contexts" link to scores of 3, 4, or 5, and words like "consistently, novel, or independent" link to scores of 6 or 7. See https://www.icpsr.umich.edu/icpsrweb/METLDB/holdings/documentation for additional information on scoring rubrics.

Within each of those domains, the research team identified talk practices within the rubric such that the score of the domain could be representative of classroom talk. For example, within the *Analysis and Problem Solving* domain, talk that supported higher-level thinking and problem solving was included within the rubric. Here, coders had looked for students explaining their thinking as well as developing arguments, providing explanations, constructing alternatives, and so forth. As such, the MET project score from the rater evaluating Analysis and Problem solving was used to represent classroom talk that involved students explaining thinking, providing explanations, constructing alternative, and so forth. Overall, within *Content Understanding*, communication of concepts and procedures was included. Within *Instructional Dialogue*, frequent conversations, open-ended questions, advanced language, and student engagement were included. Within *Quality of Feedback*, scaffolding, feedback loops, prompting thought processes, providing information, encouragement,

and affirmation were included. Lastly, within *Regard for Student Perspectives*, supports of student expression were included. Scores for *Positive Climate*, *Negative Climate*, *Teacher Sensitivity*, *Behavior Management*, *Productivity*, and *Instructional Learning Formats* were not used because they were not directly linked to classroom talk practices. Scores were averaged from two raters. Reliability across topics and throughout the school year has been shown for the CLASS (La Paro, Pianta, & Stuhlman, 2004). For our study, reliability was $\alpha = .917$ for the five items used.[1]

Protocol for language arts teaching observation (PLATOPrime; Grossman et al., 2010). Although the PLATOPrime contained six domains, only a single domain titled *Classroom Discourse* included classroom talk practices in the rubric, again rated on a scale of 1–7. Scoring criteria for Classroom Discourse focused on the opportunities students have for conversations with the teacher and among peers. This domain looks at uptake, or the extent to which the teacher engages students' ideas and prompts them to clarify and specify their understandings. The remaining five domains—*Modeling*, *Strategy Use and Instruction*, *Intellectual Challenge*, *Time Management*, and *Behavior Management*—were not used in the current study because there was no mention of talk in their scoring rubrics. Reliability (i.e., Cronbach's α) of the two items (across two raters) was found to be $\alpha = .763$.[2]

State standardized English language arts assessment (ELA). Scores from the state standardized language arts exams were used from the Spring before the study as the ELA pretest (ELA_{pre}) and from the Spring of the study year as the ELA posttest (ELA_{post}). Although the tests varied by the state in which the students learned, we compared the test blueprints and aligned the assessed constructs using the technical manuals for each state to determine comparability of the state tests (New York State Testing Program, 2010; Technical Report for 2010 FCAT Test Administrations, 2010; Technical Report for the 2010 PISA, 2010; Technical Report TCAP, 2010).[3] Our review suggests the states have similar conceptualizations and ways of assessing reading. For example, all assessments included passages with comprehension questions capturing literal and inferential comprehension of both literary and informational texts assessed via mostly multiple-choice responses and some open-ended responses that were scored via rubrics. As such, we determined that there was evidence of content comparability across state assessments.

Because different tests were used, no single reliability was provided, although technical manuals from the states show high internal consistency for the states' tests across Grades 4 and 5: 0.83–0.86 in New York (New York State Testing Program, 2010), 0.90–0.91 in Pennsylvania (Technical Report for the 2010 PISA, 2010), 0.89–0.92 in Florida (Technical Report for 2010 FCAT Test Administrations, 2010), and 0.92 in Tennessee (Technical Report TCAP, 2010). Each state's test reports also include extensive documentation of how they constructed their tests, including detailed item development processes, which suggests the validity of the comprehension measure. Additionally, these tests are practically valid, as these were the tests that states and districts were using to make high-stakes decisions that link to retention and tracking decisions. No item-level data were available in the dataset. To standardize scores across measures, each student's rank-based *z* score was used.

Stanford 9 Open-Ended Reading Assessment (SAT). The Stanford 9 Open-Ended Reading Assessment is a nationally normed achievement test (National Research Council, 1999) that involved student-constructed responses to reading. Because the SAT 9 scores are nationally normed, scores from different states are comparable. This assessment was chosen as a contrast to the multiple choice, basic skill nature of state tests and instead represents "cognitively demanding test content [that] presented students with constructed response items" (MET user guide, Inter-University Consortium for Political and Social Research, p. 26). According to the MET project user guide, "The assessment presented students with one extended reading passage [narrative] and then asked them to respond to nine, open-ended tasks (which required students to provide short, written responses to comprehension questions)." The nine tasks involved higher-order thinking responses describing, summarizing, analyzing, and evaluating the passages and then explaining their thinking behind each response. For interpretation, scale scores ($M = 608.2$, $SD = 37.4$) were used. Extensive research work documents the SAT-9s reliability and validity (Harcourt Brace & Company, 1997), with reliability coefficients reported between .94 and .96 for Grades 2 through 11 (Rogosa, 1999). The MET project selected the SAT-9 as a complement to state tests because they "included cognitively demanding content, they were reasonably well-aligned with the curriculum in the six states, had high levels of reliability, and had evidence of fairness to members of different groups of students." (Gates Foundation, 2010, p. 11).

Data Analysis

Our methodological choices were guided by our research questions, precedent from the literature, and the constraints of the MET project data and the models we used. First, our sample was limited to teachers for whom we had some data on classroom talk and student outcomes (ELA_{pre}, ELA_{post}, and SAT). As most intraclass correlations (ICCs) were above 0.05, multilevel modeling was used to deal with clustering of students nested within class sections nested within schools (see Appendix A for ICC values).[4]

To examine our first research question regarding how different measures of classroom talk relate, multilevel exploratory

[1] In the MET project, reliability for raters was low for the full CLASS measure with 8% of the total variance in scores because of main rater effects (Bill & Melinda Gates Foundation, 2012, p. 35). We took this into account via our modeling framework. As will be described later, we extracted two factors ("rater observation of teachers" and "student observation of teachers") using all 23 Tripod, CLASS, and PLATO items at the section level. The two factors from the multilevel EFA are assumed to be measurement-error-free variables and were used in the multivariate multilevel model. The multilevel composite reliability (Geldhof et al., 2014) for the two factors was 0.863 and 0.979, respectively, suggesting good reliability.

[2] Reliability for raters was found to be low (10% of the total variance in scores because of main rater effects, Bill & Melinda Gates Foundation, 2012, p. 35), which again was taken into account via our modeling framework (i.e., multilevel EFA assumed to be measurement-error-free variables and which together with the CLASS items resulted in a factor with 0.863 multilevel composite reliability (Geldhof et al., 2014).

[3] States include Florida, Tennessee, New York, and Pennsylvania, see http://www.metproject.org/downloads/Student_Assessments_92110.pdf for details. Note that a concern by the education research community with these tests is that such state assessments often measure mostly basic skills.

[4] Because a teacher taught one or two sections (4% of teachers taught two sections), clustering because of sections was chosen instead of teachers because section clustering is the lower level than teacher clustering. Also, district clustering was found to be ignorable and, therefore, not considered.

factor analysis (EFA) of the 23 indicators of classroom talk (16 Tripod, 5 CLASS, and 2 PLATO) was used to establish a measurement model for classroom talk.[5] EFA was chosen as our close review of the literature indicated no clear theoretical guidance in terms of the makeup of classroom talk. As such, methodological guidelines indicated EFA would be preferable to confirmatory factor analysis (CFA; Asparouhov & Muthén, 2009; Browne, 2001). Because school-level clustering can be ignored for the Tripod indicators (ICC for schools = 0.039 compared with nonignorable section-level clustering ICC = 0.183 as shown in Appendix A) and three-level EFA is not feasible in currently available software, two different two-level EFAs were run: (a) students nested within sections and (b) sections nested within schools. In addition, because the CLASS and PLATO are section-level observations and, thus, there were no student-level data in the rater observations of teachers, these measures were specified at the section level only. To summarize our multilevel structure, the student-level data (Level 1; the student-level Tripod scores) are nested within the section-level data (Level 2; the section-level Tripod, CLASS, and PLATO scores), and the section-level data are nested within the school-level data (Level 3; the school-level CLASS and PLATO scores; see Figure 1). To determine the best fitting model, both theory and results for fit indices were used. Fit guidelines included root mean square error of approximation index (RMSEA; Steiger & Lind, 1980) less than .06, the root mean square residual (RMSR) less than .08, and comparative fit index (CFI; Bentler, 1990) and Tucker-Lewis index (TLI; Tucker & Lewis, 1973) larger than .95. The robust weighted least squares estimator using a diagonal weight matrix (Asparouhov & Muthén, 2007) with GEOMIN rotation were used to fit multilevel EFAs in M*plus* Version 7 (Muthén & Muthén, 1998–2012). Multilevel composite reliabilities for each factor were calculated (Geldhof, Preacher, & Zyphur, 2014).

To explore our second research question, multivariate multilevel modeling was used. In the model, latent variables modeled via the measurement model from the first research question were used to predict the two measures of standardized reading performance at the student, section, and school levels. We used standardized observed outcomes as no item-level data was available for each measure and because our research question examined potential different relations between classroom talk and the different measures of standardized reading achievement, hence no creation of a latent variable from the two standardized reading outcomes (i.e., SAT and ELA). Also, variability in student background characteristics and prior skills were controlled for by entering dummy-coded demographic variables and prior scores on the state standardized ELA test as covariates.[6] The multivariate multilevel model is described in Appendix B and was fit using M*plus* with the Bayes estimator. Any participants with missing data on any student-level covariates were deleted using listwise deletion. This left 16,588 students learning within 822 class periods within 129 schools.[7] Significant coefficients for factors within the classroom talk measurement model were explored and interpreted based on 95% credible interval (CI) from the Bayes estimator. The explained variance of ELA and SAT outcomes by covariates was calculated by comparing results of the model with covariates with those of the model without covariates.

Results

Structure of Classroom Talk (RQ1)

Table 3 presents fit indices for 12 candidate models regarding the number of student-level and section-level factors using 23 indicators (16 Tripod, five CLASS, and two PLATO). Results for the single model regarding the number of section-level and school-level factors using seven indicators (five CLASS and two PLATO) are also shown. Combining the results of fit indices and factor loading interpretability (i.e., statistics and theory), the best fitting model included four factors at the student level, two factors at the section level, and one factor at the school level. Table 4 shows the factor loadings of the model and Table 5 details the structure. In Table 4, factor loadings are bolded to show the factor model in which they were ultimately assigned in the subsequent structural (see Figures 1 and 2) to answer Research Question 2. It is important to note that we used a combination of statistics and theory to identify the final model. When fit indices showed similarly good fit for multiple models and when factor loadings showed a similarly good fit for items across constructs, we considered theory and stability of items within latent constructs. For example, in the case of item S7 where students evaluated the statement "My teacher wants me to explain my answers," theory was used to decide which of the two factor loadings made more theoretical sense (i.e., the

[5] One Tripod item (S15) was not used because of negative wording that resulted in a negative discrimination value with the recoded S15.

[6] While measurement error can be a concern with observed versus latent variables, we followed the precedence of most large-scale studies (i.e., Applebee et al., 2003; Carlisle et al., 2011), which model relationships with observed, standardized reading outcomes, which would be expected to have less error than researcher measures. Also, it would have been ideal to have additional controls of student-level reading predictors like vocabulary performance, but the MET project lacked such data, and again we followed other studies by using an autoregressor (i.e., pre-test scores) to control for other factors like vocabulary or general reading achievement that might contribute to these relationships. Income differences were not controlled for because of missing data present for an entire school district related to free and reduced lunch status. Teacher background characteristics were not included as covariates because of methodological challenges and because they were not the focus of our study. As our interest included differences in classroom talk across class periods, not just teachers, the Level 2 variable was class period, not teacher. Including teacher demographics, therefore, was not possible because the data could be repetitive for some class periods taught by the same teacher. Furthermore, no school level demographics were included because such variables were not available in the MET Study database, and again, this was not the focus of our study. At the student level, White served as the reference group for all race/ethnicity dummy variables.

[7] The listwise deletion requires missing completely at random (MCAR) assumption. When MCAR assumption does not hold, parameter estimates can be biased (e.g., Enders, 2010). Comparisons of students with missing data in student-level covariates with students with no missing data in the student-level covariates revealed no psychometric differences between the groups. Specifically, we assessed MCAR using a series of *t* tests to compare missing subgroups (i.e., students with missing data in a student-level covariate vs. students with no missing data in a student-level covariate; Dixon, 1988). The series of the *t* test suggested that the group means are equivalent for student-level covariates we considered. This result supports for the MCAR assumption. In addition, to account for correlations among student-level covariates, we implemented a multivariate version of the *t* test that evaluates mean differences on every variable simultaneously across subgroups having the same missing data pattern (Little, 1988). Little's test also supports for the MCAR assumption.

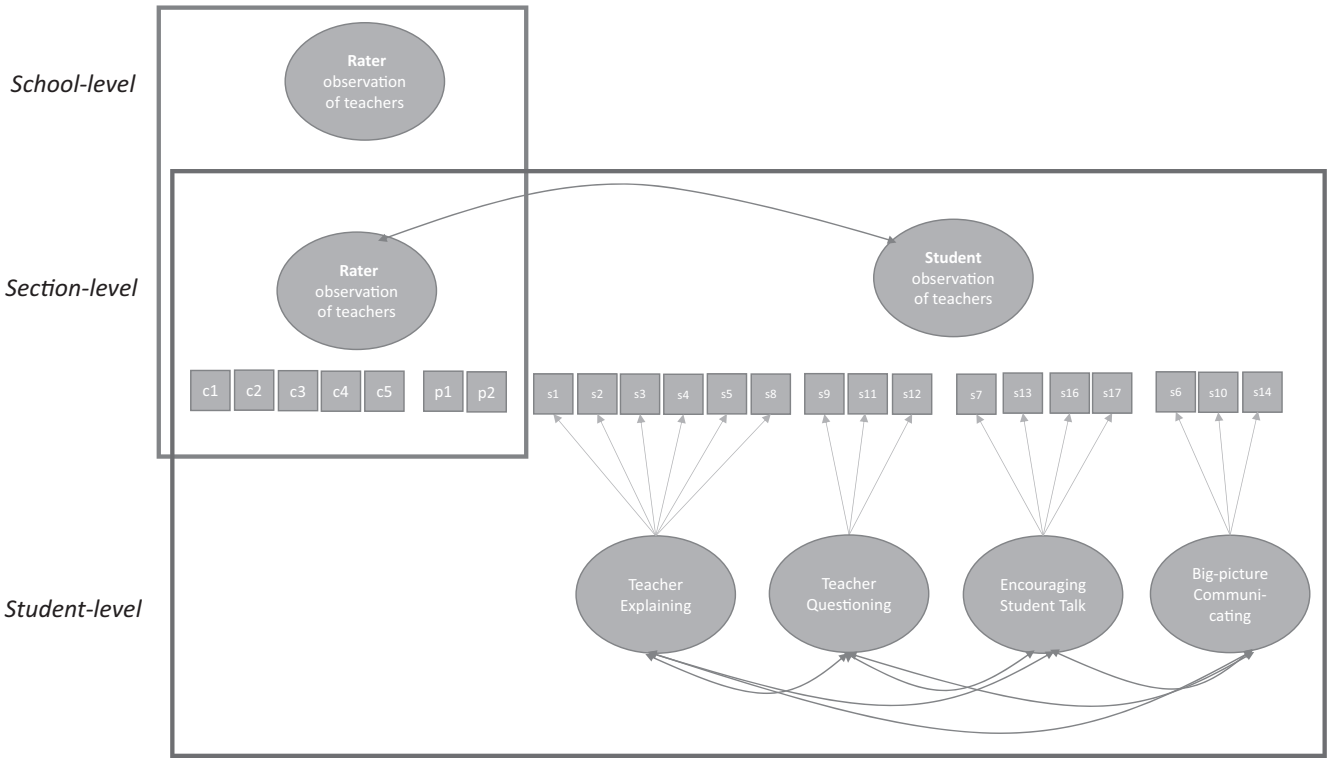

Figure 1. Multilevel measurement model compiled from two multilevel EFAs. There was no variance in Rater measures at the Student Level (teacher-level variables). Also, school level variance can be ignored for Student Observations (s1–s17; ICC = .04). Based on initial results, convergence problems related to three-level CFA, and the structure of our data, we resorted to running two EFAs (see box outlines).

factor loading for the latent variable F2 Teacher Questioning of 0.237, $p < .05$ vs. 0.324, $p < .05$ for the latent variable F3 Student Talk). In this case, because the content of the question related to student talk, in subsequent models (i.e., CFA and structural models in the multivariate multilevel model), S7 was assigned to F3. Overall, the four factors at the student level represent ratings of teacher explaining, teacher questioning, encouraging student talk, and big-picture communicating. For example, questions related to how a teacher explained classroom concepts loaded highest on the first factor, items related to teachers' questions loaded highest on the second factor, items supporting student talk loaded highest on the third factor, and items related to big-picture communicating like summarizing learning and providing a purpose for learning loaded highest on the fourth factor. The multilevel composite reliability (Geldhof et al., 2014) for each of these factors was 0.933, 0.892, 0.889, and 0.786, respectively. These factors were correlated between 0.560 and 0.701 with correlations lowest between the encouraging student talk factor and the others.

At the section level, student ratings and observer ratings represent the two factors. Ratings of classroom talk by independent observers, who were reflecting on a single lesson segment and reflecting on talk features as embedded in different instructional goals, load best on the first factor (rater observations) whereas student ratings, which consider classroom talk across content and across the year, load best on the second factor (student observations). The multilevel composite reliability (Geldhof et al., 2014) for each was 0.863 and 0.979, respectively. These factors are slightly yet significantly correlated ($r = .210$). At the school level, there is no significant variance to model for the student ratings, so the observer ratings represent the single factor at the school level. It is important to note that for the second multilevel EFA (sections nested in schools), which used the seven teacher-level indicators (five CLASS and two PLATO), only one model having a factor at each of the section and the school levels converged without an estimation problem. This may be because of the small number of section-level indicators. Even though other models having more than one factor were not obtained because of convergence problems, there is evidence supporting the model having a single factor. First, as shown in Table 4, the patterns of the factor loadings at the section level are similar between the multilevel EFA analysis using 23 indicators (16 Tripod, five CLASS, and two PLATO) and the multilevel EFA analysis using seven section-level indicators (five CLASS and two PLATO). Second, in the multilevel EFA analysis using seven section-level indicators, the between-level (i.e., school-level) fit index indicates evidence of a good fit (SRMS = 0.062 as shown in Table 3).

After the two multilevel EFA (the multilevel EFA for the student- and section-levels and the multilevel EFA for the section- and school-level) results were merged, the final measurement model for classroom talk practice includes four factors at the student level, two factors at the section level, and one factor at the school level (see Figure 1). At the student level,

Table 3
Fit Indices for Two-Level (Students Nested Within Sections) EFA Using 23 Indicators (5 CLASS + 2 PLATO + 16 Tripod) and Two-Level (Sections Nested Within Schools) EFA Using Seven Indicators (5 CLASS + 2 PLATO)

Number of factors		Fit indices				
			SRMS			
Student level	Section level	RMSEA	Within	Between	CFI	TLI
Two-level EFA with 23 indicators						
1	1	0.027	0.034	0.335	0.930	0.922
2	1	0.024	0.029	0.335	0.947	0.938
3	1	0.021	0.026	0.335	0.959	0.950
4	1	0.021	0.024	0.336	0.963	0.953
1	2	0.023	0.025	0.076	0.951	0.942
2	2	0.019	0.017	0.078	0.968	0.960
3	2	0.015	0.011	0.076	0.980	0.974
4	**2**	**0.014**	**0.008**	**0.075**	**0.984**	**0.977**
1	3	0.022	0.025	0.057	0.958	0.947
2	3	0.018	0.017	0.049	0.975	0.966
3	3	0.014	0.011	0.049	0.986	0.979
4	3	0.012	0.008	0.049	0.989	0.984
			SRMS			
Section level	School level	RMSEA	Within	Between	CFI	TLI
Two-level EFA with seven indicators						
1	**1**	**0.121**	**0.079**	**0.062**	**0.893**	**0.839**

Note. CLASS = classroom assessment scoring system; PLATO = protocol for language arts teaching observation; EFA = exploratory factor analysis; RMSEA = root mean square error of approximation; SRMS = standardized root mean square residual; CFI = comparative fit index; TLI = Tucker-Lewis index. Results of the sections nested within schools EFA (seven teacher-level indicators: five CLASS and two PLATO) converged without an estimation problem only for the model having a single factor at the section and school level. The number of factors we interpreted is in bold.

this structure indicates that students perceive different yet related patterns in talk regarding how teachers explain, question, encourage student talk, and communicate the big-picture related to learning. At the section level, results suggest student and independent observers rate classroom talk practices differently. At the school level, the lack of significant school level variance in student ratings (ICC = 0.039) suggests that independent observers witness differences in talk practices among schools, yet students do not. For example, observers would likely note differences between school cultures that emphasize accountable talk versus behavior management (where silent, independent learning is encouraged), whereas students did not seem to note such differences in school foci. Such findings confirm the need for different lenses to examine talk.

Investigating the Effects of Classroom Talk Practices on Reading (RQ2)

To explore links between the talk practices described above and standardized reading performance, the final measurement model found in RQ1 is used to explain the standardized reading outcomes controlling for prior test scores and student demographics. Results of the structural model are presented in Table 6. At the student level, results suggest that how students rate teacher explaining and teacher questioning significantly predicts both ELA_{post} and SAT ($\hat{\gamma}_{1.E} = 0.312$, 95% CI [0.188, 0.449] and $\hat{\gamma}_{1.S} = 5.957$, 95% CI [0.113, 12.078] for teacher explaining; $\hat{\gamma}_{2.E} = 0.199$, 95% CI [0.071, 0.341]; $\hat{\gamma}_{2.S} = 9.828$, 95% CI [3.711, 15.394] for teacher questioning) such that higher ratings predict higher standardized test scores. In contrast, how students rated teacher big-picture communicating was negatively related to both ELA_{post} and SAT ($\hat{\gamma}_{4.E} = -0.411$, 95% CI [−0.570, −0.267]; $\hat{\gamma}_{4.S} = -8.261$, 95% CI [−14.651, −1.396]). Also, how students rated teachers' encouraging of student talk did not significantly relate to either outcome ($\hat{\gamma}_{3.E} = 0.045$, 95% CI [−0.074, 0.173]; $\hat{\gamma}_{3.S} = -1.328$, 95% CI [−6.565, 4.240]). At the section level, controlling for prior test performance, students' overall ratings of classroom talk but not raters' observations predicted performance on the state standardized English Language Arts test ($\hat{\delta}_{2.E} = 0.188$, 95% CI [0.079, 0.302]; $\hat{\delta}_{1.E} = 0.008$, 95% CI [−0.046, 0.062]), although neither predicted performance on the open-response standardized reading test ($\hat{\delta}_{2.S} = 5.301$, 95% CI [−2.023, 12.517]; $\hat{\delta}_{1.S} = 0.131$, 95% CI [−3.137, 3.251]) when controlling for prior test scores. At the school level, there was not significant variance to model in student ratings, so student ratings were not included in the school level of the model. Observer ratings of talk-related indicators of classroom talk, though, predicted performance on the SAT ($\hat{\omega}_{1.S} = 60.922$, 95% CI [17.858, 313.364]) but not ELA_{post} ($\hat{\omega}_{1.E} = -0.005$, 95% CI [−0.508, 0.766]) controlling for prior test scores (ELA_{pre}). Note that at the student and section level, the two reading standardized tests (ELA_{post} and SAT) were significantly related, al-

Table 4
Factor Loadings and Factor Correlations for Multilevel EFAs

		Two-level (students nested within sections) EFA						Two-level (sections nested within schools) EFA	
		Level 1 (student)				Level 2 (section)		Level 1 (section)	Level 2 (school)
System	Item/measure	F1 teacher explaining	F2 teacher questioning	F3 student talk	F4 big-picture commun.	F1 rater obser.	F2 student obser.	F1 rater obser.	F1 student obser.
					Factor loading				
CLASS	C1	-	-	-	-	**0.828**[a]	−0.016	**0.805**[a]	**0.838**[a]
	C2	-	-	-	-	**0.748**[a]	0.058	**0.747**[a]	**0.732**[a]
	C3	-	-	-	-	**0.920**[a]	0.035	**0.922**[a]	**0.916**[a]
	C4	-	-	-	-	**0.907**[a]	−0.046	**0.888**[a]	**0.867**[a]
	C5	-	-	-	-	**0.792**[a]	0.021	**0.789**[a]	**0.741**[a]
PLATO	P1	-	-	-	-	**0.510**[a]	−0.029	**0.485**[a]	**0.513**[a]
	P2	-	-	-	-	**0.499**[a]	−0.008	**0.473**[a]	**0.578**[a]
Tripod	S1	**0.537**[a]	−0.102[a]	0.106[a]	0.000	0.385[a]	**0.448**[a]	–	–
	S2	**0.461**[a]	0.095[a]	0.002	−0.004	0.007	**0.890**[a]	–	–
	S3	**0.424**[a]	0.029	0.009	0.226[a]	0.054	**0.963**[a]	–	–
	S4	**0.590**[a]	0.066[a]	−0.044[a]	−0.005	0.070	**0.928**[a]	–	–
	S5	**0.233**[a]	0.044[a]	0.064[a]	0.265[a]	−0.161[a]	**0.939**[a]	–	–
	S6	0.024	0.174[a]	0.078[a]	**0.288**[a]	0.026	**0.734**[a]	–	–
	S7	−0.032[a]	0.273[a]	**0.324**[a]	0.029	0.211[a]	**0.695**[a]	–	–
	S8	**0.263**[a]	0.185[a]	0.047[a]	0.059[a]	0.061	**0.818**[a]	–	–
	S9	0.195[a]	**0.417**[a]	0.015	−0.014	−0.156[a]	**0.865**[a]	–	–
	S10	0.122[a]	0.237[a]	−0.016	**0.298**[a]	−0.141[a]	**0.989**[a]	–	–
	S11	−0.016	**0.575**[a]	0.039[a]	0.012	−0.191[a]	**0.936**[a]	–	–
	S12	0.136[a]	**0.505**[a]	0.013	0.055[a]	−0.076	**0.968**[a]	–	–
	S13	0.043[a]	0.019	**0.628**[a]	−0.021	0.403[a]	**0.556**[a]	–	–
	S14	−0.013	−0.004	0.020	**0.632**[a]	−0.252	**0.880**[a]	–	–
	S16	0.017	0.014	**0.432**[a]	0.131[a]	0.276[a]	**0.714**[a]	–	–
	S17	0.317[a]	0.014	**0.233**[a]	0.149[a]	0.216[a]	**0.868**[a]	–	–

		Factor correlations					
		F1	F2	F3	F4	F1	F2
	F1	1				1	
	F2	0.698[a]	1			0.210[a]	1
	F3	0.572[a]	0.560[a]	1		–	–
	F4	0.625[a]	0.701[a]	0.687[a]	1	–	–

Note. CLASS = classroom assessment scoring system; PLATO = protocol for language arts teaching observation; EFA = exploratory factor analysis; RMSEA = root mean square error of approximation; big-picture commun.= teacher big-picture communicating; - = not modeled because CLASS and PLATO were measured at the section level; – = not modeled because of small intraclass correlation (ICC) on Tripod items for schools (ICC = .039) and software limitation.
[a] Significance at the 5% level; factor loadings were GEOMIN rotated loadings; items used to interpret factors in bold.

though not at the school level. Overall, including the covariates in the model explained 63.8% of variance in performance on the state standardized English Language Arts test and 33.5% of variance in performance on the open-response standardized reading assessment.

Discussion

Classroom talk has been hypothesized to support literacy learning. The ways teachers and students express their ideas likely impacts reading comprehension. Much has been written unraveling specifics related to classroom talk, but what is less understood is how measurement affects the picture of classroom talk and its link to reading across a large number of upper elementary English Language Arts classrooms. Our study fills this gap.

Elementary Language Arts Classroom Talk

Findings indicate that students noticed patterns in their teachers' talk, and those patterns are similar to those found in the literature on discussions. Ratings of items assessing teachers' explanations were related, yet different from items assessing the quality of questions, encouragement of student talk, and communication of big-picture learning. Three of these categories of talk (teachers' quality of explanations, quality of questions, and encouraging student talk) have been explored in previous literature (e.g., Soter et al., 2008). However, there has been less attention in the literature to teachers' communication of big-picture ideas related to learning, which includes feedback on how to improve written work, statements on what was being learned and why, and summaries of what was learned by day. Though

Table 5
Description of Factor Structure of Classroom Talk

Level	Factor structure				
School level (one factor, seven CLASS and PLATO items)	No student ratings: No school level variance in these indicators				Observer ratings C1–5, P1–2
Section level (2 factors, 23 Tripod, CLASS, and PLATO items)	Student ratings S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S16, S17				Observer ratings C1: Analysis and problem solving C2: Content understanding C3: Instructional dialogue C4: Quality of feedback C5: Regard for Stud. perspect P1: Classroom discourse (seg 1) P2: Classroom discourse (seg 2)
Student-level (four factors, 16 Tripod items)	Teacher explaining	Teacher questioning	Teacher encouraging student talk	Teacher big-picture communicating	
	S1: My teacher is nice to me when I ask questions S2: If you don't understand something, my teacher explains it another S3: My teacher has several good ways to explain each topic that we cover in this class S4: My teacher explains difficult things clearly S5: My teacher explains things in very orderly ways S8: In this class, we learn to correct our mistakes	S9: When he/she is teaching us, my teacher asks us whether we understand S11: My teacher asks questions to be sure we are following along when he/she is teaching S12: My teacher checks to make sure we understand what he/she is teaching us	S7: My teacher wants me to explain my answers—why I think what I think S13: My teacher wants us to share our thoughts S16: Students speak up and share their ideas about class work S17: My teachers gives us time to explain our ideas	S6: When my teacher marks my work, he/she writes on my papers to help me understand how to do better S10: My teacher tells us what we are learning and why S14: My teacher takes the time to summarize what we learn each day	

Note. CLASS = classroom assessment scoring system; PLATO = protocol for language arts teaching observation.

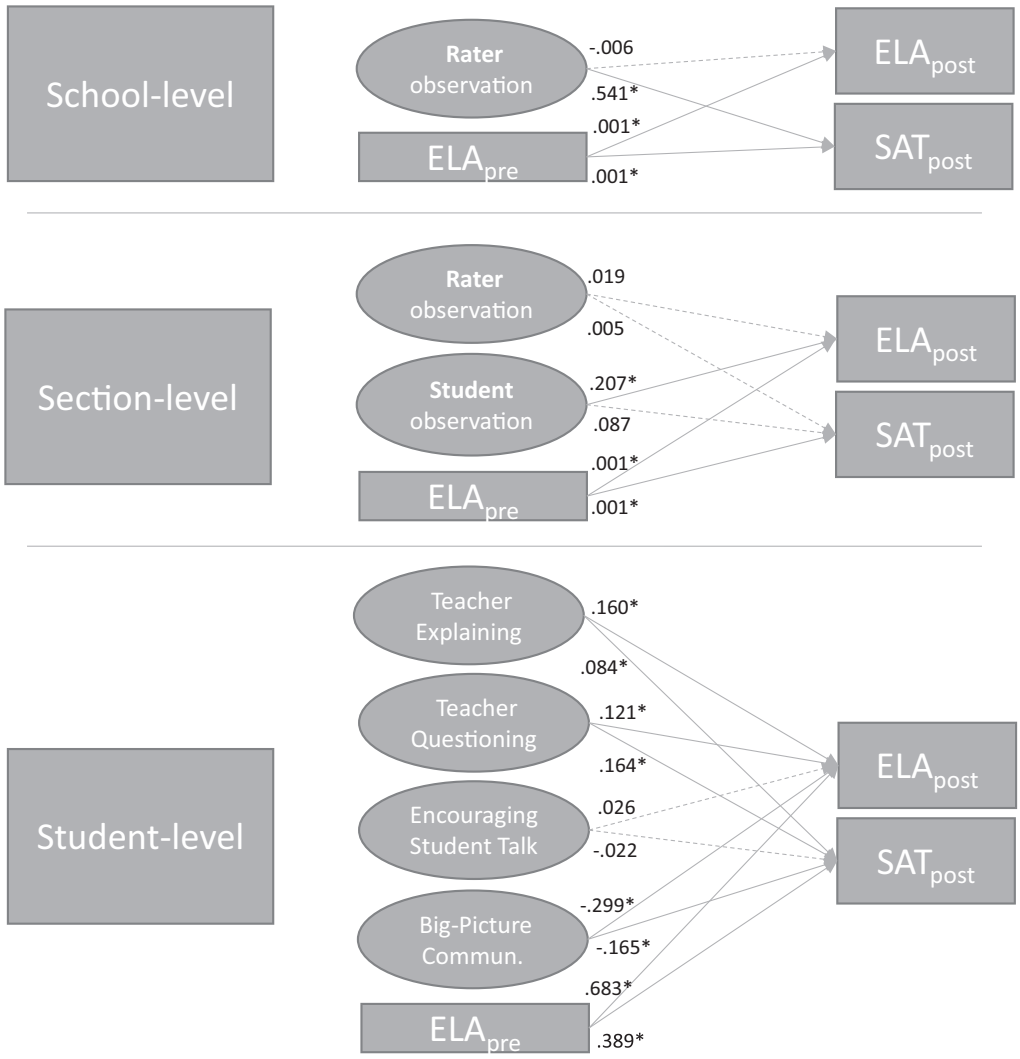

Figure 2. Results of a structural model in the multivariate multilevel model. All values are standardized. All models control for demographics. ELA = English language arts assessment; SAT = Stanford 9 Open-Ended Reading Assessment. * $p < .05$.

these aspects of teacher talk have been considered aspects of explanation in the past (Winograd & Chou, 1988), they were different from the features of explanations (i.e., clarity, order, and multiple options) in this study. This finding suggests that in the upper elementary English Language Arts context, broader ideas related to what is being learned and why it is being learned may be different from more specific aspects of explanations. This finding adds to the theoretical literature as it indicates the makeup of the interactions that are going on with more knowledgeable others that are mediating learning via linguistic supports (Vygotsky, 1978).

Findings also contrasted observers' and students' perceptions of talk, which provides additional evidence that multiple lenses exploring talk are valuable and confirms the importance of considering how classroom talk is measured. What students noticed related to classroom talk across the year and across content goals was different from what raters noticed in a segment of a lesson focused upon talk as related to different aspects of quality reading instruction. One possible explanation is that teachers may vary their talk somewhat across lessons and content areas such that what students were conveying was their perspective of the broad average of the classroom talk rather than the classroom talk occurring within a specific lesson or related to a specific content or context. For example, on a certain day an observer might see a lesson focused on vocabulary, which may highlight explanations of word meanings or questions regarding personal connections to words, yet involve less of a broad summary of learning. On a different day, an observer might rate a lesson teaching metacognitive reading strategies, which may involve discussions of why strategies may be important to the broader reading process. In contrast, a student would have experienced both of these lessons such that his or her ratings would average these experiences across the school year. Another explanation may involve individual versus globalized focus on talk (e.g., Connor et al., 2014). It may be that a student is interpreting both their individual and collective

Table 6
Results for the Multivariate Multilevel Model

Effects	ELA		SAT	
	EST	CI (LB, UB)	EST	CI (LB, UB)
Fixed effects	0.137[a]	[0.102, 0.169]	612.333[a]	[609.191, 165.421]
Intercept				
Student-level [$\gamma_{..}$]				
"Teacher explaining" [$\gamma_{1..}$]	0.312 [0.160][a]	[0.188, 0.449]	5.957 [0.084][a]	[0.113, 12.078]
"Teacher questioning" [$\gamma_{2..}$]	0.199 [0.121][a]	[0.071, 0.341]	9.828 [0.164][a]	[3.711, 15.394]
"Encouraging student talk" [$\gamma_{3..}$]	0.045 [0.026]	[−0.074, 0.172]	−1.360 [−0.022]	[−6.565, 4.244]
"Teacher big-picture communicating" [$\gamma_{4..}$]	−0.411 [−0.299][a]	[−0.570, −0.267]	−8.261 [−0.165][a]	[−14.651, −1.396]
ELA pretest scores [$\gamma_{5..}$]	0.687 [0.683][a]	[0.676, 0.698]	14.292 [0.389][a]	[13.721, 14.850]
Gender (male) [$\gamma_{6..}$]	−0.055 [−0.034][a]	[−0.073, −0.038]	−8.522 [−0.143][a]	[−9.407, −7.643)
African American [$\gamma_{7..}$]	−0.148 [−0.090][a]	[−0.180, −0.113]	−2.897 [−0.048][a]	[−4.518, −1.234]
American Indian [$\gamma_{8..}$]	−0.133 [−0.009]	[−0.296, 0.035]	−7.217 [−0.013]	[−15.783, 0.879]
Hispanic [$\gamma_{9..}$]	−0.050 [−0.026][a]	[−0.083, −0.015]	1.043 [0.015]	[−0.602, 2.750]
Asian [$\gamma_{10..}$]	0.105 [0.028][a]	[0.058, 0.154]	4.102 [0.030][a]	[1.762, 6.485]
Other [$\gamma_{11..}$]	−0.036 [−0.007]	[−0.092, 0.019]	1.690 [0.009]	[−1.038, 4.421]
ELL [$\gamma_{12..}$]	−0.094 [−0.039][a]	[−0.127, −0.060]	−0.591 [−0.007]	[−2.306, 1.107]
Special Ed [$\gamma_{13..}$]	−0.153 [−0.052][a]	[−0.187, −0.118]	−8.019 [−0.074][a]	[−9.689, −6.280]
Section-level [$\delta_{..}$]				
"Rater observations of teachers" [$\delta_{1..}$]	0.008 [0.019]	[−0.046, 0.062]	0.131 [0.005]	[−3.137, 3.251]
"Student observations of teachers" [$\delta_{2..}$]	0.188 [0.207][a]	[0.079, 0.302]	5.301 [0.087]	[−2.023, 12.517]
ELA pretest scores [$\delta_{3..}$]	0.817 [0.001][a]	[0.773, 0.862]	19.630 [0.001][a]	[16.868, 22.206]
School-level [$\omega_{..}$]				
"Rater observations of teachers" [$\omega_{1..}$]	−0.005 [−0.006]	[−0.508, 0.766]	60.922 [0.541][a]	[17.858, 313.364]
ELA pretest scores [$\omega_{2..}$]	0.872 [0.001][a]	[0.813, 0.932]	20.382 [0.001][a]	[12.895, 27.927]
Random effects				
Student-level [$\Sigma_{1(2\times2)}$]				
Variance	0.310	[0.302, 0.318]	689.012	[672.298, 706.298]
Covariance (EFA, SAT)		3.440, (3.151, 3.710)		
Section-level [$\Sigma_{2(2\times2)}$]				
Variance	0.024	[0.020, 0.029]	113.681	[97.123, 131.632]
Covariance (EFA, SAT)		0.779 (0.580, 1.00)		
School-level [$\Sigma_{3(2\times2)}$]				
Variance	0.008	[0.005, 0.013]	147.487	[47.580, 227.816]
Covariance (EFA, SAT)		−0.262 (−0.674, 0.148)		

Note. SAT = Stanford 9 Open-Ended Reading Assessment; EST = estimate; CI = 95% credibility interval from Bayes estimator; LB = lower bound; UB = upper bound; ELA = English language arts assessment; ELL = English language learner; EFA = exploratory factor analysis; values in bracket = results based on the standardization that uses the variances of the continuous latent variables ("teacher explaining," "teacher questioning," "encouraging of student talk," and "teacher big-picture communicating") as well as the variances of the student demographics and outcome variables (ELA and SAT).
[a] Significance at the 5% level.

experience with classroom talk whereas observers are more likely to pay attention to the collective experience of the class as a whole. Additionally, students may be communicating their perceptions of their relationships with teachers (i.e., their working alliance, Toste et al., 2010), which has been shown to relate to student performance. Overall, the method differences noted (e.g., different set of questions, different informants) show that multiple lenses provide meaningfully different data on classroom talk.

Link Between Classroom Talk and Reading

Findings from our study also suggest a clear relationship between classroom talk and standardized reading outcomes even when controlling for prior test scores and student demographics. And this relationship exists for more than 18,000 fourth and fifth graders learning from 745 teachers—a far larger sample than other quantitative studies of talk and comprehension (e.g., Applebee et al., 2003; Michener et al., 2018; Murphy et al., 2009). In other words, general classroom talk practices embedded within larger instruction and measured in different ways connected with end of the year reading performance on two different types of standardized reading assessments for a large sample of students. This indicates that general classroom talk practices (e.g., explanations and questions) matter in English Language Arts instruction.

Specifically, higher ratings of teacher explanations and questioning predicted higher scores on both of the standardized reading assessments. Theoretically, these findings suggest important nuances in the ways that more knowledgeable others support meaning making more broadly (Vygotsky, 1978) and also connections to reading comprehension (Alvermann et al., 2013; Gough & Tunmer, 1986). For example, these structures—explanations and questions—seem to provide the space for more knowledgeable others to model and scaffold effective language including content-specific vocabulary and complex syntactical structures as well as thinking that students can then emulate in their own meaning-making endeavors.

Michener and colleagues (2018), who also found that teacher explanations predicted upper elementary students' reading comprehension, hypothesized that, "explanations acted as linguistic exposure necessary for supporting students' linguistic comprehension for read-

ing" (p. 747). This highlights the importance of teachers explaining difficult concepts in a clear, orderly way, offering multiple explanations if need be, and using academic language not for its own sake but to explain increasingly complex content found as students transition from early elementary to middle school (Uccelli & Phillips Galloway, 2017). Here, the focus seems to be on explaining the "what" rather than the "why." The better the components of a skill are explained, the better the student seems to be able to apply that skill within the larger reading process. Similarly, effective teacher questioning is important, especially asking about understanding and using questions to help students follow along with content. It may be that MKOs' questioning as a talk practice serves as formative assessment: after eliciting student understandings, teachers can provide more effective explanations. Future work should continue to unravel what makes these specific talk structures effective in supporting reading achievement broadly.

A less expected finding was that students' perception of teachers' encouraging of student talk did not have a significant relationship with reading performance. One explanation may be found in Murphy et al.'s (2009) meta-analysis, which noted that simply increasing the proportion of student talk was not associated with increased comprehension—it appears that quality of student talk is more important than quantity. It may be that in our study, students were reporting more about their perceptions of quantity rather than quality of student talk. The prompts students were responding to when rating teacher performance focused on speaking up, sharing, and explaining, but not on critical thinking. An alternative explanation involves the variability of student perspectives: in an elementary classroom, some student talk may be on topic whereas other talk may be tangential or off topic and possibly even distracting. For example, students may have been encouraged to speak up and share, but if teachers did not keep that talk grounded in the text, students may come away believing that comprehension need not be anchored in text. Alternatively, it may be that MKOs' talk practices are more important to comprehension than the students' talk. The MKOs' explanations can simulate the authoritative voices of writers in texts, and it may be important for students' reading comprehension to listen to explanations. Overall, our findings suggest more research in this area is needed with more emphasis on examining quality of student talk.

Another finding involved teacher big-picture communicating (i.e., summarizing learning) negatively impacting reading performance. Again, we expect this may require distinguishing between quantity and quality of this kind of talk. For example, students may be able to assess the quality of teacher explanations of specific content better than assessing talk conveying why the learning is important because it is likely that students have less knowledge about connections between the learning and the larger picture of literacy and curricular development. In other words, upper elementary students might better understand *what* they are learning (i.e., teacher explanations) rather than *why* they are learning it (i.e., teacher big-picture communicating). Additionally, the wording of the items in this domain focused on quantity rather than quality, so perhaps teachers are consistently trying to communicate this big picture about what is being learned and why it is important, but that these actions fail to accurately convey the link between the learning and the larger picture. Duffy et al. (1986) showed that teachers often tended toward approaches that linked learning to knowledge acquisition (e.g., knowing terminology for a grammar rule), whereas fewer teachers achieved higher-level critical thinking (e.g., evaluating which reading comprehension strategy supports reading under what conditions). Our study suggests more work needs to be done unraveling how teachers connect learning of specific content to broader learning and links to improvements in reading performance.

Research Considerations

As part of our analysis, we were able to explore the links between student versus observer ratings of classroom talk as embedded within general instruction and the associations with two different measures of reading performance, both of which purport to measure different aspects of reading comprehension (i.e., more basic skills multiple choice vs. more cognitively demanding open response formats). We did this purposefully because key work like the RAND Reading Study Group (2002) emphasize considering differences in comprehension related to different tasks. Our results suggest that students are particularly important observers of talk that connects to reading performance. At the student level and the section level, student ratings linked to the various standardized reading assessments emphasizing the value of being present in a context throughout a year. Students were even able to see differences in talk practices within the same teacher's instruction across different class periods, and these differences explained reading performance on the state standardized language arts exam when controlling for prior test scores. In contrast, the raters' role as independent observers resulted in a less nuanced view of talk as raters did not discern differences in talk-related practices (as embedded within different high quality instructional practices) between different class periods. Raters' roles in rating talk practices within a single lesson did, though, allow them to notice meaningful differences between talk practices at schools that link to standardized reading performance, specifically performance on the open-response standardized reading assessment. While such lenses are helpful at the school level, our findings draw into question the use of rater evaluations of classroom talk to identify meaningful differences in talk practices between teachers.

We see these findings as relevant to multiple discussions within the literature on talk and comprehension. First, as noted in the literature review, classroom talk can be thought about globally as in what all students are experiencing as a community and it can also be thought about from the individual student perspective as in each student interprets that general talk uniquely and experiences different aspects of the talk, for example, a more nuanced explanation based on their expressed confusion to the teacher. Connor et al. (2014) found unique contributions of each, and our findings would support the call for exploring both as the student survey, which represented student's perspective on their unique classroom talk experiences predicted differences in student-level performances whereas raters, who are more focused on global talk, predicted school level differences in reading achievement.

Another contribution is in discussion of how to measure instructional quality, including the types of questions and the specifics of an observational protocol. While an entire paper could be devoted to this topic, we refer readers to Carlisle et al. (2011), which details different ways researchers have set out to measure literacy instructional quality and the challenges of various approaches and the various error associated with the different approaches. What is relevant to our study is that observational protocols can differ in whether counts of instruction or judgments of instruction occur as well as the expected frequency of key behaviors. For example, many of the variables related to students

outcomes are observed infrequently (in the study they discuss, less than 3 to 5% of the time), which may make observation protocols more prone to error. We followed their guidance in looking at actions via dimensions, but still, an early analysis of the Measures of Effective Teaching Project's video and classroom quality measures suggested that for the approximately 3,000 teachers in their sample, questioning and discussion techniques and communicating with students were the lowest rated areas of performance for teachers (Bill & Melinda Gates Foundation, 2013), indicating that infrequent use of classroom talk structures may also be related to our findings. Another component to consider is the focus of the tool. In our study, the student survey allowed students to communicate their perceptions of specific features of classroom talk, whereas the observer-rated items included talk practices in addition to other aspects of high quality instruction within that observational category. It may be that the more nuanced data on talk practices (i.e., specifically conveying data on features rather than a feature within a larger rubric) is more meaningful to understanding talk practices occurring within elementary classrooms and their links to reading achievement. Future research should explore whether observer ratings on specific talk features (similar to the student items) link more closely to reading gains.

Limitations and Future Directions

Our study took advantage of the large-scale database provided by the Bill and Melinda Gates Foundation's MET Study (Bill & Melinda Gates Foundation, 2012; Gates Foundation, 2010). This allowed the exploration of average trends regarding the relationship between different measures of classroom talk and reading comprehension at a scale not considered previously (745 fourth- and fifth-grade teachers and 18,844 students from five different school districts across different states and regions). It also, though, resulted in certain methodological challenges such as lack of details (like when exactly assessments were administered or even some reliability information), item-level data, and certain measures (additional predictors including SAT pretest scores). Clearly, using extant data can be challenging, but the scale of this data made up for such obstacles. Below, we discuss some of the challenges.

First, because the larger study was interested in the relationship between instructional quality and broad outcomes (reading comprehension and math), the dataset did not include predictors of reading and classroom talk like vocabulary knowledge that may underlie the general relationship. As such, while we were able to identify the relationship between classroom talk and reading comprehension at scale, we hope future research will explore variables that might impact classroom talk that could then impact reading. For example, future studies may unravel whether classroom talk explains variation in reading comprehension because it is a proxy for and reflects variation in student's vocabulary levels. Research highlights the variability in classroom and teacher talk (Applebee et al., 2003; Gámez & Lesaux, 2015), so it seems likely that these relationships are more complex, and future work would be helpful in unraveling what we see as theoretically justified predictors of both talk and comprehension that might portray the classroom talk and reading comprehension relationship in a more nuanced way than our study was able to. These variables might include vocab and oral language skill (student level), experience and quality (teacher/class period/section level), and instructional approaches and organizational structures (school level). The present study establishes the relationship between classroom talk and reading comprehension in a large scale sample, investigating potential differences because of different ways these constructs were measured. This is the first and important step in a longer line of research that should next investigate the other variables at play.

Also, because the larger study was interested in general classroom quality, a second methodological constraint involved the fact that many items were not designed to identify specific differences in classroom talk. For example, the CLASS and PLATO items used in the MET study included reference to classroom talk practices and other instructional practices, making it hard to isolate classroom talk practices. Additionally, average or composite scores for some measures (i.e., the CLASS) were provided whereas item-level data would have allowed us to look at more nuanced relationships within these measures using a more sophisticated measurement model. Additionally, although the student Tripod survey items considered talk across the academic year, the rater classroom observations (i.e., CLASS and PLATO) were based on the first 30 min of a single videotaped lesson. As such, our study likely highlighted classroom talk from the beginning and middle of a lesson, rather than end of a lesson. This means that talk like what was learned and why associated with a lesson's closure may not have been captured (Hattie, 2009). Future research should consider the whole lesson and perhaps explore different talk practices occurring within different parts of the lesson. Also, we considered the lessons submitted by teachers, which represent the classroom talk related to the specific lessons, content, and days for which the videos (up to four) were submitted. Future work should include observations of more lessons, especially as our findings show the raters' observations of the videotaped lessons differed from the students' evaluations of talk practices throughout the year.

Another methodological challenge related to the multilevel nature of the data. We had data of students nested within teachers nested within sections nested within schools. Because of redundancy, the results of our analyses must be interpreted as differences between classroom talk in different class periods. As such, we could not consider any Level 2 teacher variables as this would be redundant in our models for teachers who taught more than one class period. This was not part of our research question, but is something for future research to consider. Also, because multilevel EFA is currently limited to two-level models, we join calls for advancement in software to allow for three-level EFA. Additionally, while we used a multivariate three-level structural model based on statistics and theory in our study, future studies should continue to consider alternative models and replicate our findings as well.

Another methodological limitation involves our use of outcome variables as observed indicators rather than latent variables. While we would have preferred to use latent variables that are assumed to be free from measurement error, such an approach was not possible with our dataset as we did not have access to item-level data for the outcomes. We did have two measures of reading achievement, but we choose not to combine them into a latent variable because we wanted to explore links to these separate reading constructs. With that said, each of these observed reading outcomes were standardized reading measures and use of standardized reading outcomes as observed indicators has precedence (e.g., Applebee et al., 2003; Carlisle et al., 2011). Additionally, using such measures is practically meaningful as these are the tests that districts and states use to guide high-stakes decisions like retention and tracking. Future large scale databases, though, should consider providing item-level data so that latent variables of these separate reading constructs can be modeled.

We note here that we make a distinction between measurement error in these standardized reading outcomes and the more challenging reliability for teacher outcomes. The, literature has shown that it is challenging to have high reliability of teacher observations and the METproject's experience was no different (Bill & Melinda Gates Foundation, 2012). We echo the call for more work in reliably measuring teacher observations.

Overall, our study makes important contributions to the literature in terms of understanding the makeup of elementary English Language Arts classroom talk and links to reading performance, unraveling nuances related to how these constructs were measured and considering these questions at a scale not previously explored. Future research would benefit from replication and also more nuanced exploration of additional predictors. For example, we controlled for student demographics within our analysis, but we did not examine differences in classroom talk for classrooms with different proportions of certain types of students including students of different racial/ethnic backgrounds, different skill levels, and different language backgrounds. Additionally, we did not consider teacher or school level variables as our focus was purely on identifying the makeup of this talk and its relationship to reading comprehension. Furthermore, our study explored relationships between talk across topics, but future research would benefit from looking specifically at how these talk practices are enacted within specific classrooms and how they are enacted differently across content and across learning groups of different backgrounds and characteristics. Overall, our study suggests that classroom talk, specifically how teachers explain and question, matters to reading achievement, and indicated that continued research on classroom talk and its connection to student outcomes is warranted.

References

Alexander, R. J. (2008). *Essays on pedagogy*. London, UK: Routledge.

Alvermann, D. E., Unrau, N. J., & Ruddell, R. B. (Eds.). (2013). *Theoretical models and processes of reading* (6th ed.). Newark, DE: International Reading Association.

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,* 685–730. http://dx.doi.org/10.3102/00028312040003685

Asparouhov, T., & Muthén, B. (2007, July). Computationally efficient estimation of multilevel high-dimensional latent variable models. *Proceedings of the 2007 JSM meeting in Salt Lake City, UT, Section on Statistics in Epidemiology* (pp. 2531–2535).

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16,* 397–438. http://dx.doi.org/10.1080/10705510903008204

Bailey, A. L. (2007). *The language demands of school: Putting academic English to the test*. New Haven, CT: Yale University Press.

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays* (C. Emerson & M. Holquist, Trans.). Austin: University of Texas Press.

Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2013). Examining the structure of reading comprehension: Do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing, 26,* 349–379. http://dx.doi.org/10.1007/s11145-012-9372-9

Bentler, P. M. (1990). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 107,* 238–246. http://dx.doi.org/10.1037/0033-2909.107.2.238

Bill & Melinda Gates Foundation. (2012). *Gathering feedback for teaching*. Bill & Melinda Gates Foundation. Retrieved from http://k12education.gatesfoundation.org/download/?Num=2530&filename=MET_Gathering_Feedback_for_Teaching_Summary1.pdf

Bill & Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three year study*. Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Boyd, M., & Rubin, D. (2006). How contingent questioning promotes extended student talk: A function of display questions. *Journal of Literacy Research, 38,* 141–169. http://dx.doi.org/10.1207/s15548430jlr3802_2

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36,* 111–150. http://dx.doi.org/10.1207/S15327906MBR3601_05

Carlisle, J., Kelcey, B., Berebitsky, D., & Phelps, G. (2011). Embracing the complexity of instruction: A study of the effects of teachers' instruction on students' reading comprehension. *Scientific Studies of Reading, 15,* 409–439. http://dx.doi.org/10.1080/10888438.2010.497521

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

Connor, C. M., Spencer, M., Day, S. L., Giuliani, S., Ingebrand, S. W., McLean, L., & Morrison, F. J. (2014). Capturing the complexity: Content, type, and amount of instruction and quality of the classroom learning environment synergistically predict third graders' vocabulary and reading comprehension outcomes. *Journal of Educational Psychology, 106,* 762–778. http://dx.doi.org/10.1037/a0035921

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific studies of reading, 10,* 277–299.

Daniel, S. M., Martin-Beltrán, M., Peercy, M. M., & Silverman, R. (2015). Moving beyond yes or no: Shifting from over-scaffolding to contingent scaffolding in literacy instruction with emergent bilingual students. *TESOL Journal*. Advance online publication. http://dx.doi.org/10.1002/tesj.213

Dickinson, D. K., & Porche, M. V. (2011). Relation between language experiences in preschool classrooms and children's kindergarten and fourth-grade language and reading abilities. *Child Development, 82,* 870–886. http://dx.doi.org/10.1111/j.1467-8624.2011.01576.x

Dixon, W. J. (1988). *BMDP statistical software*. Los Angeles: University of California Press.

Duffy, G. G., Roehler, L. R., & Rackliffe, G. (1986). How teachers' instructional talk influences students' understanding of lesson content. *The Elementary School Journal, 871,* 3–16. http://dx.doi.org/10.1086/461476

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Ferguson, R. F. (2008). *The TRIPOD Project framework*. Cambridge, MA: Harvard University.

Florit, E., & Cain, K. (2011). The simple view of reading: Is it valid for different types of alphabetic orthographies? *Educational Psychology Review, 23,* 553–576. http://dx.doi.org/10.1007/s10648-011-9175-6

Gámez, P. B., & Lesaux, N. K. (2015). Early-adolescents' reading comprehension and the stability of the middle school classroom-language environment. *Developmental Psychology, 51,* 447–458. http://dx.doi.org/10.1037/a0038868

Gates Foundation. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://k12education.gatesfoundation.org/download/?Num=2537&filename=Preliminary_Findings-Research_Paper.pdf

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19,* 72–91. http://dx.doi.org/10.1037/a0032138

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7,* 6–10. http://dx.doi.org/10.1177/074193258600700104

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,* 293–303. http://dx.doi.org/10.3102/0013189X14544542

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010, May). *Measure for measure: The relationship between measures of instructional practice in middle school English Language Arts and teachers' value-added scores* (NBER Working Paper No. 16015). http://dx.doi.org/10.3386/w16015

Hamre, B. K., & Pianta, R. C. (2007). Learning opportunities in preschool and early elementary classrooms. In R. C. Pianta, M. J. Cox, & K. L. Snow (Eds.), *School readiness and the transition to kindergarten in the era of accountability* (pp. 49–83). Baltimore, MD: Brookes Publishing.

Harcourt Brace & Company. (1997). *Stanford Achievement Test Series—Ninth Edition: Technical data report*. San Antonio, TX: Harcourt Brace & Company.

Hattie, J. (2009). *Visible learning: A synthesis of Meta-analyses in education*. London, UK: Routledge.

Hogan, K. E., & Pressley, M. E. (1997). Scaffolding scientific competencies within classroom communities of inquiry. In K. E. Hogan & M. E. Pressley (Eds.), *Scaffolding student learning: Instructional approaches and issues*. Boston, MA: Brookline Books.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2,* 127–160. http://dx.doi.org/10.1007/BF00401799

Langer, J. A. (1995). *Envisioning literature: Literary understanding and literature instruction*. New York, NY: Teachers College Press.

La Paro, K., Pianta, R., & Stuhlman, M. (2004). Classroom assessment scoring system (CLASS): Findings from the pre-k year. *The Elementary School Journal, 104,* 409–426. http://dx.doi.org/10.1086/499760

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83,* 1198–1202. http://dx.doi.org/10.1080/01621459.1988.10478722

McElhone, D. (2012). Tell us more: Reading comprehension, engagement, and conceptual press discourse. *Reading Psychology, 33,* 525–561. http://dx.doi.org/10.1080/02702711.2011.561655

Mercer, N. (1995). *The guided construction of knowledge: Talk amongst teacher and learners*. Clevedon, England: Multilingual Matters.

Mercer, N. (2000). *Words and minds: How we use language to think together*. London: Psychology Press.

Mercer, N. (2008). The seeds of time: Why classroom dialogue needs a temporal analysis. *Journal of the Learning Sciences, 17,* 33–59. http://dx.doi.org/10.1080/10508400701793182

Mercer, N., & Littleton, K. (2007). *Dialogue and the development of children's thinking: A sociocultural approach*. London, UK: Routledge. http://dx.doi.org/10.4324/9780203946657

Michener, C. (2014). *Features of dialogic instruction in upper elementary classrooms and their relationships to student reading comprehension* (Unpublished doctoral dissertation). Boston College, Chestnut Hill, MA.

Michener, C. J., Proctor, C. P., & Silverman, R. D. (2018). Features of instructional talk predictive of reading comprehension. *Reading and Writing, 31,* 725–756. http://dx.doi.org/10.1007/s11145-017-9807-4

Murphy, P. K., Wilkinson, I. A. G., Soter, A. O., Hennessey, M. N., & Alexander, J. F. (2009). Examining the effects of classroom discussion on students' comprehension of text: A meta-analysis. *Journal of Educational Psychology, 101,* 740–764. http://dx.doi.org/10.1037/a0015576

Muthén, L. K., & Muthén, B. O. (1998–2012). Mplus (Version 7) [Computer software]. Retrieved from www.statmodel.com

National Research Council. (1999). *Uncommon measures: Equivalence and linking among educational tests*. Washington, DC: The National Academics Press.

New York State Testing Program. (2010). *New York State Testing Program 2010: English Language Arts, Grades 3–8 Technical Report*. Monterey, CA: CTB/McGraw-Hill.

Ninio, A., & Bruner, J. (1978). The achievement and antecedents of labelling. *Journal of Child Language, 5,* 1–15. http://dx.doi.org/10.1017/S0305000900001896

Nystrand, M. (1997). *Opening dialogue: Understanding the dynamics of language and learning in the English classroom*. New York, NY: Teachers College Press.

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

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

Nystrand, M., Wu, L., Gamoran, A., Zeiser, S., & Long, D. A. (2003). Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse Processes, 35,* 135–198. http://dx.doi.org/10.1207/S15326950DP3502_3

Perfetti, C. A. (1988). Verbal efficiency in reading ability. In G. E. MacKinnon, T. G. Waller, & M. Daneman (Eds.), *Reading research: Advances in theory and practice* (Vol. 6, pp. 109–143). New York, NY: Academic Press, Inc.

Porter, A. C., & Brophy, J. (1988). Synthesis of research on good teaching: Insights from the work of the Institute for Research on Teaching. *Educational Leadership, 45,* 74–85.

RAND Reading Study Group. (2002). *Reading for understanding: Toward a research and development program in reading comprehension*. Pittsburgh, PA: Office of Educational Research and Improvement. Retrieved from http://www.rand.org/pubs/monograph_reports/MR1465/MR1465.pdf

Rex, L. A., & McEachen, D. (1999). If anything is odd, inappropriate, confusing, or boring, it's probably important" The emergence of inclusive academic literacy through English classroom discussion practices. *Research in the Teaching of English, 3,* 65–129.

Roehler, L. R., & Cantlon, D. J. (1997). Scaffolding: A powerful tool in social constructivist classrooms. In K. E. Hogan & M. E. Pressley (Eds.), *Scaffolding student learning: Instructional approaches and issues*. Cambridge, MA: Brookline Books.

Roehler, L. R., & Duffy, G. G. (1984). Direct explanation of comprehension processes. In G. G. Duffy, L. R. Roehler, & J. Mason (Eds.), *Comprehension instruction: Perspectives and suggestions* (pp. 265–280). New York, NY: Longman.

Rogosa, D. (1999). *Accuracy of individual scores expressed in percentile ranks: Classical Test Theory calculations*. Stanford, CA: Stanford University.

Rosenshine, B. (1983). Teaching functions in instructional programs. *The Elementary School Journal, 83,* 335–351. http://dx.doi.org/10.1086/461321

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23,* 441–474.

Shanahan, T., Callison, K., Carriere, C., Duke, N. K., Pearson, P. D., Schatschneider, C., & Torgesen, J. (2010). *Improving reading comprehension in kindergarten through 3rd grade: A practice guide* (NCEE 2010–4038). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from whatworks.ed.gov/publications/practiceguides

Silverman, R. D., Proctor, C. P., Harring, J. R., Doyle, B., Mitchell, M. A., & Meyer, A. G. (2014). Teachers' instruction and students' vocabulary

and comprehension: An exploratory study with English monolingual and Spanish–English bilingual students in Grades 3–5. *Reading Research Quarterly, 49,* 31–60. http://dx.doi.org/10.1002/rrq.63

Sinclair, J. M., & Coulthard, M. (1975). *Towards an analysis of discourse: The English used by teachers and pupils*. Oxford, UK: Oxford UP.

Snow, C. E., & Kim, Y. (2007). Large problem spaces: The challenge of vocabulary for English language learners. In R. K. Wagner, A. E. Muse, & K. R. Tannenbaum (Eds.), *Vocabulary acquisition: Implications for reading comprehension* (pp. 123–129). New York, NY: Guilford Press.

Soter, A., Wilkinson, I., Murphy, P., Rudge, L., Reninger, K., & Edwards, M. (2008). What the discourse tells us: Talk and indicators of high-level comprehension. *International Journal of Educational Research, 47,* 372–391. http://dx.doi.org/10.1016/j.ijer.2009.01.001

StataCorp. (2017). *Stata statistical software: (Release 15)*. College Station, TX: StataCorp LLC. Retrieved from https://www.stata.com

Steiger, J. H., & Lind, J. (1980, May). *Statistically-based tests for the number of common factors*. Paper presented at the Annual Spring Meeting of the Psychometric Society, IA.

Technical Report for 2010 FCAT Test Administrations. (2010). *Reading and Mathematics Technical Report for 2010 FCAT Test Administrations*. San Antonio, TX: Pearson. Retrieved from http://fcat2.fldoe.org/fcatpub5.asp

Technical Report for the 2010 PISA. (2010). *Technical Report for the 2010 Pennsylvania System of School Assessment*. Data Recognition Corporation. Retrieved from https://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PSSA/Technical%20Reports/2010%20PSSA%20Technical%20Report.pdf

Technical Report Tennessee Comprehensive Assessment Program (TCAP), Academic Year 2009–2010. (2010). *Technical Report Tennessee Comprehensive Assessment Program (TCAP) Achievement Test/ English Linguistically Simplified Assessment (ELSA) Test, Academic Year 2009–2010*. Pearson/TN DOE.

Toste, J. R., Heath, N. L., & Dallaire, L. (2010). Perceptions of classroom working alliance and student performance. *Alberta Journal of Educational Research, 56,* 371–387.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10. http://dx.doi.org/10.1007/BF02291170

Uccelli, P., & Phillips Galloway, E. (2017). Academic language across content areas: Lessons from an innovative assessment and from students' reflections about language. *Journal of Adolescent & Adult Literacy, 60,* 395–404. http://dx.doi.org/10.1002/jaal.553

Vygotsky, L. (1978). Interaction between learning and development. *Readings on the Development of Children, 23,* 34–41.

Vygotsky, L. S. (1986). *Thought and language*. Cambridge, MA: MIT Press.

Wells, G. (1993). Reevaluating the IRF sequence: A proposal for the articulation of theories of activity and discourse for the analysis of teaching and learning in the classroom. *Linguistics and Education, 5,* 1–37. http://dx.doi.org/10.1016/S0898-5898(05)80001-4

Winograd, P., & Chou, H. V. (1988). Direct instruction of reading comprehension strategies: The nature of teacher explanation. In C. Weinstein, E. Goetz, & P. Alexander (Eds.) *Learning and study strategies: Issues in assessment, instruction, and evaluation* (pp. 121–139). San Diego, CA: Academic Press, Inc.

Wolf, M. K., Crosson, A. C., & Resnick, L. B. (2005). Classroom talk for rigorous reading comprehension instruction. *Reading Psychology, 26,* 27–53. http://dx.doi.org/10.1080/02702710490897518

(*Appendices follow*)

Appendix A

ICC for Each Measure

System	Variance components			ICC	
	Sections	Schools	Residuals	Sections	Schools
		Measures of classroom talk practices			
CLASS	—	1.833	3.524	—	0.342
PLATO	—	0.200	0.468	—	0.299
Tripod	11.859	3.331	70.647	0.138	0.039
		Measures of reading achievement			
ELA_{pre}	0.080	0.137	0.698	0.087	0.150
ELA_{post}	0.091	0.137	0.696	0.098	0.148
SAT	162.98	281.65	951.60	0.117	0.202

Note. CLASS = classroom assessment scoring system; PLATO = protocol for language arts teaching observation; ELA = English language arts assessment. For the student-level measures (Tripod [except "S15" variable], ELA [pretest and posttest] and SAT), sum scores were used to calculate intraclass correlations (ICCs) from a three-level random intercept multilevel linear model. For the two teacher-level measures of CLASS and PLATO, the two-level (teachers nested within sections) random intercept model was. Models were fit using Stata mixed command (StataCorp, 2017). Also, — = not modeled (because CLASS and PLATO were measured at the section level, variances and ICC were not reported for CLASS and PLATO).

Appendix B

Description of Multivariate Multilevel (Three-Level; Random Intercept) Model

The student-level (Level 1) model is

$$\begin{bmatrix} E_{jkg.post} \\ S_{jkg.post} \end{bmatrix} = \begin{bmatrix} \gamma_{0kg.E} \\ \gamma_{0kg.S} \end{bmatrix} + \begin{bmatrix} \gamma_{1.E} & 0 \\ 0 & \gamma_{1.S} \end{bmatrix} \begin{bmatrix} \theta_{jkg.TE} \\ \theta_{jkg.TE} \end{bmatrix} + \begin{bmatrix} \gamma_{2.E} & 0 \\ 0 & \gamma_{2.S} \end{bmatrix} \begin{bmatrix} \theta_{jkg.TQ} \\ \theta_{jkg.TQ} \end{bmatrix} + \begin{bmatrix} \gamma_{3.E} & 0 \\ 0 & \gamma_{3.S} \end{bmatrix} \begin{bmatrix} \theta_{jkg.ST} \\ \theta_{jkg.ST} \end{bmatrix} + \begin{bmatrix} \gamma_{4.E} & 0 \\ 0 & \gamma_{4.S} \end{bmatrix} \begin{bmatrix} \theta_{jkg.TC} \\ \theta_{jkg.TC} \end{bmatrix} + \begin{bmatrix} \gamma_{5.E} & 0 \\ 0 & \gamma_{5.S} \end{bmatrix} \begin{bmatrix} E_{jkg.pre} - E_{.kg.pre} \\ E_{jkg.pre} - E_{.kg.pre} \end{bmatrix} + \sum_{l=6}^{13} \begin{bmatrix} \gamma_{l.E} & 0 \\ 0 & \gamma_{l.S} \end{bmatrix} \times \begin{bmatrix} SCH_{jkg.l} \\ SCH_{jkg.l} \end{bmatrix} + \begin{bmatrix} \varepsilon_{jkg.E} \\ \varepsilon_{jkg.S} \end{bmatrix}, \begin{bmatrix} \varepsilon_{jkg.E} \\ \varepsilon_{jkg.S} \end{bmatrix} \sim MN(\mathbf{0}_{(2\times1)}, \Sigma_{1(2\times2)}),$$

where

j is an index for a student ($j = 1, \ldots, J$),

k is an index for a section ($k = 1, \ldots, K$),

g is an index for a school ($g = 1, \ldots, G$),

l is an index for student-level demographic variables ($l = 6, \ldots, 13$),

$E_{jkg.post}$ is posttest ELA scores,

S_{jkg} is posttest SAT scores,

$\theta_{jkg.TE}$ is "teacher explaining" factor,

$\theta_{jkg.TQ}$ is "teacher questioning" factor,

$\theta_{jkg.ST}$ is "encouraging of student talk" factor,

$\theta_{jkg.TC}$ is "teacher big-picture communicating" factor,

$E_{jkg.pre} - E_{.kg.pre.}$ is the within-section deviation score of the student pretest ELA scores from the section mean,

$SCH_{jkg.l}$ is the lth student-level covariates (i.e., Gender, African American, American Indian, Hispanic, Asian, Others, ELL, and Special Ed),

$\gamma_{0jkg..}$ is a *random* intercept for each measure (ELA scores or SAT scores),

$\gamma_{1..}$ is the effect of "teacher explaining" factor,

$\gamma_{2..}$ is the effect of "teacher questioning" factor,

$\gamma_{3..}$ is the effect of "encouraging of student talk" factor,

$\gamma_{4.}$ is the effect of "teacher big-picture communicating" factor,

$\gamma_{5.}$ is the effect of pretest ELA scores on each measure (ELA scores or SAT scores) at the student level,

$\gamma_{l.}$ is the effect of the lth student-level covariates for each measure (ELA scores or SAT scores), and

$\varepsilon_{jkg..}$ is a student-level residual.

The section-level (Level 2) model is

$$\begin{bmatrix} \gamma_{0kg.E} \\ \gamma_{0kg.S} \end{bmatrix} = \begin{bmatrix} \delta_{00g.E} \\ \delta_{00g.S} \end{bmatrix} + \begin{bmatrix} \delta_{1.E} & 0 \\ 0 & \delta_{1.S} \end{bmatrix} \begin{bmatrix} \theta_{kg.RO} \\ \theta_{kg.RO} \end{bmatrix} + \begin{bmatrix} \delta_{2.E} & 0 \\ 0 & \delta_{2.S} \end{bmatrix} \times \begin{bmatrix} \theta_{kg.SO} \\ \theta_{kg.SO} \end{bmatrix} + \begin{bmatrix} \delta_{3.E} & 0 \\ 0 & \delta_{3.S} \end{bmatrix} \begin{bmatrix} E_{.kg.pre} - E_{.g.pre} \\ E_{.kg.pre} - E_{.g.pre} \end{bmatrix} + \begin{bmatrix} \varepsilon_{kg.E} \\ \varepsilon_{kg.S} \end{bmatrix}, \begin{bmatrix} \varepsilon_{kg.E} \\ \varepsilon_{kg.S} \end{bmatrix} \sim MN(\mathbf{0}_{(2\times1)}, \Sigma_{2(\times2)})$$

where

$\theta_{kg.RO}$ is "rater observations of teachers" factor at the section level,

$\theta_{kg.SO}$ is "student observations of teachers" factor at the section level,

$E_{.kg.pre} - E_{.g.pre}$ is the within-school deviation ELA score of the section mean from the school mean,

(*Appendices continue*)

$\delta_{00g..}$ is a *random* intercept for each measure (ELA scores or SAT scores),

$\delta_{1..}$ is the effect of "rater observations of teachers" factor at the section level,

$\delta_{2..}$ is the effect of "student observations of teachers" factor at the section level,

$\delta_{3..}$ is the effect of pretest ELA scores on each measure (ELA scores or SAT scores) at the section level, and

$\varepsilon_{kg..}$ is a section-level residual.

The school-level (Level 3) model is

$$\begin{bmatrix}\delta_{00g.E}\\ \delta_{00g.S}\end{bmatrix} = \begin{bmatrix}\omega_{000.E}\\ \omega_{000.S}\end{bmatrix} + \begin{bmatrix}\omega_{1.E} & 0\\ 0 & \omega_{1.S}\end{bmatrix}\begin{bmatrix}\theta_{g.RO}\\ \theta_{g.RO}\end{bmatrix} + \begin{bmatrix}\omega_{2.E} & 0\\ 0 & \omega_{2.S}\end{bmatrix}\begin{bmatrix}E_{..g.pre}\\ E_{..g.pre}\end{bmatrix} + \begin{bmatrix}\varepsilon_{g.E}\\ \varepsilon_{g.S}\end{bmatrix}, \begin{bmatrix}\varepsilon_{g.E}\\ \varepsilon_{g.S}\end{bmatrix} \sim MN(\mathbf{0}_{(2\times1)}, \Sigma_{3(2\times2)}), \quad (1)$$

where

$\theta_{g.RO}$ is "rater observations of teachers" factor at the school level,

$E_{..g.pre}$ is pretest ELA scores at the school level,

$\omega_{000..}$ is an intercept (i.e., overall mean) for each measure,

$\omega_{1..}$ is the effect of "rater observations of teachers" factor at the school level,

$\omega_{2..}$ is the effect of pretest EFA scores on each measure (ELA scores or SAT scores) at the school level, and

$\varepsilon_{g..}$ is a school-level residual.

A measurement model for teacher talk practice is the three-level *confirmatory* factor model where there are four factors at the student-level (i.e., "teacher explaining" factor; "teacher questioning" factor; "encouraging of student talk"; "teacher big-picture communicating" factor), two factors at the section level (i.e., "rater observations of teachers" factor; "student observations of teachers" factor), and one factor at the school level (i.e., "rater observations of teachers" factor):

$$\boldsymbol{y}_{jkg} = \boldsymbol{\mu} + \boldsymbol{\nu}_{jkg} + \boldsymbol{\nu}_{kg} + \boldsymbol{\nu}_{g}, \quad (2)$$

where $\boldsymbol{y}_{jkg}$ is 23-dimensional (C1-C5, P1, P2, S1-S17 [but S15]) vector of indicators, $\boldsymbol{\mu}$ is a 23-dimensional vector of grand means (or intercept), $\boldsymbol{\nu}_{jkg}$ is a 23-dimensional vector containing latent student scores, $\boldsymbol{\nu}_{kg}$ is a 23-dimensional vector containing latent teacher scores, and $\boldsymbol{\nu}_{g}$ is a 23-dimensional vector containing latent school scores. These latent scores have two sources of variation: common factors and unique factors:

$$\boldsymbol{\nu}_{jkg} = \boldsymbol{\lambda}_S\boldsymbol{\theta}_{jkg} + \boldsymbol{\zeta}_{jkg}, \quad (3)$$

where $\boldsymbol{\lambda}_S$ is a 23 by 4 student-level factor loading matrix, $\boldsymbol{\theta}_{jkg} = [\theta_{jkg.TE}, \theta_{jkg.TQ}, \theta_{jkg.ST}, \theta_{jkg.TC}]'$ is a 4-dimensional vector of student-level latent variable scores, and $\boldsymbol{\zeta}_{jkg}$ is a 23-dimensional vector of student-level unique factors. The section-level measurement model is

$$\boldsymbol{\nu}_{kg} = \boldsymbol{\lambda}_T\boldsymbol{\theta}_{kg} + \boldsymbol{\zeta}_{kg}, \quad (4)$$

where $\boldsymbol{\lambda}_T$ is a 23 by 2 section-level factor loading matrix, $\boldsymbol{\theta}_{kg} = [\theta_{kg.RO}, \theta_{kg.SO}]'$ is a 2-dimensional vector of section-level latent variable scores, and a 23-dimensional vector of section-level unique factors. The school-level measurement model is

$$\boldsymbol{\nu}_{g} = \boldsymbol{\lambda}_Z\theta_{g.RO} + \boldsymbol{\zeta}_{g}, \quad (5)$$

where $\boldsymbol{\lambda}_Z$ is a 23 by 1 school-level factor loading matrix, $\theta_{g.RP}$, is a 1-dimensional vector of school-level latent variable score, and a 23-dimensional vector of school-level unique factors.

Specifically, the (confirmatory) dimensionality structure in $\boldsymbol{\lambda}_S$, $\boldsymbol{\lambda}_T$, and $\boldsymbol{\lambda}_Z$ is specified as follows:

$$\boldsymbol{\lambda}_S = \begin{array}{c} \\ C1\\ C2\\ C3\\ C4\\ C5\\ P1\\ P2\\ S1\\ S2\\ S3\\ S4\\ S5\\ S6\\ S7\\ S8\\ S9\\ S10\\ S11\\ S12\\ S13\\ S14\\ S16\\ S17 \end{array} \begin{bmatrix} & & & & \\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ \lambda_{S.s1} & 0 & 0 & 0\\ \lambda_{S.s2} & 0 & 0 & 0\\ \lambda_{S.s3} & 0 & 0 & 0\\ \lambda_{S.s4} & 0 & 0 & 0\\ \lambda_{S.s5} & 0 & 0 & 0\\ 0 & & 0 & \lambda_{S.s6}\\ 0 & 0 & \lambda_{S.s7} & 0\\ \lambda_{S.s8} & 0 & 0 & 0\\ 0 & \lambda_{S.s9} & 0 & 0\\ 0 & & 0 & \lambda_{S.s10}\\ 0 & \lambda_{S.s11} & 0 & 0\\ 0 & \lambda_{S.s12} & 0 & 0\\ 0 & 0 & \lambda_{S.s13} & 0\\ 0 & & 0 & \lambda_{S.s14}\\ 0 & 0 & \lambda_{S.s16} & \lambda_{S.s16}\\ 0 & 0 & \lambda_{S.s17} & 0 \end{bmatrix},$$

(*Appendices continue*)

$$\boldsymbol{\lambda}_T = \begin{array}{c} C1 \\ C2 \\ C3 \\ C4 \\ C5 \\ P1 \\ P2 \\ S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \\ S7 \\ S8 \\ S9 \\ S10 \\ S11 \\ S12 \\ S13 \\ S14 \\ S16 \\ S17 \end{array} \begin{bmatrix} \lambda_{T.c1} & 0 \\ \lambda_{T.c2} & 0 \\ \lambda_{T.c3} & 0 \\ \lambda_{T.c4} & 0 \\ \lambda_{T.c5} & 0 \\ \lambda_{T.p1} & 0 \\ \lambda_{T.p2} & 0 \\ 0 & \lambda_{T.s1} \\ 0 & \lambda_{T.s2} \\ 0 & \lambda_{T.s3} \\ 0 & \lambda_{T.s4} \\ 0 & \lambda_{T.s5} \\ 0 & \lambda_{T.s6} \\ 0 & \lambda_{T.s7} \\ 0 & \lambda_{T.s8} \\ 0 & \lambda_{T.s9} \\ 0 & \lambda_{T.s10} \\ 0 & \lambda_{T.s11} \\ 0 & \lambda_{T.s12} \\ 0 & \lambda_{T.s13} \\ 0 & \lambda_{T.s14} \\ 0 & \lambda_{T.s16} \\ 0 & \lambda_{T.s17} \end{bmatrix}, \quad \boldsymbol{\lambda}_Z = \begin{array}{c} C1 \\ C2 \\ C3 \\ C4 \\ C5 \\ P1 \\ P2 \\ S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \\ S7 \\ S8 \\ S9 \\ S10 \\ S11 \\ S12 \\ S13 \\ S14 \\ S16 \\ S17 \end{array} \begin{bmatrix} \lambda_{Z.c1} \\ \lambda_{Z.c2} \\ \lambda_{Z.c3} \\ \lambda_{Z.c4} \\ \lambda_{Z.c5} \\ \lambda_{Z.p1} \\ \lambda_{Z.p2} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

and

Parameters in Equations 1 through 5 were estimated simultaneously.

Appendix C

Examples of PLATOPrime Scoring Descriptors

Intellectual Challenge

Focuses on the intellectual rigor of the activities and assignments in which students engage

Classroom Discourse*

Focuses on the opportunities students have for conversations with the teacher and among peers.

Strategy Use and Instruction

Focuses on the teacher's ability to teach strategies connected to learning to read, write, speak, listen, and engage with literature.

Modeling

Focuses on the degree to which a teacher visibly enacts strategies, skills, and processes targeted in the lesson. The teacher might model metacognitive or discussion strategies, a think aloud on how to identify theme, demonstrate how to support a statement with textual evidence.

Time Management

Focuses on the teacher's efficient organization of classroom routines and materials to ensure that instructional time is maximized and little class time is lost to transitions or student behavior.

Behavior Management

Focuses on the degree to which student behavior facilitates academic work.

* indicator Used in the Analysis.