Automated Scoring of Students' Small-Group Discussions to Assess Reading Ability

Audra E. Kosh[1, 2]

Jeffrey A. Greene[1]

P. Karen Murphy[3]

Hal Burdick[2]

Carla M. Firetto[3]

Jeff Elmore[2]

[1]University of North Carolina at Chapel Hill

[2]MetaMetrics, Inc.

[3]The Pennsylvania State University

Published: Summer 2018

Author Note

This research was supported by the Institute of Educational Sciences, U.S. Department of Education, through Grant R305A130031 to Pennsylvania State University. Any opinions, findings, and conclusions or recommendations expressed are those of the author(s) and do not represent the views of the Institute or the U.S. Department of Education.

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Abstract

We explored the feasibility of using automated scoring to assess upper-elementary students' reading ability through analysis of transcripts of students' small-group discussions about texts. Participants included 35 fourth-grade students across two classrooms that engaged in a literacy intervention called Quality Talk. During the course of one school year, data were collected at ten time points for a total of 327 student-text encounters, with a different text discussed at each time point. To explore the possibility of automated scoring, we considered which quantitative discourse variables (e.g., variables to measure language sophistication and latent semantic analysis variables) were the strongest predictors of scores on a multiple-choice and constructed-response reading comprehension test. Convergent validity evidence was collected by comparing automatically-calculated quantitative discourse features to scores on a reading fluency test. After examining a variety of discourse features using multilevel modeling, results showed that measures of word rareness and word diversity were the most promising variables to use in automated scoring of students' discussions.

Keywords: assessment, automated scoring, reading ability

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Automated Scoring of Students' Small-Group Discussions to Assess Reading Ability

The Common Core State Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) in English Language Arts (ELA) call for students to demonstrate a variety of complex skills such as integrating information from multiple texts, identifying reasons and evidence, and engaging in collaborative discussions, in addition to foundational skills such as fluently reading and comprehending text. The broad range of skills represented in these standards raise the need for assessment formats beyond traditional multiple-choice tests (Lane & Iwatani, 2016). Moreover, limitations of existing ELA assessments have led to a variety of stakeholders calling for multiple measures of student ability, particularly interim and formative assessments as well as performance assessments with authentic, real-life tasks to measure complex skills (Gallup, Inc., 2016; Lane & Iwatani, 2016). At the same time, teachers, students, and parents have expressed dissatisfaction with the amount of time spent on testing, with one reason being that more time on testing equates to less time on instruction (Bennett, 2016).

One possible approach to addressing these challenges of assessing ELA skills is to supply teachers, schools, or states with multiple options for test formats, and stakeholders could choose which options best fit the needs of their unique population and purpose. As an added benefit of expanding the range of assessment options, multiple measures of a construct are more likely to provide a comprehensive representation of the construct as opposed to a single assessment format (Correnti, Matsumura, Hamilton, & Wang, 2012). A challenge to producing new assessment formats is the need to develop complex scoring procedures such as creating detailed rubrics and training human raters (Lane & Iwatani, 2016).

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

The purpose of this paper is to explore an innovative format for assessing upper elementary school students' reading ability that can expand the range of reading assessment options by serving as an alternative measure to supplement other traditional reading tests. For the purpose of this study, we conceptualized reading ability as fluently reading and comprehending a text; however, when referring to a measure associated with a particular literacy skill such as fluency or comprehension, we use the respective terms reading fluency or reading comprehension.

The assessment format we sought includes three desirable properties:

1) Minimizing the amount of instructional time spent on assessment by collecting assessment data during already-occurring instructional activities,
2) Utilizing authentic tasks occurring in a classroom's natural context, and
3) Permitting automated scoring in order to reduce the resources required to score tests and reduce the time between test administration and score reporting.

One way to achieve these assessment properties would be to develop a method of measuring reading ability by automatically scoring students' small-group discussions about texts. Using small-group discussions as data for reading assessment meets the first two desirable assessment properties we described but requires research to achieve the third property. For the first property (i.e., minimizing the amount of instructional time spent on assessment), in the case of reading instruction, a common pedagogical strategy is conducting small- or whole-group discussions about texts (Murphy, Wilkinson, Soter, Hennessey, & Alexander, 2009), thereby allowing assessment data to stem from already-occurring instructional activities. For the second property (i.e., utilizing authentic tasks), engaging in collaborative discourse is a typical component of ELA classrooms and represents a skill that students will use throughout their

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

education and career, for example by listening to others' reasoning, building upon ideas, and asking questions (Murphy, Firetto, Wei, Li, & Croninger, 2016). Achieving the third property (i.e., automated scoring) was the focus of this exploratory study: we sought to investigate automatically-calculated features of transcripts of students' small-group discussions as predictors of upper-elementary students' reading ability. Our research question was: which quantitative discourse features, when applied to transcripts of student talk about texts, best predict upper-elementary students' reading ability, thereby allowing for automated scoring of students' discussions? This research question builds from theory relating classroom discussions to student performance in reading.

Theoretical Foundations Relating Classroom Discussion to Reading Comprehension

The use of classroom discussion as a pedagogical tool can be grounded by any one of a number of theoretical frameworks. In practice, discussion approaches focused on improving students' comprehension are often built upon a combination of multiple frameworks, including social constructivist, cognitive, and sociocognitive frameworks (Croninger, Li, Cameron, & Murphy, in press). For example, Vygotsky's social constructivist perspective (1978) described how the social interactions that take place within discussions help students gain a stronger understanding of the text. Alternatively, from a cognitive perspective, text-based discussions provide ample opportunities for individual learners to acquire, refine, and use knowledge while actively engaging in cognitive processes such as organizing, retrieving, and elaborating (Piaget, 1928; Woolfolk Hoy, Davis, & Anderman, 2013). Finally, while the sociocognitive perspective is situated at the intersection of both social constructivist and cognitive frameworks, it uniquely grounds the vital and influential role of the teacher in enhancing students' comprehension through discussion (Bandura, 1977). Indeed, teachers can support students' comprehension by

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

modeling critical-analytic thinking, challenging students with misconceptions, and prompting them to elaborate their thinking aloud within small-group discussions (Murphy & Knight, 2016; Wei & Murphy, in press). In sum, the weaving of such robust and comprehensive foundations provides a strong theoretical foundation for the prediction that small-group classroom discussions can facilitate students' reading comprehension.

Previous Research on Analysis of Multi-Party Interactions

Our research question focuses on interactions within groups of students, which presents a unique challenge in terms of producing an individual-level measure based on analyzing a group-level interaction. One context where previous researchers have analyzed multi-party interactions is discussion threads in online environments. After manual coding of students' interactions in a web-based graduate course, Song and McNary (2011) found that the most common interactions were ones where students clarified ideas or made suggestions but also that the number of discussion board posts only had a weak relationship to students' grades in the course. Moving beyond manual annotations, other scholars have used automated methods to analyze oral meetings. Renals (2011) investigated automatic analysis of meetings with multiple parties, including methods of automatically summarizing meeting content. Germesin and Wilson (2009) worked to automatically identify cases when an individual agreed with a statement made by another person in the group by applying machine learning to various features of meetings, such as the types of words used, the timing and duration of speech, and the structure of combinations of features across an entire meeting.

Previous Research on Alternate Reading Assessments

We reviewed research on various types of reading assessments that utilize oral-administration or automated scoring, as our assessment format includes these two features.

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Orally-administered tests often use oral fluency as a proxy for reading ability due to theory and prior research demonstrating the inherent relationship between expressive language (e.g., speaking and writing) and receptive language (e.g., reading and listening). Isbell, Sobol, Lindauer, and Lowrance (2004) demonstrated that three- and four-year-olds grew in language complexity (e.g., mean length of utterance and word diversity) and storytelling ability from a wordless picture book after participating in a reading program. Moving to slightly older children, Kendeou, van den Broek, White, and Lynch (2009) showed that oral language skills (e.g., listening comprehension, vocabulary) predicted second-graders' reading comprehension skills. Research specific to classroom discourse has also shown that literacy discussions associated with high-level thinking produced student oral responses with more complex language (e.g., greater mean length of utterance; Soter et al., 2008) and stronger oral reading fluency skills (Li et al., 2016).

Orally-administered tests, which follow different formats depending on the students' age, are particularly common for measuring indicators of young children's reading ability such as phonological awareness, listening comprehension, sight word identification, and vocabulary definitions (Nation, Cocksey, Taylor, & Bishop, 2010). Another format of oral assessment, more applicable to lower-elementary students, is orally narrating or retelling a story or orally answering comprehension questions after being read a story (Petersen, Gillam, & Gillam, 2008; Roth, Speece, & Cooper, 2002). Orally-administered tests become less common in upper-elementary grades as reading instruction places less emphasis on word recognition and decoding and more emphasis on text comprehension. Further, by upper-elementary school, students typically have acquired the basic reading and writing skills needed to take a written reading comprehension test. In research on upper-elementary school and beyond, researchers using oral

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

assessments tend to focus on the relationship between oral reading fluency and scores on written reading comprehension tests rather than using an oral assessment format to measure reading comprehension (Denton et al., 2011; Hunley, Davies, & Miller, 2013).

Previous Research on Automated Scoring in Reading and Writing

Researchers have introduced several new, innovative assessment methods using automated scoring whereby students respond to questions during reading (Magliano, Millis, Levinstein, & Boonthum, 2011; Millis, Magliano, & Todaro, 2006; Millis, Magliano, Wiemer-Hastings, Todaro, & McNamara, 2011). Automated scoring often relies on latent semantic analysis (LSA), a technique that measures the semantic similarity of words based on which words frequently occur together in large corpora of texts (Landauer, McNamara, Dennis, & Kintsch, 2007). Magliano and Millis (2003) found that LSA variables used to measure semantic similarity between verbal text comprehension prompts and undergraduate students' responses to those prompts predicted scores on a comprehension test. Other tasks for which automated scoring has been attempted include open-ended short-answer questions (Brew & Leacock, 2013), computer-generated cloze tasks whereby a student must identify the missing word from a sentence in a passage (Stenner, Fisher, Stone, & Burdick, 2013), and students' summaries of texts (Dascalu et al., 2015).

Automated scoring methods, particularly those for assessing writing ability, are commonly developed and validated by training a model on human ratings of essays, often using the aforementioned LSA measures or additional language diversity measures as predictor variables. In Burdick and colleagues' (2013) development of an automated scoring method of students' written compositions, the authors found that vocabulary diversity was a strong predictor of human ratings on compositions. Educational Testing Service's e-Rater essay scoring

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

tool also incorporates lexical complexity measures in addition to other variables such as the use of prompt-specific vocabulary (Attali & Burstein, 2006). Other automated scoring methods use LSA to compare the semantic similarity of words (e.g., Pearson's Intelligent Essay Assessor) in order to score writing on the meaning of words rather than just word diversity or grammar (Landauer, Laham, & Foltz, 2003).

These aforementioned findings show that students demonstrated greater reading and writing ability when using more sophisticated language, and for writing, when using words with similar meaning to the topic of the prompt. However, previous automated scoring studies have not addressed two components: scoring of individual students participating in group activities or scoring of oral language to measure reading ability. Our study sought to fill these two gaps in automated scoring methodology. Building on previously-described literature, we hypothesized that students would score higher on reading ability tests when they spoke with more complex language. We also explored whether or not students would score higher on reading ability tests when their discussions stayed on-topic with the ideas in the text that was being discussed, recognizing that although on-topic responses while reading a text have predicted scores on text comprehension measures (Magliano & Millis, 2003; Millis, Magliano, & Todaro, 2006), similar on-topic measures may not generalize to small-group student discussions occurring after reading a text. These hypotheses guided the search for automated scoring metrics applicable to small-group, classroom-based discussions of texts.

Method

Participants

Participants included two fourth-grade classes with 17 and 18 students each, for a total of 35 students. Separate teachers taught each class in a single, semi-rural, mid-Atlantic private

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

school. Although demographic data were not collected, video data of students suggested that gender was roughly equal and that most students were Caucasian. Approximately 30 percent of students at the school qualified for free/reduced lunch. Students participated in the study over the course of an entire school year, with data collected at ten roughly equally-spaced time points between November, 2013 and May, 2014. After accounting for typical student absences and one incident of video malfunction during one discussion group at one time point, the final data set contained 327 student-text encounters.

Procedures

Data were collected while students participated in Quality Talk, a literacy intervention aimed at increasing students' critical-analytic thinking and high-level comprehension skills (Murphy, Greene, & Firetto, 2014). In Quality Talk, students read a text and then engaged in an open-ended group discussion about the text almost every week, with the goal of using discussion to increase comprehension. A key difference between Quality Talk discussions and other forms of narrative expression used as assessment is that students did not retell stories in Quality Talk; rather, students made connections between their own knowledge or personal lives and the texts, asked various forms of meaningful questions, constructed thoughtful arguments, and made intertextual comparisons. Quality Talk encouraged students to talk about, around, and with the text, meaning that students were encouraged to expand their discussions to include ideas both directly and indirectly related to the text. Throughout the year-long intervention, teachers delivered explicit instruction, using provided materials, to students about asking questions and providing responses (Murphy et al., 2016). Additionally, teachers received initial and ongoing professional development about the discussion model (e.g., components of Quality Talk or how

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

to fade their own participation in discussions in order to release interpretive authority to students) and how to analyze their own discussions as a form of feedback for discourse improvement.

Discussion groups contained between four and six students along with the students' classroom teacher. Students were assigned to discussion groups in ways that sought to ensure roughly equivalent groups in terms of reading ability and gender. Table 1 presents mean AIMSweb R-CBM scores (i.e., the number of words students correctly read aloud in one minute) at the beginning of the school year by group and by class membership (Pearson, 2012). A one-way ANOVA with AIMSweb R-CBM scores as the dependent variable and discussion group as the grouping variable, with one student removed from analysis who was not assigned a reading group at the time of AIMSweb R-CBM administration, indicated no statistically significant differences in reading fluency across groups, $F(5,28) = 1.09$, $p = .39$. Although this result should be interpreted cautiously due to few students in each group, the means presented in Table 1 demonstrate that discussion groups had comparable reading fluency at the beginning of the school year with some variation due to outliers in AIMSweb R-CBM scores and due to the need to balance gender within groups. For the most part, group membership remained constant throughout the school year, with a few changes made to accommodate student personality conflicts. Each discussion was roughly held to a twenty-minute time limit, thereby limiting the length of dialogue the group as a whole could produce at any time point.

Four main activities occurred at each data collection time point. First, students independently read a text from the fourth-grade Reading Street™ Common Core curriculum. Second, students completed a journal activity where they brainstormed and wrote down questions about the text designed to provoke thoughtful discussion among peers. Third, students participated in a small-group Quality Talk discussion with their teacher about the text, and,

fourth, students completed a posttest targeted to the text from the group discussion. Furthermore, as part of the Quality Talk intervention, students received bi-weekly Quality Talk mini-lessons about different types of discussion questions and argumentation structure.

Materials and Measures

Materials and measures included a different text at each time point along with the corresponding transcript of group discussion about the text, researcher-created posttests to assess text comprehension, and the AIMSweb R-CBM oral reading fluency measure (Pearson, 2012). Although the Quality Talk intervention was implemented across fourteen data collection time points, several changes were made to instruments during the first four time points as the researchers pilot tested various instrument formats to determine the most appropriate measures. As a result, this study excludes the first four time points and instead only considers the last ten time points which had consistent measures.

Texts. Students read and discussed a different text from the fourth-grade Reading Street™ Common Core curriculum almost every week, with comprehension data collected roughly every two weeks. Table 2 includes the titles, genre, text complexity as measured with The Lexile® Framework for Reading (Stenner, Burdick, Sanford, & Burdick, 2007), and total number of words of each text at the ten data collection time points. Lexile® measures of commonly used texts in first- through twelfth-grade typically range from below 200 Lexiles to above 1600 Lexiles, where one Lexile is a unit equal to 1/1000th of the difference in text comprehensibility between an early-reader basal primer and an encyclopedia (Stenner, Burdick, Sanford, & Burdick, 2007). The mean Lexile® measure of texts used in the study was 837 Lexiles, which corresponds to the text complexity of typical fourth-grade texts (MetaMetrics, Inc., 2016).

Transcripts of group discussions. Video recordings of group discussions were professionally transcribed with anonymized identifiers for individual students. Professional transcribers watched video recordings while transcribing in order to guarantee accurate identification of the speaker in the transcript. Transcripts were cleaned to ensure student confidentiality and to increase consistency for text analysis measures. Cleaning criteria resulted in deletion of the following types of occurrences: personally-identifying information such as references to family members, utterances where an entire talk-turn was inaudible due to interruption or quiet speaking, and utterances that consisted of only one stutter or nonsensical word (e.g., hm, um, uh, etc.). Cleaned transcripts were then parsed into individual text samples consisting of a single student's dialogue at a single time point.

Posttests. Multiple-choice and short-answer posttests served as measures of reading comprehension for each unique Reading Street™ text. After small-group discussions, students completed a comprehension posttest comprised of two dichotomously-scored locate/recall four-option multiple-choice questions (e.g., "What did the husband do to show his love for Iemanjá's daughter?") and three integrate/interpret short-answer questions (e.g., "Why was Iemanjá angry with the servants for begging?"), all modeled after the cognitive targets identified in the reading framework for the National Assessment of Educational Progress (National Assessment Governing Board, 2013). Short-answer questions were scored by two raters on the number of idea units presented in each question (Wolfe & Goldman, 2005), with scores capped at two idea units per question for a total of eight possible points on posttests. Raters resolved any scoring discrepancies together until achieving 100-percent agreement. Table 3 presents descriptive statistics for posttest performance at each time point. Internal consistency of posttests was not calculated due to the limited number of participants responding to each posttest.

AIMSweb R-CBM. In addition to reading comprehension posttests**,** we used the AIMSweb R-CBM (Pearson, 2012) as a measure of oral reading fluency administered at three time points during the school year: September 16-17, January 13, and March 31. The AIMSweb R-CBM counts how many words a student correctly reads aloud from a passage in one-minute. Table 4 presents descriptive statistics for student performance at each of the three time points by class and group.

Development of Automatically-Calculated Quantitative Discourse Features

We began by selecting over a dozen variables to apply to transcripts of students' discussions and then determining which of those variables most strongly predicted students' reading comprehension posttest scores. We considered numerous text complexity variables, described below, in order to operationalize language sophistication for the hypothesis that students would demonstrate greater reading ability when they spoke with more sophisticated language. For the exploration of whether students who stayed on-topic with the text showed greater reading ability, we used LSA to compare the meaning of words from student talk to words in the text to produce a measure of how closely students' dialogue mirrored the meanings of words in the text. As a first step in the work to predict reading ability with quantitative student discussion metrics, we considered variables that did not require accounting for dependencies among members of the group because of the need for an assessment score to produce an inference about student ability that is as independent as possible of the effects of discussion-group peers.

Selection of variables for analysis. In alignment with the hypothesis that students would demonstrate greater reading ability when they spoke with more sophisticated language, quantitative discourse measures were generated for each student's dialogue at each time point.

The discourse measures included nine text complexity features found by Fitzgerald and colleagues (2015) to be the most important in determining text complexity of elementary texts. The variables included:

1) decoding demand (i.e, the degree of orthographic complexity corresponding to the difficulty of decoding the words)
2) number of syllables in words
3) average age of acquiring a word's meaning
4) abstractness (i.e., extent to which words refer to concepts that cannot be seen)
5) word rareness (i.e., the inverse of the number of times words appear in a large corpus of existing texts commonly used in classrooms; for example, a statement such as "apples taste good" uses words that more commonly occur in English texts as compared to words in a statement such as "persimmons trigger scrumptiousness")
6) intersentential complexity (i.e., sentence complexity including word, phrase, and letter repetition across adjacent sentences)
7) phrase diversity (i.e., word, phrase, and letter repetition across multiple sentences)
8) text density (i.e, amount of information within a text)
9) noncompressability (i.e., degree to which text can be compressed, where redundant and less complex text corresponds to text that is more compressible).

For more detail on how variables were calculated, see Fitzgerald et al. (2015).

In addition to text complexity measures, we also looked at quantitative measures of expression previously used in expressive language research. Namely, a variety of type-token ratios (TTRs) were calculated, which compare the number of unique words (i.e., types) to the total number of words (i.e., tokens). In this study, counts of types and tokens excluded words

from a stop list consisting of the most frequently used words with low semantic value (e.g., and, he, she, the, but, in). Spoken words that failed to appear on EDL Core Vocabulary lists (Taylor et al., 1989) or in an existing corpus of 540 million words appearing in K-12 text were excluded from type and token counts. These included proper nouns (e.g., character names, city names), slang and grammatically incorrect words (e.g., yeah, funner), and stutters (e.g., um, hm). TTRs were calculated without regard to individual utterances; that is, the TTR for a participant is the total number of unique words spoken during a discussion divided by the total word count for the student's spoken dialogue during the discussion. For example, if a student said the same word in different talk turns throughout a discussion, that word was counted exactly one time as a unique word.

Because the raw TTR (i.e., types divided by tokens) has received much criticism due to its tendency to decrease as the number of words increases (Hess, Sefton, & Landry, 1986; Malvern, Richards, Chipere, & Durán, 2004; Richards, 1987), we examined transformed TTRs such as the log TTR (i.e., log of types divided by log of tokens; Herdan, 1960), root TTR (i.e., types divided by the square root of tokens; Herdan, 1960), and the moving-average TTR, which calculates a TTR on repeated samples of a specified window size across the entire text and then averages those TTRs (Covington & McFall, 2010). The moving-average TTR was calculated for window sizes of 30, 50, and 100 words.

In addition to lexical diversity measures, we also calculated mean length of utterance (MLU), which measures the average number of words per talk turn, due to prior research showing that longer talk turns are associated with discussions promoting reading comprehension (Soter et al., 2008). MLU also allows a standardized calculation of how much students say that is not necessarily correlated with the total number of words in the talk sample. For example, a

student could say 50 words during a single utterance or 50 words across 10 separate utterances resulting in different MLUs of 50 and 5, respectively.

It is important to note that transcript segmentation affects MLU, which is particularly challenging in the case of group discourse when students frequently interrupt each other. Although researchers have proposed several ways for segmenting transcripts (Foster, Tonkyn, & Wigglesworth, 2000; Nutter, 1981; Passonneau & Litman, 1997), we did not modify transcript segmentation beyond the segmentation determined by the professional transcriber in order to preserve the automated nature of the assessment format by minimizing subjective data cleaning requirements. Accordingly, an utterance consisted of a student's words until either the student finished speaking or was interrupted by another group member, as determined by line breaks in the transcript. Although transcripts for this study did receive some data cleaning as previously described, discourse features were calculated in an entirely automated fashion once applied to a clean transcript.

To explore whether students demonstrated greater reading ability when their discussions followed the topics in the text, LSA cosines between the words students spoke and the words in the story were generated for each students' dialogue at each time point. Because LSA results can vary based on the semantic space, we conducted LSA in two separate semantic spaces: the English 100k and Touchstone Applied Science Associates (TASA) semantic spaces obtained from the Semantic Space Online Repository (Günther, 2015). The LSAfun package in R was used to calculate LSA cosines for both semantic spaces (Günther, Dudschig, & Kaup, 2014).

Multilevel modeling of posttest scores. We tested all of the aforementioned text complexity variables and LSA cosines for the extent to which they predicted reading comprehension posttest scores. Two-level multilevel modeling was used to account for the

hierarchical structure of the data, with 35 students as level-2 units and 327 student-text encounters as level-1 units across ten data collection time points. A third-level for the student's teacher was not possible because there were only two classrooms included in the study. We also included control variables for features of the text that students read, including text difficulty as measured with The Lexile® Framework for Reading (Stenner, Burdick, Sanford, & Burdick, 2007), the total number of words in the text, and the text genre (i.e., fiction or non-fiction). Multilevel modeling was conducted in version 7 of HLM for Windows.

Results

Multilevel Modeling Results

Table 5 presents results for multilevel models. The null model (Model 1) showed that reading comprehension posttest scores had an intraclass correlation coefficient of .185, meaning that 18.5 percent of the variation in posttest scores was due to students and thus multilevel modeling was an appropriate data analysis technique. Next, we built a model solely with a variable for data collection time point, *Time* (Model 2). *Time* was coded from zero to nine, respectively corresponding to the ten data collection points. *Time* was statistically significant ($p < .001$) such that students, on average, gained 0.22 points on posttest scores per data collection time point for a total of a 2.2 point gain over the ten time points. Due to its statistical significance, *Time* was included in all models when testing discourse variables to account for student growth above and beyond the effect of time. Following a model building approach, we built a model for each discourse variable coupled with *Time*. Discourse variables were entered into the model group-centered in order to aid interpretation. Based on analyzing variance components, the intercept was treated as a random effect, but *Time* and discourse variables were fixed.

After examining all text complexity and LSA variables, two models had discourse variables that statistically significantly predicted reading comprehension posttest scores after accounting for variation due to students and growth over time. These models were those with the variables *Word Rareness* (Model 3) and *Root TTR* (Model 4), respectively explaining 31.2 percent and 30.7 percent of level-1 variance combined with the effect of *Time*. When both *Word Rareness* and *Root TTR* were included in the same model with *Time*, neither variable was a statistically significant predictor of posttest scores (Model 5), likely because *Word Rareness* and *Root TTR* were highly correlated (r=.908). Thus, either Model 3 or Model 4 was the most parsimonious model. Table 6 shows descriptive statistics for *Word Rareness* and *Root TTR*, and Table 7 shows the correlation matrix of variables used in the multilevel models. For descriptive statistics of all text variables by time point, see the appendix. LSA cosines, mean length of utterance, and the other aforementioned text complexity variables were not statistically significant predictors of reading comprehension posttest scores above and beyond *Time, Word Rareness*, and *Root TTR*. Thus, we found that students that spoke with more rare words and diverse vocabulary scored higher on reading comprehension posttests but that the extent to which students stayed on-topic during discussion with the topics in the text was not a statistically significant predictor of reading comprehension posttest scores.

Text complexity variables (i.e., Lexile® measure of the story, total number of words in the story, and story genre) were not statistically significant predictors of reading comprehension posttest scores above and beyond the effects of quantitative discourse features and time point. Additionally, the statistical significance of discourse variables in all models did not change after controlling for the previously-stated text complexity variables of the stories. All results are available by contacting the first author.

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Validation with AIMSweb R-CBM Scores

Given the similarity in statistical significance of *Word Rareness* and *Root TTR*, we compared each variable with AIMSweb R-CBM scores as an independent measure of reading fluency in order to produce convergent validity evidence for each model and thus determine whether either *Word Rareness* or *Root TTR* was more appropriate as an indicator of reading ability. AIMSweb R-CBM assessments were only administered at three time points (i.e., once in fall, winter, and spring) as opposed to the ten discussion time points; therefore, AIMSweb R-CBM scores were not used as a dependent variable in multilevel models with discourse measures as predictors. Instead, *Word Rareness* and *Root TTR* measures were correlated to scores from the chronologically nearest AIMSweb R-CBM administration (Table 8).

Time points one, two through five, and six through ten respectively were correlated with fall, winter, and spring AIMSweb R-CBM scores. During exactly half of the time points, AIMSweb R-CBM scores more strongly correlated with *Root TTR* than they correlated with posttests. Similarly, during four out of ten time points, AIMSweb R-CBM scores had a greater correlation with *Word Rareness* than they did with posttests. The differences between the correlation of AIMSweb R-CBM scores with posttest scores and the correlation of AIMSweb R-CBM scores with either *Word Rareness* or *Root TTR* at all time points were not statistically significant, as tested using Steiger's Z-test (Lee & Preacher, 2013). These results suggest that each of the explored models presented here, which utilized *Word Rareness* and *Root TTR*, perform roughly as well as traditional reading comprehension posttests when relating to AIMSweb R-CBM scores, which is a widely-accepted measure of reading ability and has ample reliability and validity evidence (Pearson, 2012).

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Summary of Results

In conclusion, after considering how various quantitative discourse variables predicted posttest scores and correlated to AIMSweb R-CBM scores, we found *Word Rareness* and *Root TTR* were the most appropriate variables to use in automated scoring of discussions to measure reading ability. Both variables produced similar results, explaining about 31 percent of variance in posttest scores in separate models with *Time*. The extent to which students stayed on-topic with the ideas in the text during discussion, as operationalized with LSA cosines produced in two different semantic spaces, was not a statistically significant predictor of reading comprehension posttest scores.

Discussion

We explored the feasibility of automatically scoring transcripts of fourth-grade students' small-group discussions as a new method of assessing reading ability. Our results showed that automatically-calculated quantitative discourse variables, as applied to students' talk about texts, were statistically significant predictors of reading ability measures. Specifically, the average rareness of words occurring in a large corpus of texts (i.e., *Word Rareness*) and the number of unique words divided by the square root of the total number of words (i.e., *Root TTR*) were statistically significant predictors of reading ability after controlling for variation due to students, growth over time, and text features such as text complexity, text length, and genre.

These results support a proof of concept for automated scoring of students' classroom reading discussions to measure reading ability. To be clear, we do not recommend that model coefficients in this study be directly applied to score future transcripts of student talk; rather, we presented analyses for the purpose of identifying which quantitative discourse variables were

most closely related to reading ability, so that these variables can be explored in future automated scoring work.

Limitations

An initial limitation, common to most studies of this type, concerns the degree to which findings from this study can be extrapolated to other contexts. Our methods were applied in the unique context of one fourth-grade discourse-based reading intervention, Quality Talk, and thus we cannot generalize results to other types of classroom discussions or grade levels. More research is needed to examine the viability of this method across contexts.

One challenge with identifying quantitative discourse features that relate to reading ability was that *Word Rareness* and *Root TTR* were positively correlated with total number of words. That is, values of *Word Rareness* and *Root TTR* indicated more sophisticated talk when students talked more. Although this finding was theoretically intuitive (i.e., students that talk more about a text show greater understanding of the text), the correlation is problematic when scoring discussions because students have a fixed, pre-determined amount of time to talk with their group and the amount of talk in a group is zero-sum (i.e., if one student talks more, then others must talk less). Thus, we did not see an increase in *Word Rareness* or *Root TTR* over time despite the fact that posttest scores did increase over time. This issue limits correlations between quantitative discourse features and posttest scores because discourse features cannot continue to steadily rise along with posttest scores when they are correlated with a student's amount of talk. However, other variables we considered that were nearly independent of number of words (e.g., moving-average type-token ratio) were not statistically significant predictors of posttest scores, thereby indicating that amount of talk confounds with posttest scores. Relatedly, students demonstrated a large range of the number of words and utterances spoken across time points (see

Table A11), raising concern about the minimum number of spoken words necessary to report a reliable score. Further research should investigate how scores are affected when a student does not have ample opportunity to participate in discussion due to other dominating members of a discussion group.

We also note, as a limitation, that we did not report reliability of posttest scores because several authors have recommended between 200 and 400 participants as the minimum sample size required to calculate metrics of internal consistency (Charter, 1999; Kline, 1986), whereas our study had 35 participants completing each posttest, and the study design did not permit an alternate measure of reliability such a parallel forms or test-retest reliability. Finally, as limitation related to the methods of this study, all correlations between AIMSweb R-CBM scores and discourse features were relatively low. Future work should explore additional convergent validity criteria due to possibility that reading fluency is a sufficiently different construct than reading comprehension as measured through spoken language sophistication.

Three additional limitations should be noted as pertaining to the general idea of using discussion as a measure of reading ability. First, the discussion-based assessment format we proposed requires transcribed data, and the cost of obtaining transcribed data may outweigh the costs saved from avoiding traditional test development stages (e.g., item writing, form design, human scoring). In the future, automatically calculating assessment scores from student talk has more potential for widespread applicability as automatic speech recognition technologies improve, including voice-to-text tools and speaker diarization (i.e., identifying which speaker spoke when), thus decreasing the resources needed to obtain accurate transcriptions. We recognize that automatic speech recognition technologies in their current state may not suffice

for transcribing group discussions, yet, nevertheless, we presented this exploratory study as an example of possible future assessment methods.

Second, it is possible that discussion itself increases reading comprehension (Murphy et al., 2016), thereby indicating that the discussion-based assessment format could impact the exact construct it attempts to measure. However, to some extent, this concern exists for any type of test, because research on the testing effect has shown that learning improves when a portion of learning time is spent retrieving information (Carrier & Pashler, 1992; Roediger & Butler, 2011). Future research could experimentally test this possibility by administering parallel comprehension tests both before and after discussions.

Third, differences in the instructional environment due to the teacher, students, or dynamics between students and their teacher could result in different types of group discussions which could influence automatically-calculated scores in a way that does not support the validity of score inferences. For example, construct-irrelevant individual student differences (e.g., personality, gender, English language fluency, relationships with classroom peers) may influence how willing a student is to speak in a small-group, which would in turn affect automatically-calculated scores due to the issue discussed previously of how the scores are correlated with total number of words spoken. Additionally, teachers' preferences and personalities in the classroom may cause some teachers to intervene more or less in discussion or ask different types of questions, which would result in non-standard administration of the assessment. For these reasons, we recommend using this particular assessment method as a formative assessment that informs instruction at the classroom level rather than a summative assessment used to compare students across classrooms. As an eventual goal, a user interface for teachers could allow a teacher to upload the text the class is discussing, and then the interface would perform the

speech-to-text translation and the calculations necessary to produce a student score in real-time. Teachers could use formative assessment scores to identify students who need extra learning supports, such as through targeted vocabulary enrichment or activities to further develop reading comprehension.

Conclusion

We are in the early stages of measuring reading ability through automated analysis of students' classroom discussions. Our results showed that students who spoke with more rare words and more diverse vocabulary scored higher on reading comprehension tests as compared to students that used more common words or repeated the same words. These initial results are promising and warrant further exploratory work on automated scoring of students' discussions. Because work of this nature is relatively new, many of the variables we considered stemmed from studies of text difficulty when students read a text (e.g., difficulty to decode a word) as opposed to the complexity of students' verbalizations when orally speaking about a text, which could explain why several of the variables we considered were statistically insignificant in models predicting reading ability based on student talk.

Before using automatically-calculated scores of discussions as a proxy for other reading ability tests, in the future researchers should (a) determine how quantitative discourse features compare to other validated reading tests beyond AIMSweb R-CBM scores, (b) broaden the sample to include more participants across multiple grade levels, (c) analyze quantitative features in the context of literacy discussions beyond Quality Talk, and (d) develop a method for handling dependency of talk within a group, including exploring discourse features applicable to multi-party interactions (Renals, 2011; Song & McNary, 2011).

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Bandura, A. (1977). *Social learning theory.* Englewood Cliffs, NJ: Prentice Hall.

Bennett, R. E. (2016). Opt out: An examination of issues (Report No. RR-16-13). Princeton, NJ: Educational Testing Services.

Brew, C., & Leacock, C. (2013). Automated short answer scoring: Principles and prospects. In M. D. Shermis & J. Burstein, (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 136-152). New York, NY: Routledge.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633-642.

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, 21(4), 559-566.

Correnti, R., Matsumura, L. C., Hamilton, L. S., & Wang, E. (2012). Combining multiple measures of students' opportunities to develop analytic, text-based writing skills. *Educational Assessment, 17(2)*, 132-161.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*,*17*(2), 94-100.

Croninger, R. M. V., Li, M., Cameron, C, & Murphy, P. K. (in press). Classroom discussions: Building the foundation for productive talk. In P. Karen Murphy (Vol. Ed.) & P. A. Alexander (Series Ed.), *Educational Psychology Insights Series, Classroom discussions*

in education: Promoting productive talk about text and content. New York, NY: Routledge.

Dascalu, M., Stavarache, L. L., Dessus, P., Trausan-Matu, S., McNamara, D. S., & Bianco, M. (2015). Predicting comprehension from students' summaries. In C. Contati, N. Heffernam, A. Mitrovic, & M.F. Verdejo (Eds.), *Artificial Intelligence in Education* (pp. 95-104). Switzerland: Springer International.

Denton, C.A., Barth, A. E., Fletcher, J. M., Wexler, J., Vaughn, S., Cirino, P.T.,…Francis, D. J. (2011). The relations among oral and silent reading fluency and comprehension in middle school: Implications for identification and instruction of students with reading difficulties. *Scientific Studies of Reading*, *15*(2), 109-135.

Fitzgerald, J., Elmore, J., Koons, H., Hiebert, E. H., Bowen, K., Sanford-Moore, E. E., & Stenner, A. J. (2015). Important text characteristics for early-grades text complexity. *Journal of Educational Psychology*, *107*(1), 4-29.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354-375.

Gallup, Inc. (2016). Make assessments work for all students: Multiple measures matter. Northwest Evaluation Association. Retrieved from https://www.nwea.org/resources/make-assessment-work-students-multiple-measures-matter/

Germesin, S., & Wilson, T. (2009). Agreement detection in multiparty conversation. In *Proceedings of the 2009 International Conference on Multimodal Interfaces* (pp. 7-13). Cambridge, MA: Association for Computing Machinery.

Günther, F. (2015). Semantic space online repository. Retrieved July 9, 2015 from http://www.lingexp.uni-tuebingen.de/z2/LSAspaces/

Günther, F., Dudschig, C., & Kaup, B. (2014). LSAfun - An R package for computations based on latent semantic analysis. *Behavior Research Methods, 47,* 930-934.

Herdan, G. (1960). *Type-token mathematics: A textbook of mathematical linguistics*. The Hague, Netherlands: Mouton.

Hess, C. W., Sefton, K. M., & Landry, R. G. (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech, Language, and Hearing Research*, *29*(1), 129-134.

Hunley, S. A., Davies, S. C., & Miller, C. R. (2013). The relationship between curriculum-based measures in oral reading fluency and high-stakes tests for seventh-grade students. *Research in Middle Level Education*, *36*(5), 1-8.

Isbell, R., Sobol, J., Lindauer, L., & Lowrance, A. (2004). The effects of storytelling and story reading on the oral language complexity and story comprehension of young children. *Early Childhood Education Journal*, *32*(3), 157-163.

Kendeou, P., Van den Broek, P., White, M. J., & Lynch, J. S. (2009). Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills. *Journal of Educational Psychology, 101*(4), 765-778.

Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design*. New York: Methuen.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the intelligent essay assessor. In M. D. Shermis & J. C. Burstein (Eds.), *Automated*

essay scoring: A cross-disciplinary perspective (pp. 82-106). Mahwah, NJ: Lawrence Erlbaum Associates.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2013). *Handbook of latent semantic analysis*. New York, NY: Routledge.

Lane, S. & Iwatani, E. (2016). Design of performance assessments in education. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 274-293). New York, NY: Routledge.

Lee, I. A., & Preacher, K. J. (2013, September). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]. Available from http://quantpsy.org.

Li, M., Murphy, P.K., Wang, J., Mason, L.H., Firetto, C.M., Wei, L., Chung, K.S. (2016), Promoting reading comprehension and critical-analytic thinking: A comparison of three approaches with fourth and fifth Graders. *Contemporary Educational Psychology.46,* 101-115.

Magliano, J. P., Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure and latent semantic analysis. *Cognition and Instruction*, *21*(3), 251-283.

Magliano, J. P., Millis, K. K., Levinstein, I., & Boonthum, C. (2011). Assessing comprehension during reading with the reading strategy assessment tool (RSAT). *Metacognition and Learning*, *6*(2), 131-154.

Malvern, D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke, UK: Palgrave Macmillan.

MetaMetrics, Inc. (2016). Lexile measures and grade levels. Retrieved from https://lexile.com/about-lexile/grade-equivalent/

Millis, K., Magliano, J., & Todaro, S. (2006). Measuring discourse-level processes with verbal protocols and latent semantic analysis. *Scientific Studies of Reading*, *10*(3), 225-240.

Millis, K., Magliano, J., Wiemer-Hastings, K., Todaro, S., & McNamara, D. S. (2011). Assessing and improving comprehension with latent semantic analysis. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 207-225). New York, NY: Routledge.

Murphy, P. K., Firetto, C. M., Wei, L., Li, M., & Croninger, R. M. V. (2016). What REALLY works: Optimizing classroom discussions to promote comprehension and critical-analytic thinking. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 27-35.

Murphy, P. K., Greene, J. A., & Firetto, C. M. (2014). *Quality Talk: Developing students' discourse to promote critical-analytic thinking, epistemic cognition, and high-level comprehension*. (Technical Report No. 1). The Pennsylvania State University.

Murphy, P. K., Knight, S. L. (2016). Exploring a century of advancements in the science of learning. *Review of Research in Education*, *40*(1), 402–456. 10.3102/0091732X16677020

Murphy, P. K., Wilkinson, I. A., 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*(3), 740-764.

National Assessment Governing Board. (2013). *Reading Framework for the 2013 National Assessment of Educational Progress.* Washington, DC: U.S. Department of Education. Retrieved from: www.edpubs.gov/document/ed005373p.pdf

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Washington, DC: Authors.

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Nutter, N. (1981). Relative merit of mean length of t-unit and sentence weight as indices of syntactic complexity in oral language. *English Education*, *13*(1), 17-19.

Passonneau, R. J., & Litman, D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, *23*(1), 103-109.

Pearson, Inc. (2012). Aimsweb technical manual. Bloomington, MN: Pearson.

Petersen, D. B., Gillam, S. L., & Gillam, R. B. (2008). Emerging procedures in narrative assessment: The index of narrative complexity. *Topics in Language Disorders*, *28*(2), 115-130.

Piaget, J. (1928). *The child's conception of the world*. London: Routledge and Kegan Paul.

Renals, S. (2011). Automatic analysis of multiparty meetings. *Sadhana*, 36(5), 917-932.

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of Child Language*, *14*(2), 201-209.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20-27.

Roth, F. P., Speece, D. L., Cooper, D. H., & De La Paz, S. (1996). Unresolved mysteries: How do metalinguistic and narrative skills connect with early reading? *The Journal of Special Education*, *30*(3), 257-277.

Song, L., & McNary, S. W. (2011). Understanding students' online interaction: Analysis of discussion board postings. *Journal of Interactive Online Learning*, *10*(1), 1-14.

Soter, A. O., Wilkinson, I. A. G., Murphy, P. K., Rudge, L., Reninger, K., & Edwards, M. (2008). What the discourse tells us: Talk and indicators of high-level comprehension. *International Journal Educational Research, 47*, 372-391.

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2007). The Lexile Framework for Reading Technical Report. Durham, NC: MetaMetrics.

Stenner, A. J., Fisher, W. P., Stone, M. H., & Burdick, D. S. (2013). Causal rasch models. *Frontiers in Psychology, 4*(536): 1-14.

Taylor, S. E., Frackenpohl, H., White, C. E., Nieroroda, B. W., Browning, C. L., & Brisner, E. P. (1989). *EDL core vocabularies in reading, mathematics, science, and social studies*. Orlando, FL: Steck-Vaughn Company.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wei, L., & Murphy, P. K. (in press). Teacher and student roles: Walking the gradually changing line of responsibility. In P. Karen Murphy (Vol. Ed.) & P. A. Alexander (Series Ed.), *Educational Psychology Insights Series, Classroom discussions in education: Promoting productive talk about text and content.* New York, NY: Routledge.

Wolfe, M. B. W., & Goldman, S. R. (2005). Relations between adolescents' text processing and reasoning. *Cognition and Instruction, 23*(4), 467-502.

Woolfolk Hoy, A., Davis, H., & Anderman, E. (2013). Theories of learning and teaching in TIP. *Theory Into Practice, 52*, 9–21.

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Tables

Table 1

Mean Fall AIMS R-CBM Scores by Class and Group

	Group 1 (6 students)	Group 2 (5 students)	Group 3 (6 students)
Class X	116.17	145.40	135.67
Class Y [a]	141.33	157.20	134.67

[a] One student from Class Y was not assigned a group during the first discussion week and was omitted from scores reported in the table.

Table 2

Texts Used in Discussion

Time Point	Date of Discussion	Title	Genre	Lexile® Measure	Number of Words
1	November 20	*The Man Who Named the Clouds*	Non-fiction	920L	1,723
2	December 12	*How Night Came from the Sea*	Fiction	950L	1,281
3	January 8	*Paul Bunyan*	Fiction	1000L	2,550
4	January 23	*Encantado: Pink Dolphin of the Amazon*	Non-fiction	770L	1,882
5	February 7	*Navajo Code Talkers*	Non-fiction	1170L	1,891
6	February 26	*Smokejumpers*	Non-fiction	900L	1,685
7	March 27	*Cliff Hanger*	Fiction	480L	1,190
8	April 10	*Moonwalk*	Fiction	630L	1,527
9	May 1	*Jim Thorpe's Bright Path*	Non-fiction	880L	2,413
10	May 15	*A Gift From the Heart*	Fiction	670L[a]	1,368

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Table 3

Descriptive Statistics for Reading Comprehension Posttest Scores

Posttest Scores	N[a]	*M*	*SD*	Min	Max
Time Point 1	29	3.97	1.27	2	7
Time Point 2	32	4.50	1.05	3	7
Time Point 3	35	5.20	1.18	3	8
Time Point 4	32	4.84	1.37	1	7
Time Point 5	32	4.59	1.39	2	8
Time Point 6	34	5.59	0.99	4	8
Time Point 7	34	5.68	1.12	4	8
Time Point 8	31	5.52	1.03	4	8
Time Point 9	34	6.18	1.22	3	8
Time Point 10	34	6.15	1.16	4	8
All	327	5.25	1.35	1	8

[a] N refers to the number of students completing a posttest at a given time point.

Table 4

Descriptive Statistics for AIMSweb R-CBM

		Group 1				Group 2				Group 3			
	Time Point	N[a]	M (*SD*)	Min	Max	N	M (*SD*)	Min	Max	N	M (*SD*)	Min	Max
Class X	Fall	6	116.2 (34.1)	58	153	5	145.4 (29.6)	118	184	6	135.7 (18.8)	126	174
	Winter	6	133.7 (38.1)	74	185	5	152.6 (19.3)	136	179	6	165.3 (23.6)	149	210
	Spring	6	141.5 (46.6)	70	204	5	171.2 (18.4)	150	200	6	170.2 (21.6)	152	212
Class Y	Fall	6	141.3 (40.6)	71	179	5	157.2 (19.8)	124	174	6	134.7 (33.1)	77	175
	Winter	6	156 (44.5)	78	198	5	173.4 (19.7)	146	195	6	140.3 (35.8)	94	196
	Spring	6	166.7 (48.8)	95	238	5	191 (24.8)	153	222	6	156.7 (37.3)	107	212

[a] N refers to the number of students completing the AIMSweb R-CBM in each discussion group. One student was not assigned a group during the fall AIMSweb R-CBM administration and is excluded from descriptive statistics.

Table 5

Multilevel Models for Reading Comprehension Posttest Outcome Variable

	Model 1		Model 2		Model 3		Model 4		Model 5	
Variable	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.	95% CI
Fixed Effects										
Intercept	5.21***	[4.97,5.45]	4.02***	[3.70,4.34]	4.05***	[3.73,4.38]	4.05***	[3.73,4.38]	4.06***	[3.73,4.38]
Time	--	--	0.22***	[0.18,0.25]	0.21***	[0.17,0.25]	0.21***	[0.17,0.25]	0.21***	[0.17,0.25]
Word Rareness	--	--	--	--	2.24***	[0.91,3.58]	--	--	1.17	[-1.28,3.63]
Root TTR	--	--	--	--	--	--	0.39***	[0.16,0.63]	0.22	[-0.21,0.65]
Random Effects										
Intercept	0.34***	[0.28,0.41]	0.38***	[0.32,0.45]	0.37***	[0.3,0.44]	0.37***	[0.3,0.44]	0.37***	[0.3,0.44]
Within Student	1.52	[1.38,1.65]	1.09	[0.98,1.21]	1.05***	[0.94,1.17]	1.05***	[0.94,1.16]	1.05***	[0.94,1.16]
N Level-1 Units[a]	345		345		327		327		327	
Deviance	1165.04		1066.61		999.81		1003.14		998.14	
Chi-Square	109.68***		151.36***		145.31***		145.39***		145.34***	

Note: Model 1 is the null model, Model 2 includes *Time* as a level 1 predictor, Model 3 includes *Time* and *Word Rareness*, Model 4 includes *Time* and *Root TTR*, Model 5 includes *Time*, *Word Rareness*, and *Root TTR*.

*$p < .05$. **$p < .01$. ***$p < .001$.

[a] Level-1 units in these models are student-text encounters.

Table 6

Descriptive Statistics for Independent Variables in Multilevel Models

	N[a]	*M*	*SD*	Min	Max
Word Rareness	327	1.13	0.14	0.72	1.46
Root TTR	327	3.29	0.74	1.11	5.78

[a]N refers to student-text encounters.

Table 7

Correlation Matrix of Dependent Variables in Multilevel Models (N=327 student-text encounters)

	Posttest Score	Word Rareness	Root TTR
Posttest Score	1	.145**	.172**
Word Rareness		1	.908**
Root TTR			1

**$p < .01$

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Table 8

Correlations Between Posttests Scores, Root TTR, and Word Rareness with AIMSweb R-CBM Scores

		AIMSweb Scores		
Time Point		Fall	Winter	Spring
1	Posttest	0.23		
	Root TTR	0.38*		
	Word Rareness	0.38*		
2	Posttest		0.03	
	Root TTR		0.11	
	Word Rareness		0.17	
3	Posttest		0.33	
	Root TTR		-0.11	
	Word Rareness		-0.12	
4	Posttest		0.40*	
	Root TTR		0.2	
	Word Rareness		0.19	
5	Posttest		0.32	
	Root TTR		0.41*	
	Word Rareness		0.28	
6	Posttest			0.05
	Root TTR			0
	Word Rareness			0.09
7	Posttest			-0.02
	Root TTR			0.29
	Word Rareness			0.15
8	Posttest			0.31
	Root TTR			0.18
	Word Rareness			0.18
9	Posttest			0.23
	Root TTR			0.07
	Word Rareness			0.14
10	Posttest			0.37
	Root TTR			0.42*
	Word Rareness			0.33

* $p < .05$

Note: Gray-highlighted values indicate that AIMSweb R-CBM scores have a greater correlation with *Root TTR* or *Word Rareness* than they do with posttest scores.

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Appendix

Table A1

Descriptive Statistics for Independent Variables at Time Point 1

Variable	N[a]	*M*	*SD*	Min	Max	Skewness	Kurtosis
Types	29	56.40	32.90	4.00	140.00	0.43	2.51
Tokens	29	314.60	227.30	13.00	975.00	1.07	3.85
TTR	29	0.20	0.04	0.10	0.30	0.51	4.71
Root TTR	29	3.10	0.90	1.10	4.50	-0.26	2.12
Log TTR	29	0.70	0.04	0.50	0.70	-1.77	7.55
Rolling TTR with window size 30	29	0.80	0.04	0.70	0.90	-0.35	2.49
Rolling TTR with window size 50	29	0.70	0.10	0.60	0.90	0.38	3.54
Rolling TTR with window size 100	29	0.60	0.10	0.50	0.90	1.89	8.75
Mean types per utterance	29	1.70	0.50	0.80	2.90	0.27	2.44
Mean tokens per utterance	29	9.30	3.90	3.20	16.90	0.50	2.11
Text Density	29	0.80	0.04	0.60	0.90	-2.33	11.19
Phrase Diversity	29	0.90	0.03	0.80	0.90	-1.16	4.31
Noncompressibility	29	0.60	0.10	0.50	0.90	1.58	4.41
Intersentential complexity	29	73.90	34.00	29.40	182.20	1.23	4.71
Average age of word acquisition	29	3.90	0.20	3.50	4.40	0.48	2.91
Word abstractness	29	473.50	19.20	440.70	535.10	0.94	4.97
Word rareness	29	1.10	0.20	0.80	1.30	-0.33	1.68
Average number of syllables in words	29	1.50	0.10	1.40	1.70	0.33	2.12
Word decoding demand	29	5.40	0.40	4.40	6.30	-0.15	3.76
LSA from English 100k semantic space	29	0.90	0.03	0.80	1.00	-0.99	3.87
LSA from TASA semantic space	29	0.50	0.10	0.20	0.70	-0.55	2.81

[a] N refers to students completing discussions.

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Table A2

Descriptive Statistics for Independent Variables at Time Point 2

Variable	N[a]	*M*	*SD*	Min	Max	Skewness	Kurtosis
Types	32	59.20	24.00	10.00	124.00	0.49	3.12
Tokens	32	342.40	163.70	27.00	654.00	0.23	2.07
TTR	32	0.20	0.05	0.10	0.40	1.78	6.91
Root TTR	32	3.20	0.60	1.90	4.80	0.45	3.37
Log TTR	32	0.70	0.02	0.70	0.80	0.17	2.67
Rolling TTR with window size 30	32	0.80	0.03	0.70	0.80	-0.45	3.85
Rolling TTR with window size 50	32	0.70	0.04	0.60	0.80	0.38	3.72
Rolling TTR with window size 100	32	0.60	0.10	0.50	0.80	1.82	8.57
Mean types per utterance	32	1.70	0.50	0.90	2.80	0.31	2.11
Mean tokens per utterance	32	9.20	2.80	5.00	15.10	0.46	2.25
Text Density	32	0.80	0.03	0.70	0.90	-2.18	8.43
Phrase Diversity	32	0.90	0.03	0.80	0.90	-1.89	9.70
Noncompressibility	32	0.60	0.10	0.50	0.80	1.65	5.20
Intersentential complexity	32	65.00	28.80	29.20	179.10	2.11	8.82
Average age of word acquisition	32	3.80	0.20	3.40	4.40	0.78	4.12
Word abstractness	32	454.90	12.00	420.90	476.90	-0.59	3.69
Word rareness	32	1.10	0.10	0.90	1.30	-0.02	2.65
Average number of syllables in words	32	1.50	0.10	1.30	1.70	0.80	5.11
Word decoding demand	32	5.40	0.30	4.80	6.20	0.64	3.49
LSA from English 100k semantic space	32	0.90	0.04	0.70	0.90	-3.84	19.33
LSA from TASA semantic space	32	0.20	0.10	0.04	0.40	1.19	4.97

[a] N refers to students completing discussions.

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Table A3

Descriptive Statistics for Independent Variables at Time Point 3

Variable	N[a]	*M*	*SD*	Min	Max	Skewness	Kurtosis
Types	35	64.20	28.90	12.00	133.00	0.39	2.62
Tokens	35	345.80	182.50	50.00	757.00	0.44	2.49
TTR	35	0.20	0.04	0.10	0.30	0.99	3.98
Root TTR	35	3.40	0.70	1.70	4.90	-0.04	2.86
Log TTR	35	0.70	0.03	0.60	0.80	-0.70	3.61
Rolling TTR with window size 30	35	0.80	0.03	0.70	0.90	0.18	2.43
Rolling TTR with window size 50	35	0.70	0.04	0.70	0.80	0.01	2.13
Rolling TTR with window size 100	35	0.60	0.04	0.50	0.70	0.69	3.48
Mean types per utterance	35	1.80	0.50	0.90	3.20	0.44	2.90
Mean tokens per utterance	35	9.40	3.10	4.30	18.70	0.76	4.00
Text Density	35	0.80	0.10	0.40	0.90	-4.16	22.34
Phrase Diversity	35	0.90	0.02	0.80	0.90	-1.04	7.00
Noncompressibility	35	0.60	0.04	0.50	0.70	1.31	5.48
Intersentential complexity	35	61.70	16.50	30.00	99.20	0.48	2.61
Average age of word acquisition	35	3.80	0.10	3.50	4.20	0.78	4.85
Word abstractness	35	438.00	19.60	407.50	476.50	0.21	1.89
Word rareness	35	1.20	0.10	0.90	1.40	-0.13	2.81
Average number of syllables in words	35	1.40	0.10	1.20	1.60	-0.64	3.42
Word decoding demand	35	5.20	0.30	4.40	5.80	-0.36	3.22
LSA from English 100k semantic space	35	1.00	0.02	0.90	1.00	-0.45	2.19
LSA from TASA semantic space	35	0.40	0.10	0.20	0.60	-0.49	3.05

[a] N refers to students completing discussions.

Table A4

Descriptive Statistics for Independent Variables at Time Point 4

Variable	N[a]	*M*	*SD*	Min	Max	Skewness	Kurtosis
Types	32	60.50	29.30	8.00	133.00	0.60	3.00
Tokens	32	324.10	189.00	28.00	871.00	0.97	3.87
TTR	32	0.20	0.03	0.20	0.30	0.95	3.49
Root TTR	32	3.30	0.70	1.50	4.60	-0.17	2.87
Log TTR	32	0.70	0.03	0.60	0.70	-1.16	5.05
Rolling TTR with window size 30	32	0.80	0.03	0.70	0.80	-0.81	4.40
Rolling TTR with window size 50	32	0.70	0.04	0.60	0.80	-0.64	3.84
Rolling TTR with window size 100	32	0.60	0.05	0.50	0.80	1.15	5.41
Mean types per utterance	32	1.50	0.30	0.90	2.30	0.78	3.15
Mean tokens per utterance	32	7.60	2.30	3.60	14.10	0.74	3.65
Text Density	32	0.80	0.04	0.70	0.90	-1.21	5.10
Phrase Diversity	32	0.90	0.02	0.80	0.90	-0.44	4.31
Noncompressibility	32	0.60	0.10	0.50	1.10	3.52	17.06
Intersentential complexity	32	59.00	17.40	30.10	119.20	1.40	6.33
Average age of word acquisition	32	3.80	0.20	3.50	4.10	-0.19	2.00
Word abstractness	32	448.50	16.90	408.40	480.80	-0.10	2.78
Word rareness	32	1.20	0.10	0.80	1.40	-0.46	3.25
Average number of syllables in words	32	1.50	0.10	1.30	1.70	-0.08	2.81
Word decoding demand	32	5.30	0.30	4.50	5.80	-0.61	3.24
LSA from English 100k semantic space	32	1.00	0.03	0.80	1.00	-3.19	14.65
LSA from TASA semantic space	32	0.50	0.10	0.30	0.80	-0.18	2.74

[a] N refers to students completing discussions.

Table A5

Descriptive Statistics for Independent Variables at Time Point 5

Variable	N[a]	*M*	*SD*	Min	Max	Skewness	Kurtosis
Types	32	61.30	27.70	12.00	115.00	0.18	2.09
Tokens	32	344.70	190.90	35.00	737.00	0.42	2.50
TTR	32	0.20	0.05	0.10	0.30	1.30	4.68
Root TTR	32	3.30	0.70	2.00	4.30	-0.14	2.02
Log TTR	32	0.70	0.02	0.70	0.70	-0.32	2.39
Rolling TTR with window size 30	32	0.80	0.03	0.70	0.90	0.20	2.59
Rolling TTR with window size 50	32	0.70	0.04	0.60	0.80	1.14	4.97
Rolling TTR with window size 100	32	0.60	0.10	0.50	0.80	2.28	10.80
Mean types per utterance	32	2.00	0.70	0.80	4.10	0.64	3.37
Mean tokens per utterance	32	10.10	3.30	4.10	17.50	0.47	2.72
Text Density	32	0.80	0.03	0.70	0.80	-0.80	3.60
Phrase Diversity	32	0.90	0.03	0.80	1.00	-1.10	4.32
Noncompressibility	32	0.60	0.10	0.50	0.90	1.71	5.29
Intersentential complexity	32	77.30	21.90	36.20	161.50	1.73	8.39
Average age of word acquisition	32	4.10	0.20	3.80	4.50	0.23	2.59
Word abstractness	32	469.60	12.50	448.20	496.00	0.15	2.31
Word rareness	32	1.10	0.10	0.90	1.30	-0.15	2.11
Average number of syllables in words	32	1.60	0.10	1.50	1.80	0.10	2.76
Word decoding demand	32	5.50	0.30	4.90	6.40	1.14	5.09
LSA from English 100k semantic space	32	0.90	0.10	0.70	1.00	-3.63	18.28
LSA from TASA semantic space	32	0.50	0.10	0.10	0.70	-0.68	3.56

[a] N refers to students completing discussions.

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Table A6

Descriptive Statistics for Independent Variables at Time Point 6

Variable	N[a]	*M*	*SD*	Min	Max	Skewness	Kurtosis
Types	34	66.20	37.80	16.00	165.00	0.80	3.02
Tokens	34	364.00	234.60	62.00	863.00	0.62	2.38
TTR	34	0.20	0.04	0.10	0.30	0.30	2.06
Root TTR	34	3.40	0.90	2.00	5.80	0.49	2.87
Log TTR	34	0.70	0.03	0.70	0.80	-0.23	2.04
Rolling TTR with window size 30	34	0.80	0.03	0.70	0.90	0.29	3.74
Rolling TTR with window size 50	34	0.70	0.03	0.70	0.80	0.64	2.82
Rolling TTR with window size 100	34	0.60	0.04	0.50	0.70	0.80	3.44
Mean types per utterance	34	1.90	0.70	0.90	4.00	0.93	3.68
Mean tokens per utterance	34	9.60	3.50	3.90	19.20	0.74	3.24
Text Density	34	0.80	0.03	0.70	0.90	-0.57	2.74
Phrase Diversity	34	0.90	0.04	0.70	0.90	-2.09	7.88
Noncompressibility	34	0.60	0.10	0.50	0.80	2.01	7.80
Intersentential complexity	34	65.40	18.80	26.40	125.70	0.90	4.78
Average age of word acquisition	34	3.90	0.20	3.10	4.30	-1.13	5.10
Word abstractness	34	452.50	21.20	399.20	484.70	-0.74	2.78
Word rareness	34	1.20	0.20	0.80	1.50	-0.34	2.53
Average number of syllables in words	34	1.50	0.10	1.30	1.60	-0.18	2.41
Word decoding demand	34	5.30	0.20	4.80	5.70	0.16	2.74
LSA from English 100k semantic space	34	0.90	0.03	0.90	1.00	-1.43	4.70
LSA from TASA semantic space	34	0.50	0.10	0.10	0.70	-1.29	5.07

[a] N refers to students completing discussions.

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Table A7

Descriptive Statistics for Independent Variables at Time Point 7

Variable	N[a]	*M*	*SD*	Min	Max	Skewness	Kurtosis
Types	34	62.60	22.40	20.00	110.00	0.19	2.70
Tokens	34	381.30	171.80	73.00	663.00	-0.02	1.94
TTR	34	0.20	0.05	0.10	0.30	1.53	6.14
Root TTR	34	3.20	0.50	1.80	4.30	-0.11	3.25
Log TTR	34	0.70	0.03	0.60	0.80	-0.37	5.32
Rolling TTR with window size 30	34	0.80	0.02	0.70	0.80	-0.03	2.77
Rolling TTR with window size 50	34	0.70	0.03	0.60	0.80	-0.21	2.84
Rolling TTR with window size 100	34	0.60	0.03	0.50	0.70	0.26	2.58
Mean types per utterance	34	2.00	1.10	0.50	6.80	2.62	12.86
Mean tokens per utterance	34	11.10	5.30	4.20	32.20	2.07	8.61
Text Density	34	0.80	0.03	0.70	0.80	-1.06	4.28
Phrase Diversity	34	0.90	0.03	0.80	0.90	-1.43	5.82
Noncompressibility	34	0.60	0.10	0.50	0.80	2.67	11.94
Intersentential complexity	34	67.90	18.80	32.50	105.70	0.39	2.40
Average age of word acquisition	34	3.80	0.20	3.50	4.00	-0.15	2.09
Word abstractness	34	457.90	13.70	433.60	487.60	0.08	2.37
Word rareness	34	1.10	0.10	0.90	1.30	-0.32	2.43
Average number of syllables in words	34	1.40	0.10	1.30	1.60	0.49	2.67
Word decoding demand	34	4.90	0.20	4.40	5.70	0.72	5.22
LSA from English 100k semantic space	34	0.90	0.02	0.90	1.00	-0.38	1.86
LSA from TASA semantic space	34	0.40	0.10	0.30	0.60	0.25	2.64

[a] N refers to students completing discussions.

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Table A8

Descriptive Statistics for Independent Variables at Time Point 8

Variable	N[a]	*M*	*SD*	Min	Max	Skewness	Kurtosis
Types	31	70.80	31.20	10.00	137.00	0.34	2.97
Tokens	31	465.50	264.60	40.00	1200.00	1.04	4.00
TTR	31	0.20	0.03	0.10	0.20	0.62	4.36
Root TTR	31	3.30	0.70	1.60	4.90	-0.23	3.66
Log TTR	31	0.70	0.03	0.60	0.70	-0.92	3.68
Rolling TTR with window size 30	31	0.80	0.03	0.70	0.80	-0.10	2.20
Rolling TTR with window size 50	31	0.70	0.03	0.70	0.80	0.30	3.18
Rolling TTR with window size 100	31	0.60	0.03	0.50	0.70	0.39	3.81
Mean types per utterance	31	1.60	0.70	0.80	3.80	1.59	4.82
Mean tokens per utterance	31	9.90	4.60	3.30	22.20	1.27	4.21
Text Density	31	0.80	0.03	0.70	0.80	-1.26	4.59
Phrase Diversity	31	0.90	0.03	0.80	0.90	-1.43	5.22
Noncompressibility	31	0.60	0.10	0.50	0.70	0.77	2.96
Intersentential complexity	31	72.50	27.90	44.40	186.40	2.52	10.31
Average age of word acquisition	31	3.80	0.20	3.50	4.20	-0.09	2.08
Word abstractness	31	465.70	15.90	428.90	493.80	-0.16	2.37
Word rareness	31	1.20	0.10	0.90	1.40	-0.40	2.60
Average number of syllables in words	31	1.50	0.10	1.10	1.70	-1.62	8.24
Word decoding demand	31	5.00	0.30	4.10	5.70	-0.24	4.22
LSA from English 100k semantic space	31	0.90	0.03	0.80	1.00	-0.74	2.73
LSA from TASA semantic space	31	0.40	0.10	0.10	0.60	-0.53	2.11

[a] N refers to students completing discussions.

Table A9

Descriptive Statistics for Independent Variables at Time Point 9

Variable	N[a]	*M*	*SD*	Min	Max	Skewness	Kurtosis
Types	34	59.20	34.70	11.00	131.00	0.43	2.23
Tokens	34	332.30	237.70	36.00	1126.00	1.09	4.73
TTR	34	0.20	0.10	0.10	0.40	1.96	8.27
Root TTR	34	3.20	0.90	1.60	4.90	0.08	2.23
Log TTR	34	0.70	0.03	0.60	0.80	-0.54	3.45
Rolling TTR with window size 30	34	0.80	0.03	0.70	0.90	1.05	5.01
Rolling TTR with window size 50	34	0.70	0.04	0.60	0.80	1.03	4.62
Rolling TTR with window size 100	34	0.60	0.10	0.50	0.80	1.95	7.93
Mean types per utterance	34	1.70	0.60	0.70	3.30	0.74	3.07
Mean tokens per utterance	34	8.90	3.30	3.80	18.00	0.68	2.93
Text Density	34	0.80	0.03	0.70	0.80	-0.37	2.46
Phrase Diversity	34	0.90	0.02	0.80	0.90	-1.10	4.54
Noncompressibility	34	0.60	0.10	0.50	0.80	1.74	5.54
Intersentential complexity	34	66.10	19.10	29.30	120.10	0.76	3.95
Average age of word acquisition	34	3.90	0.20	3.30	4.10	-1.14	4.03
Word abstractness	34	461.10	17.30	424.60	503.30	0.50	3.76
Word rareness	34	1.10	0.20	0.70	1.40	-0.33	2.36
Average number of syllables in words	34	1.50	0.10	1.30	1.80	-0.13	3.20
Word decoding demand	34	5.20	0.30	4.50	5.90	-0.39	4.43
LSA from English 100k semantic space	34	0.90	0.04	0.80	1.00	-1.98	7.58
LSA from TASA semantic space	34	0.40	0.20	0.10	0.70	-0.38	2.47

[a] N refers to students completing discussions.

Table A10

Descriptive Statistics for Independent Variables at Time Point 10

Variable	N[a]	*M*	*SD*	Min	Max	Skewness	Kurtosis
Types	34	70.60	32.30	11.00	136.00	0.08	2.36
Tokens	34	422.20	241.40	27.00	1019.00	0.55	2.99
TTR	34	0.20	0.10	0.10	0.40	2.23	8.13
Root TTR	34	3.40	0.70	2.10	4.70	-0.24	2.32
Log TTR	34	0.70	0.02	0.70	0.70	-0.32	2.72
Rolling TTR with window size 30	34	0.80	0.04	0.70	0.90	1.46	5.54
Rolling TTR with window size 50	34	0.70	0.05	0.70	0.90	2.05	8.73
Rolling TTR with window size 100	34	0.60	0.10	0.50	0.90	2.63	11.50
Mean types per utterance	34	1.60	0.60	0.70	3.70	1.28	5.90
Mean tokens per utterance	34	8.80	3.20	4.20	16.70	0.60	2.42
Text Density	34	0.80	0.10	0.60	0.90	-2.28	8.74
Phrase Diversity	34	0.90	0.02	0.80	0.90	-1.18	5.04
Noncompressibility	34	0.60	0.10	0.50	0.80	1.88	6.09
Intersentential complexity	34	68.60	26.80	31.20	169.00	1.70	7.08
Average age of word acquisition	34	3.90	0.20	3.50	4.20	-0.28	2.49
Word abstractness	34	455.60	14.60	426.70	485.30	0.21	2.48
Word rareness	34	1.10	0.20	0.70	1.40	-0.73	2.94
Average number of syllables in words	34	1.50	0.10	1.30	1.70	0.25	3.52
Word decoding demand	34	5.30	0.20	4.80	5.60	-0.65	2.41
LSA from English 100k semantic space	34	0.90	0.03	0.80	1.00	-2.65	10.87
LSA from TASA semantic space	34	0.30	0.20	0.04	0.60	-0.30	2.18

[a] N refers to students completing discussions.

Table A11

Range of Amount of Student Talk Across Discussion Time Points

	Total Number of Words Spoken			Total Number of Utterances		
	Min	Max	Range	Min	Max	Range
Class X Student 1	242	952	710	29	80	51
Class X Student 2	135	751	616	21	103	82
Class X Student 3	28	297	269	5	44	39
Class X Student 4	69	370	301	7	27	20
Class X Student 5	149	856	707	20	72	52
Class X Student 6	108	534	426	14	72	58
Class X Student 7	43	636	593	11	61	50
Class X Student 8	58	1412	1354	9	106	97
Class X Student 9	319	1516	1197	20	75	55
Class X Student 10	239	867	628	32	87	55
Class X Student 11	112	521	409	13	55	42
Class X Student 12	49	365	316	13	43	30
Class X Student 13	78	673	595	10	78	68
Class X Student 14	440	1403	963	31	86	55
Class X Student 15	175	821	646	11	37	26
Class X Student 16	66	324	258	5	26	21
Class X Student 17	288	1202	914	26	82	56
Class Y Student 1	18	345	327	5	41	36
Class Y Student 2	283	930	647	27	80	53
Class Y Student 3	245	632	387	19	53	34
Class Y Student 4	42	647	605	8	102	94
Class Y Student 5	101	576	475	9	53	44
Class Y Student 6	355	977	622	34	78	44
Class Y Student 7	463	1277	814	34	78	44
Class Y Student 8	290	1050	760	27	81	54
Class Y Student 9	288	997	709	30	94	64
Class Y Student 10	58	455	397	13	80	67
Class Y Student 11	351	664	313	27	68	41
Class Y Student 12	205	826	621	20	68	48
Class Y Student 13	110	595	485	20	75	55
Class Y Student 14	455	1064	609	31	105	74
Class Y Student 15	144	636	492	17	70	53
Class Y Student 16	51	335	284	5	35	30
Class Y Student 17	117	892	775	12	95	83

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174

Class Y Student 18	187	530	343	20	56	36

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice*, *37*, 20-34. 10.1111/emip.12174