Journal of Educational Psychology

Beyond Text Complexity: Production-Related Sources of Text-Based Variability in Oral Reading Fluency

Van Rynald T. Liceralde, Anastassia Loukina, Beata Beigman Klebanov, and John R. Lockwood

Online First Publication, October 4, 2021. http://dx.doi.org/10.1037/edu0000532

CITATION

Liceralde, V. R. T., Loukina, A., Beigman Klebanov, B., & Lockwood, J. R. (2021, October 4). Beyond Text Complexity: Production-Related Sources of Text-Based Variability in Oral Reading Fluency. *Journal of Educational Psychology*. Advance online publication. http://dx.doi.org/10.1037/edu0000532

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Educational Psychology

© 2021 American Psychological Association
ISSN: 0022-0663

https://doi.org/10.1037/edu0000532

Beyond Text Complexity: Production-Related Sources of Text-Based Variability in Oral Reading Fluency

Van Rynald T. Liceralde[1, 2], Anastassia Loukina[1], Beata Beigman Klebanov[1], and John R. Lockwood[1]
[1] Educational Testing Service, Princeton, New Jersey, United States
[2] Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill

Variability in oral reading fluency (ORF), an indicator of foundational reading skills, has been linked to characteristics of texts. Such *text-based variability* in ORF has been traditionally attributed to text complexity, but substantial text-based variability has still been observed after accounting for text complexity. We consider that *oral* reading requires pronouncing the text aloud, which makes it subject to the same articulatory and prosodic constraints as other types of speech productions. Thus, texts with similar levels of complexity may still differ in expected durations when read aloud because of the texts' segmental and prosodic structure, leading to differences in reading rate. We propose that these *production-related effects* are also important sources of text-based ORF variability. Data from upper elementary school students in the United States reading a large variety of passages from a popular fiction book showed that a composite measure of production-related effects (i.e., reading rate estimates derived from a text-to-speech synthesis system) explained a substantial amount of text-based ORF variability. Follow-up exploratory analyses indicated that these production-related effects are robust. Because text complexity metrics consist of features that also tap into production constraints, our results motivate an updated interpretation of text complexity effects on ORF and highlight the importance of accounting for production-related effects on ORF, which remain to be acknowledged in the ORF literature as potential sources of text-based variability.

Educational Impact and Implications Statement

Although the achievement of oral reading fluency (ORF) is a milestone in reading development on its own, it is also a popular indicator of other critical reading skills, such as reading comprehension. ORF is measured by having students read some text aloud for some time, and the average number of words they correctly read in a minute is calculated. One well-known complication of this measure is that the measure partly depends on the properties of the text being read. More complex texts are expected to take longer to read, so text complexity metrics are often used when selecting passages to measure ORF. In this article, we show that the relationship between text complexity and ORF is not straightforward because some properties of texts could make them both more complex and slower to utter, but other properties could increase the complexity but make them *faster* to utter; thus, the *oral* and the *reading* parts of "oral reading" both need to be modeled to successfully explain the effects of text properties on ORF.

Keywords: grade school, oral reading fluency, speech production, text complexity, text effects

Successful reading requires the mastery of foundational phonological and orthographic skills. A standard proxy of these skills in children is *oral reading fluency* (ORF), which is measured as *words correct per minute* (WCPM) to account for both speed and accuracy in reading (Fuchs et al., 2001). Reading both swiftly and accurately reflects a mastery of decoding skills, which indicates the successful mapping of phonological codes to orthographic codes. In turn, ORF has been shown to strongly relate to reading comprehension because the mastery of these skills shifts the cognitive load during reading from decoding processes to higher-level

Van Rynald T. Liceralde https://orcid.org/0000-0001-8447-2424

Anastassia Loukina https://orcid.org/0000-0002-8213-3362

John R. Lockwood is now at Duolingo, Pittsburgh, Pennsylvania, United States.

We thank Tenaha O'Reilly, Don Powers, and Zuowei Wang whose careful reading greatly improved the article. We also thank Art Graesser and participants at the 2019 Annual Meeting of the National Council on Measurement in Education for important discussions on an earlier version of this work.

Correspondence concerning this article should be addressed to Van Rynald T. Liceralde, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, 235 East Cameron Avenue, Chapel Hill, NC 27599, United States. Email: vrtliceralde@gmail.com

semantic and discourse integration (Fuchs et al., 2001, 1988; Kim et al., 2010, 2011; Kim & Wagner, 2015; Pikulski & Chard, 2005; Torgesen et al., 2001).

Because ORF is considered a measure of children's foundational reading skills, it is important to understand systematic sources of variability in ORF, such as the characteristics of texts that children read. The most prominent of these text characteristics is *text complexity*. The appeal of text complexity as a source of systematic text-based ORF variability is intuitive: More complex texts tend to be read more slowly because processing more complex material takes more time. One way that standard formal assessments of ORF have reduced text-based variability is by using texts that are constructed to have comparable text complexity. However, substantial text-based variability in WCPM has still been observed in these texts (Ardoin et al., 2005; Compton et al., 2004; Francis et al., 2008), suggesting that text complexity is not the only source of systematic text-based variability in ORF.

In this article, we pursue two new avenues in studying text-based ORF variability. First and most importantly, we consider that *oral* reading requires pronouncing the text aloud, which makes it subject to the same articulatory and prosodic constraints as other types of speech productions. Despite much work showing developmental differences in the production and use of prosody when reading aloud (e.g., Ardoin et al., 2013; Kim & Wagner, 2015; Miller & Schwanenflugel, 2006), text features that constrain prosody (e.g., the segmental composition of the text, the amount of phrasal boundaries) have not been incorporated in studies and accounts of text-related ORF effects in educational psychology (e.g., Barth et al., 2014; Compton et al., 2004; Francis et al., 2008). We propose that these *production-related effects*—which the phonetic and speech production literatures treat as fundamental constraints on the duration of utterances—are also important sources of text-based ORF variability. That is, texts that have similar levels of complexity or the same number of words or syllables may still have different expected durations when read aloud due to differences in texts' production demands, and this would lead to differences in reading rate.

Second, we extend the empirical basis of research on text-based ORF variability by considering materials beyond those selected or constructed for formal assessments of ORF. Understanding the impact of text features on ORF in texts beyond standardized materials would help evaluate whether ORF can also be measured on materials that are read for knowledge or for pleasure, which would have wide-ranging practical implications.

We first explain the reasons for considering materials beyond standardized test passages and then focus on production-related effects in oral reading, which we consider to be the main contribution of this article.

Data From Extended Book Reading

Rather than constructing reading materials or using standardized test passages, as is common practice when formally assessing ORF, we consider a setting where the oral reading data come from a classroom-based oral reading program that uses a popular novel. To our knowledge, this is the first attempt to analyze text-based variability in ORF using data from children casually reading a popular fiction book for an extended period of time.

Using such data extends the empirical basis for studying the relationships between text features and children's oral reading because reading full-length fiction books constitutes a different reading experience from reading test passages. Passages in fiction books exhibit substantial variation in text complexity, above and beyond what children would encounter in a typical ORF assessment. For example, passages from *Harry Potter and the Sorcerer's Stone* (Rowling, 1997) exhibit Grade 2 to Grade 11 complexity (Beigman Klebanov et al., 2017), and passages from *Black Beauty* exhibit Grade 2 to Grade 9 complexity (see Table A8 of Milone, 2014). This makes them quite different from standardized test passages, which are written to specification. For example, authors of passages from the *Dynamic Indicators of Basic Early Literacy Skills* (Good & Kaminski, 2002), a standard ORF assessment, are instructed to meet a specified target in terms of the interplay between average sentence length and average word length as reflected in the Flesch-Kincaid formula (Flesch, 1948; Kincaid et al., 1975), a classic text complexity metric. The authors are also told to use grade-level appropriate vocabulary and avoid dialogue, slang, and too much humor, among other criteria (Biancarosa et al., 2019; pp. 26–27).

Clearly, texts for assessing reading skills exhibit different characteristics from fiction books because these texts were written with different goals. Nevertheless, the same text features could affect the oral reading of both types of texts. The constraints placed on standardized test passages, however, can limit variation in text features in ways that make it difficult to appreciate the range of text-based effects on ORF without a contrasting set of materials where such constraints are not imposed. This motivates our use of a popular fiction book in the current study to investigate potential sources of text-based ORF variability beyond text complexity.

In terms of educational practice, the ability to measure ORF on a wider variety of texts, including texts written for reading rather than for testing, would increase educators' flexibility in measuring ORF by providing an expanded set of materials to measure ORF. Moreover, if students' fluency can be assessed while they are reading to learn or simply for pleasure, this could save valuable class time that would otherwise be spent on reading special ORF assessment materials.

Beyond Text Complexity: Production-Related Effects in ORF

The effects of text complexity on ORF have been primarily attributed to how its component features affect the processing of the text. As a historical example, Flesch-Kincaid (Flesch, 1948; Kincaid et al., 1975) uses average sentence length (in number of words) and average word length (in number of syllables) to estimate the complexity of a text. Because longer words tend to be less frequent and more morphologically complex (New et al., 2006) and longer sentences tend to be more syntactically complex (Wang, 1970), texts that consist of longer words and longer sentences would take longer to process on average (Lewis & Vasishth, 2005; Wang, 1970).

Similar explanations have been used to describe how text features in more recent and comprehensive text complexity metrics such as Lexile (Stenner et al., 2006), Coh-Metrix (Graesser et al., 2011), and TextEvaluator (Sheehan et al., 2014) might affect

ORF. Theories on narrative comprehension and discourse processing posit that text complexity is related to ORF because its component features affect how quickly various levels of mental representations are accessed and integrated during reading.

To overview, the event segmentation theory (Zacks et al., 2009; Zacks & Swallow, 2007), the event index model (Zwaan et al., 1995), and the structure building framework (Gernsbacher, 1997) converge on the idea that readers continuously monitor texts for when ideas or events start and end. In doing so, readers form "event models" for the text which consist of expectations about what the upcoming text is likely to talk about. When the upcoming text is aligned with the reader's expectations, the reader's event model is reinforced and the processing of upcoming text is facilitated. When the two begin to drift apart, the reader resolves this separation by noting its location as an event boundary. The reader then updates their event model with this additional structure, and they momentarily suppress their predictions and become more careful in processing upcoming text until the event model becomes stable enough to produce reliable predictions for upcoming text again.

Based on these accounts, text complexity is related to ORF because its component features, such as narrativity, lexical cohesion, and concreteness, affect readers' ability to form event models. For example, narrativity would affect ORF because it affects the predictability of events in the text. Texts that are more narrative/story-like rather than informational/expository would be read faster because the predictable structure of narrative texts makes it advantageous for readers to rely on their expectations and prior knowledge of such texts to facilitate their reading (Perfetti, 1994). Lexical cohesion would affect ORF because it facilitates the connections between potential event boundaries. Texts that have "deeper" cohesion, where independent ideas are more tightly linked with various linguistic devices such as connectives (e.g., *because*, *therefore*; Graesser et al., 2011), would be read faster than texts that are less cohesive. Connecting ideas would keep past information activated as readers update their discourse model (O'Brien et al., 1995; Suh & Trabasso, 1993), and this sustained activation of overlapping content would facilitate processing of the text (Kintsch, 1988), which would lead to faster reading. Lexical concreteness would affect ORF because it generates imagery that strengthens event memory representations (Marschark et al., 1994; McDaniel et al., 1995) and reinforces the reader's event model. Texts with more concrete words would be read faster because they widen event segments and reduce the need to update event models by keeping them stable. Consistent with these expectations, empirical studies have shown that ORF is strongly related to text complexity (Barth et al., 2014; Betts et al., 2009; Fuchs et al., 2001; Hintze et al., 1998). In Barth et al. (2014), text complexity and various component features accounted for about half of the text-based ORF variability of middle-school students.

Because of the strong relations between text complexity and ORF, text complexity metrics are often used to select or create passages for measuring ORF (e.g., Barth et al., 2014; Biancarosa et al., 2019). Given that a substantial part of text-based ORF variability remains to be explained after controlling for the effects of text complexity and/or their component features (Ardoin et al., 2005; Barth et al., 2014; Compton et al., 2004; Francis et al., 2008), it is surprising that other effects beyond those related to the processing of the text have yet to be closely studied as potential sources of this variability.

Drawing on extensive literature on speech production, we consider in this article that production-related effects are another important source of text-based ORF variability. Because oral reading involves reading *aloud*, properties of the text that affect its production generally affect how the text is read and should thus be considered as possibly contributing to text-based ORF variability.

There is abundant evidence that systematic differences in the duration of utterances arise because of segmental (subword) and prosodic effects (for reviews, see Hirschberg, 2002; Loukina et al., 2018; White, 2014), and these differences manifest at multiple linguistic levels.[1] At the segmental level, different segments have different intrinsic durations and they are subject to consistent effects of phonetic context (Klatt, 1976; Peterson & Lehiste, 1960; van Santen, 1992): for example, high vowels tend to be shorter than low vowels, vowels tend to be shorter when followed by a voiceless consonant than when followed by a voiced consonant (Crystal & House, 1988; House & Fairbanks, 1953), and segments tend to be longer in word-initial positions (Turk & Shattuck-Hufnagel, 2000, 2007). At the word level, syllables bearing the lexical stress tend to be lengthened, particularly in monosyllabic words, but this lengthening attenuates and spreads to unstressed syllables in polysyllabic words (Turk & Shattuck-Hufnagel, 2000; White & Turk, 2010). In addition, words tend to be uttered more quickly when they are frequent, predictable, and have been repeated (Bell et al., 2009; Zhao & Jurafsky, 2009). At the phrase/clause/sentence level, prosodic boundaries are indicated by phrase-final lengthening and often by a pause, with sentence-final pauses being the longest (Bailly & Gouvernayre, 2012; Burrows et al., 2005; Pfitzinger & Reichel, 2006; Turk & Shattuck-Hufnagel, 2000; 2007). Thus, two texts containing the same number of words or syllables may still have different expected durations when read aloud because of differences in segmental and prosodic structure, and this would lead to differences in reading rate. Importantly, children exhibit these production effects, even as developing readers (Miller & Schwanenflugel, 2006).

These findings collectively suggest that effects beyond text complexity—production-related effects, specifically—ought to be considered when studying sources of text-based ORF variability to develop a fuller picture of where this variability arises. However, capturing production-related effects is challenging because there are many text properties that affect production and they interact in complex ways (Hirschberg, 2002; White, 2014). Fortunately, many of these features are already incorporated in timing models used in modern *text-to-speech synthesis* (TTS) systems: the durations of segments and pauses produced by a TTS system for a given text are computed based on models that incorporate properties such as segment identity and context, stress, part-of-speech context, and prominence (e.g., Capes et al., 2017). Thus, these systems generate renditions that serve as good estimates of how a comprehensive set of text properties would interact to constrain how the text would be uttered. We used TTS-based estimates of reading rates for texts to capture production-related effects on oral

[1] Some of these effects are universal, others are language-specific. In this review, we focus on the ones that have been shown to exist in English, the language considered in this article.

reading and to estimate how much text-based ORF variability is attributable to production-related effects.

Research Questions

In this study, we analyzed data from upper elementary school children in northeast United States who read out loud *Harry Potter and the Sorcerer's Stone* (HP1; Rowling, 1997), a popular fiction book, with a reading app in a relaxed, naturalistic setting. Using these data, we sought to answer the following questions about text-based ORF variability:

1. How much text-based variability is present in WCPM based on reading passages from a fiction book?
2. How much of the text-based variability in WCPM can text complexity explain on its own?
3. How much of the text-based variability in WCPM can production-related effects explain on their own?
4. Do production-related effects explain text-based WCPM variability above and beyond text complexity?

Method

Participants

Fifty-six students (26 females [46%]; $M_{\text{age}} = 9.61$; $SD = 1.06$; no age information for seven students[2]) were part of a larger sample that participated in this naturalistic reading study that was conducted in the context of summer camp and afterschool reading activities (see Procedures). Of these 56 students, 46 (82%) were from English-speaking households, one was from a Spanish-speaking household, and nine did not have this information available. Twenty-seven of these students (eight females [30%]; $M_{\text{age}} = 9.58$, $SD = 1.32$) participated in a camp held during the summer of 2017 in New Jersey. The other 29 students were similar in age (18 females [62%]; $M_{\text{age}} = 9.63$, $SD = 0.78$) and participated in two camps held during the summer of 2018 in the New York metropolitan area.

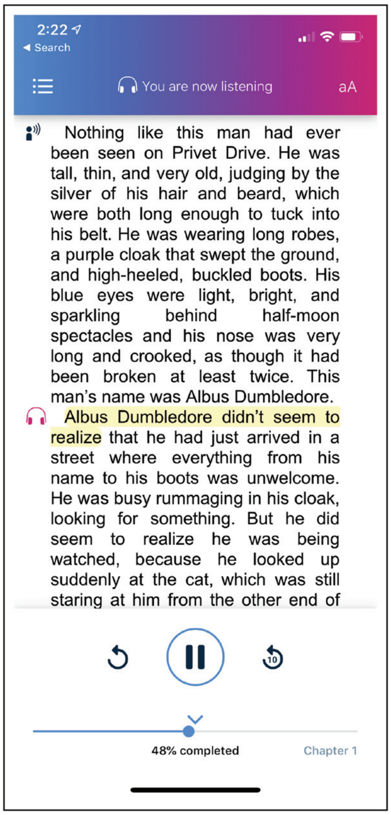
The sampled students are in upper elementary grade levels, at which point readers are continuing to develop (as opposed to just starting or to have largely established) their skills at reading efficiency, phonological decoding, and vocabulary (Daane et al., 2005; Torgesen et al., 2001). At this stage, readers are generally transitioning from "learning to read" to "reading to learn," where readers increasingly focus on using their reading skills to learn in various subject areas.

Materials: Reading App

The oral reading data were collected through *Relay Reader* (formerly *MyTurnToRead*; Beigman Klebanov et al., 2019; Madnani et al., 2019), an app designed to facilitate sustained reading among developing readers by providing them with additional practice to transition from "learning to read" to "reading to learn" (see Figure 1 for a screenshot of the mobile version). It encourages readers to read for meaning and pleasure through the use of (audio)books and technology to enhance engagement, alleviate frustration, and obtain immediate feedback. The app provides an interleaved and interactive oral reading context, where the reader takes turns reading passages aloud (see "reader" icon in Figure 1) with a model narrator (i.e., the audiobook narrator; "headphones" icon) and regularly answers comprehension questions about the passages to monitor understanding of the text. To create the passages, the book is divided into paragraphs and consecutive paragraphs are combined until a desired passage length (in number of words; determined by the researchers for this study) is achieved. Thus, each passage always starts and ends with a paragraph break: For example, in Figure 1, one paragraph is sufficient for the passage

Figure 1

A Screenshot of the iOS Version of Relay Reader



Nothing like this man had ever been seen on Privet Drive. He was tall, thin, and very old, judging by the silver of his hair and beard, which were both long enough to tuck into his belt. He was wearing long robes, a purple cloak that swept the ground, and high-heeled, buckled boots. His blue eyes were light, bright, and sparkling behind half-moon spectacles and his nose was very long and crooked, as though it had been broken at least twice. This man's name was Albus Dumbledore.

Albus Dumbledore didn't seem to realize that he had just arrived in a street where everything from his name to his boots was unwelcome. He was busy rummaging in his cloak, looking for something. But he did seem to realize he was being watched, because he looked up suddenly at the cat, which was still staring at him from the other end of

Note. The text is highlighted during the narrator's turn (which is indicated by the "headphones" icon) to encourage the reader to silently read along with the narrator. Image used by permission; © 2019 Educational Testing Service. www.ets.org. Text used by permission; "Harry Potter and the Sorcerer's Stone" by J.K. Rowling. © 1997 by J.K. Rowling. See the online article for the color version of this figure.

[2] All sites/summer camps grouped the students by age; students with no age information were enrolled in the same groups as students for whom age information was available.

assigned to the reader, given the desired passage length. The reader's oral reading is recorded in the app.

Procedure

Data Collection A

Data Collection A took place between July 2017 and October 2017 across five locations/sites in New Jersey. A prototype version of the app was locally installed in laptops that were brought to each site during the data collection period. The data collection was supervised by the project staff.

Data collection at each site occurred over five consecutive days, where each day consisted of a 20-minute session. The first session involved assessing students' prior knowledge of HP1 (i.e., Have they read the book and/or seen the movie before?) and other reading-related activities. The subsequent four sessions involved reading with the app. During a reading session, each student individually read with the app through headphones with built-in microphones. Students were told to read aloud as they typically would.

All students started at the beginning of the book in the first reading session. The session started with students listening to the recording of Jim Dale, a renowned British actor who narrates the audiobook (Rowling & Dale, 2016), reading the first passage of the book. After Jim Dale's "turn," students indicated that they were ready to start their turn, after which the app started recording their oral reading. Students took as much time as they needed for a turn, after which they clicked a button to indicate that they were done reading. No feedback was provided to students during their reading turns, and students could not rerecord their reading of previous passages. Students "took turns" with Jim Dale. When students moved on to the next chapter, the beginning passage was always read by Jim Dale regardless of who read the final passage of the previous chapter.

After 20 minutes into the reading session, the app automatically timed out and students answered three multiple-choice or yes/no comprehension questions about the passages they read during the session. These questions were fact-based, in that they were always about information explicitly described in the passages (e.g., What did the "put-outer" do? Who did Dumbledore plan for Harry to live with? Did Harry enjoy spending time with Mrs. Figg? How did the letters arrive on Saturday? Why did Harry have no friends at school?). These questions were intended to check that students were paying attention to the story while they were reading aloud.

Each subsequent reading session always started with students listening to Jim Dale read the passage that they last read but did not finish in the previous session. Other than this, the subsequent reading sessions proceeded the same way as the first.

Data Collection B

Data Collection B took place between June 2018 and August 2018 in two sites in the New York metropolitan area. For this data collection, a new beta version of the app (Madnani et al., 2019) was installed on tablets that were either available at the site or provided to the students by the research staff.

In both sites, regular reading sessions with the app were scheduled as part of the camp program. The reading sessions were scheduled and monitored by the camp instructors. One program ran for 6 weeks and included a reading session with the app for 20–50 minutes, four days a week, with fewer days in the first week of the camp. The second program ran for a total of 8 weeks (different children were enrolled for a different number of weeks) with a variable reading schedule depending on other camp activities; each reading session included about half an hour of reading and half an hour of related games and activities.

The structure of the reading sessions was the same as in Data Collection A: Students alternated between listening to the narration and reading aloud during their own turn. However, unlike Data Collection A, the sessions did not automatically time out after 20 minutes of reading. Students were also asked two multiple-choice reading comprehension questions for every two turns they made; the turn count reset at the end of each session, so that at the start of a new session, students still made two turns before answering questions. As in Data Collection A, students could not rerecord their reading of previous passages; they also could not reanswer comprehension questions they already answered.

At the end of the data collection period, students were asked if they have read the book or seen the movie before. They also indicated how much they agreed with the statement "The *Harry Potter* book was boring" on a four-point scale (*Strongly disagree*, *Disagree*, *Agree*, *Strongly agree*) as a rough measure of overall interest and engagement with the book.

Focal Measures

Oral Reading Fluency: WCPM

All recordings of children's reading were transcribed by a professional transcription agency. The transcribers were given the audio recording and the passage text and were asked to indicate which words were substituted, deleted, and inserted by the student. The transcribers also indicated any task-irrelevant speech. Finally, the transcriptions contained timestamps in seconds indicating the beginning of the first uttered word in the passage and the end of the last uttered word in the passage.

We used these transcriptions to compute WCPM. To do this, we first computed the number of words in each passage that have not been marked as deleted or substituted. We then computed the time (in seconds) it took the student to read the passage by taking the difference in the timestamps of the first and the last word uttered for that turn. With these numbers, we computed WCPM as shown in Equation 1.[3]

$$\text{WCPM} = \frac{\text{total number of words read correctly}}{\text{time spent reading in seconds}} \times 60 \qquad (1)$$

[3] WCPM is typically estimated by either having the reader read a test passage for one minute (Barth et al., 2014; Daane et al., 2005; Good & Kaminski, 2002) or by having the reader read a passage and dividing the number of correctly read words by the duration of the turn in minutes (Bernstein et al., 2017; Daane et al., 2005), as done here. Daane et al. (2005) compared the two approaches in a nationally representative study of Grade 4 students in the United States and reported that both were positively correlated with comprehension ability.

Text Complexity: TextEvaluator Scores

To examine the effects of text complexity, we used TextEvaluator as our measure of the text complexity of HP1 passages. *TextEvaluator* (formerly *SourceRater*) is a state-of-the-art tool that has been validated against expert judgements of text complexity (Napolitano et al., 2015; Sheehan et al., 2014). TextEvaluator extracts a wide range of linguistic features from passages—which includes simpler text complexity measures such as Flesch-Kincaid—and combines these features into 8 dimensions based on factor analysis (i.e., academic vocabulary, concreteness, degree of narrativity, interactive/conversational style, level of argumentation, lexical cohesion, syntactic complexity, and word unfamiliarity). It then computes text complexity scores based on combinations of these dimensions that have been optimized for literary, informational, and mixed texts. Thus, the tool provides a comprehensive account of features that have been shown to affect text complexity. TextEvaluator is highly correlated with various text complexity metrics such as Lexile (Stenner et al., 2006), Coh-Metrix (Graesser et al., 2011), Reading Maturity Metric (RMM; Landauer et al., 2011), ATOS (Milone, 2014), Degrees of Reading Power (DRP; Zeno et al., 1995), and Reader-specific Practice (REAP; Collins-Thompson & Callan, 2004), and it generally outperformed these metrics when evaluated against standard benchmarks and teachers' estimates of grade level difficulty (Nelson et al., 2012; Sheehan et al., 2014; Toyama et al., 2017). TextEvaluator scores are scaled from 100 to 2,000, with higher numbers indicating higher text complexity.

Production Constraints: Reading Rate Estimates From TTS Synthesis

To capture production constraints, we need a model that takes into account text properties that would predict how fast texts would be read in terms of words per minute, such as the nature of individual segments, the distribution of stressed syllables, and prosodic boundaries. One way to account for these properties is by building a model that predicts the relative duration of each segment and the location and duration of pauses in a given text. We can then use this model's predictions to compute the expected words-per-minute rate for the text.

Because there is a considerable number of text properties that affect speech production, building a model that captures these properties is a challenging task. Fortunately, accurate prediction of segment and pause durations based on text properties is a task that has many applications, including speech technologies, and it has been actively explored in research on text-to-speech synthesis (Tokuda et al., 2016; van Santen, 1994; Yoshimura et al., 1999; Zen et al., 2009). Modern text-to-speech (TTS) systems incorporate complex timing models for estimating the duration of each segment based on segmental, prosodic, and other factors. Therefore, we used a state-of-the-art TTS engine to model production-related effects. To make it easy to reproduce our results, we used Apple Inc.'s widely available and built-in TTS engine (OS X 10.11.6). Like other TTS engines, Apple's systems use text features such as stress, part-of-speech, context, and sentence type to model timing patterns (among other things) and are evaluated against the judgment of native listeners (see, e.g., Capes et al., 2017) for a description of a similar system and its evaluation).

We used the male-Alex voice with default settings to synthesize audio files of the HP1 passages and computed reading rate estimates of the passages using these audio files. To compute the expected reading rate for each passage in words per minute, the total number of words in the passage was divided by the duration of the passage's synthesized audio in seconds from the start of the first word to the end of the last word and then multiplied by 60. Because these reading rate estimates were derived from a large set of features that have been shown to affect utterance duration, these estimates can be thought to come from a digital model of oral sentence reading that adheres to phonological and prosodic constraints in production.

Data Description

The dataset used in this study consisted of 964 reading turns from the 56 students described in the Participants section, and it came from a larger dataset that was preprocessed to address measurement issues related to collecting data in naturalistic settings (see Appendix A for details). The dataset contained 346 different passages from the first ten chapters of the book (mean length = 175.00 words; *SD* = 77.82; Min = 14.00; Max = 436.00). Because students read at their own pace and passages were assigned to them based on where they started reading at the beginning of a reading session (i.e., where they were "bookmarked" in the book), students read different passages from the 346 total passages. Thus, the dataset is cross-classified, where students read 1 to 93 passages (e.g., some students only read one passage, some read 93 passages, but no one read all 346 passages; *Mdn* = 11.00; *SD* = 18.61) and passages were read by one to 13 students (e.g., some passages were only read by 1 student, some were read by 13 students, but no passage was read by all 56 students; *Mdn* = 2.00; *SD* = 2.55).

Looking within each Data Collection Procedure, the data from Data Collection A contained two to 19 turns per student (*Mdn* = 10.00; *SD* = 5.25), resulting in 282 reading turns available for analysis. Because this data collection period only lasted for five days, students were only able to read early passages in the book. The data from Data Collection B contained 1 to 93 turns per student (*Mdn* = 14.00; *SD* = 23.87), resulting in 682 reading turns available for analysis. Because this data collection lasted for a longer period, students read farther into the book on average (i.e., around Chapter 3) and showed greater variability in how far and how much they read. Table 1 summarizes the nesting structure of these turns across students within sites.

Results

Descriptive Statistics

Student-Level Characteristics

The top section of Table 1 reports descriptive information about the students in our sample. Whereas students' mean WCPM were similar across data collection procedures and sites (see also statistical models below), there was clear between-student variability in WCPM (*M* = 99.72; *SD* = 24.74) and oral reading accuracy (*M* = 0.93; *SD* = 0.06). Their WCPM means are well within the interquartile range of expected WCPM values for their age group based on norms (Hasbrouck & Tindal, 2017). Overall, students' correctly

Table 1

Descriptive Statistics (Frequencies, Means, and Standard Deviations) of Student-Level and Turn-Level Characteristics Across Sites

	Site								
Student-level statistics	A1	A2	A3	A4	A5	B1	B2	Overall mean/total	Missing
Demographics									
$N_{students}$	3	3	8	8	5	11	18	56	
Female	1	1	2	3	1	8	10	26 (46%)	
English L1 at Home	3	2	5	8	5	8	15	46 (82%)	9
Familiar with HP1	1	1	6	1	3	1	7	20 (36%)	16
Age (years)	8.51 (0.91)	8.88 (0.70)	10.05 (1.10)	8.87 (0.99)	11.18 (0.74)	9.40 (0.75)	9.80 (0.78)	9.61 (1.06)	7
Oral Reading									
n_{turns} per Student	14.33 (4.73)	12.00 (5.57)	10.00 (6.63)	7.88 (3.56)	12.00 (4.85)	9.36 (5.92)	32.17 (26.62)	17.21 (18.61)	
Mean WCPM	100.91 (10.22)	110.75 (27.41)	100.45 (22.14)	92.84 (26.66)	121.42 (35.66)	93.80 (21.94)	98.00 (24.38)	99.72 (24.74)	
Mean Accuracy	0.94 (0.00)	0.89 (0.07)	0.96 (0.02)	0.91 (0.07)	0.97 (0.03)	0.90 (0.08)	0.94 (0.03)	0.93 (0.06)	
Comprehension of *Harry Potter*									
N Questions Answered	10.00 (3.46)	12.00 (0.00)	9.00 (2.27)	7.88 (3.18)	9.60 (2.51)	30.82 (26.19)	82.11 (43.08)		
Comprehension Accuracy	0.89 (0.10)	0.75 (0.14)	0.62 (0.25)	0.71 (0.24)	0.92 (0.09)	0.80 (0.16)	0.67 (0.16)	0.73 (0.19)	
HP Interest[a]									
The Harry Potter book was boring.						1.60 (0.89)	2.12 (1.12)	1.97 (1.06)	11
Turn-Level Statistics									
N_{turns}	43	36	80	63	60	103	579	964	
Turn WCPM	99.08 (17.53)	114.32 (26.30)	108.27 (20.45)	98.69 (30.69)	120.87 (31.77)	98.27 (17.53)	106.05 (33.93)	105.84 (31.60)	
Turn Accuracy	0.94 (0.04)	0.90 (0.08)	0.97 (0.04)	0.93 (0.07)	0.98 (0.04)	0.92 (0.04)	0.95 (0.06)	0.95 (0.06)	

Note. Standard deviations and percentages are shown in parentheses. WCPM = words correct per minute; HP1 = Harry Potter and the Sorcerer's Stone.

[a] Means and standard deviations are reported for these variables as summary statistics but they are treated as categorical in mixed-effects models.

answered comprehension questions about the HP1 passages considerably above chance levels (M = 0.73; SD = 0.19). Collectively, these statistics indicate that the students analyzed here were generally not struggling readers and were generally paying attention to the story while reading aloud.

Regarding their familiarity with HP1, 20 (36%) of the students indicated that they have read the book or seen the movie before. As for their interest and engagement with HP1 (at least for Data Collection B), students generally disagreed that HP1 was boring (67%), suggesting that they liked the story overall.

As a surface validity check on our measures, we also inspected the correlations between these student-level characteristics (see Table 2). As expected, students' accuracy on HP1 reading comprehension questions were also strongly correlated with their mean WCPM (r = 0.53), consistent with the idea that more fluent readers have automated decoding skills that allow them to focus on understanding of the text, leading to higher comprehension accuracy (Fuchs et al., 2001; Kim et al., 2010, 2011). This result also indicates that students were reading the text with the goal of understanding it and not just to read it swiftly out loud.

It is also notable that the number of reading turns a student made is largely unrelated to their reading fluency, their accuracy for the reading comprehension questions, and their interest in the book: The number of turns made is only weakly and marginally related to WCPM (r = 0.23; p = .08), whereas it was not correlated with their accuracy on the comprehension questions and interest in the book. Thus, stronger and more fluent readers did not necessarily read more text or found the text more interesting than their weaker and less fluent peers.

Passage-Level Features

Table 3 shows the descriptive statistics for the text features obtained on the 346 passages in the dataset. TextEvaluator scores for the passages (treated as literary texts) ranged from 50 to 1150 (M = 574.60, SD = 182.84), which corresponds to below Grade 1 to Grade 12 range of text complexity and is consistent with the observed distribution of text complexity across this book (Beigman Klebanov et al., 2017). TTS reading rate estimates ranged from 85.06 to 189.16 words per minute (M = 154.77; SD = 13.16). This range is similar to variation in oral reading rates observed in adult readers (Loukina et al., 2018). Lastly, TextEvaluator scores are positively correlated with TTS estimates: more complex passages are estimated to be read faster based on TTS (r = 0.42, p < .001).

Table 2

Correlations Between Student Characteristics

Variable	Correlation 1	2	3	4	5
1. Age	1				
2. WCPM	0.22	1			
3. n_{turns}	0.06	0.23	1		
4. *Harry Potter* Accuracy	0.06	0.53**	0.10	1	
5. The *Harry Potter* book was boring.[a]	−0.44	0.08	−0.15	−0.37	1

Note. WCPM = word count per minute.
[a] Because this variable is on an ordinal scale, these coefficients refer to Spearman correlations; N = 18.
** $p \leq .01$.

Table 3

Descriptive Statistics for Text Feature Measures Obtained on HP1 Passages (N = 346)

Measure	*M*	*SD*	Min	Max	*r*	
TextEvaluator	574.60	182.84	50.00	1,150.00	1	
TTS	154.77	13.16	85.06	189.16	0.42***	1

Note. HP1 = Harry Potter and the Sorcerer's Stone; TTS = text-to-speech synthesis.
*** $p \leq .001$.

Turn-Level Characteristics

As part of the Data Preprocessing Procedure, we only considered and analyzed turns where WCPM was within 50 and 219 WCPM and accuracy was above 70% because turns that fail these criteria probably do not consist of bona fide, uninterrupted, and complete readings of a passage (see Appendix A for details). The bottom section of Table 1 summarizes the descriptive statistics for turns that met these criteria. For these turns, average WCPM were similar across data collection procedures and sites but there was substantial variability across turns, suggesting the presence of both text-based and student-based variability in WCPM. Overall average turn accuracy has little variability and is well above 90%, the level of accuracy below which would begin to suggest frustration from having difficulties in reading (Hasbrouck & Tindal, 2017). These results indicate that the reading turns used in this study reflect data in which students are reading at appropriate instructional levels, in that they are challenged but are not frustrated from reading the assigned material.

Planned Analyses: Mixed-Effects Models

We fit a series of mixed-effects models[4] to address our research questions. The first model we fit was a baseline model that controlled for data collection procedure as a fixed effect and included sites, students, and passages as random effects (where students were nested within sites; Model 1 in Table 4). The R syntax corresponding to this baseline model is given in Equation 2.

$$\begin{aligned}\texttt{wcpm} \sim \texttt{DataCollection} &+ (1|\texttt{site}) \\ &+ (1|\texttt{site:student}) \\ &+ (1|\texttt{passage}) \end{aligned} \tag{2}$$

With this model, we estimated the amount of student-based and text-based variability in our data to answer RQ1. This model showed that 9% of the variance in WCPM is attributable to passages, 64% to students/readers, and 27% to residual variance, which leaves no variance in WCPM attributable to site.[5] These

[4] Also known as multilevel or random-effects models. All mixed-effects models reported in this study were fit with restricted maximum likelihood using version 1.1-13 of the lme4 package (Bates et al., 2015) in R (R Core Team, 2017). Significance tests for parameter estimates were performed using version 3.0 of the lmerTest package (Kuznetsova et al., 2017).

[5] The site variance estimate of zero across all fitted models suggests that the variability among sites has already been captured by variation among students within sites. Nonetheless, we decided to retain the site term when possible to properly acknowledge the actual nesting structure of the current data. Note, however, that the size of the dataset precludes us from inferring that there is really no variation in WCPM across sites.

Table 4
Summary of Planned Mixed-Effects Models Fit to WCPM Data

Model	Model 1 Null model		Model 2 TextEvaluator		Model 3 TTS		Model 4 TextEvaluator + TTS		
Fixed effects	β	*t*	β	*t*	β	*t*		β	*t*
Intercept	103.86	21.09***	103.73	20.92***	102.19	21.00***		102.00	20.69***
Data Collection	−7.26	−1.07	−6.82	−1.00	−4.79	−0.72		−4.39	−0.65
TextEvaluator			−0.27	−0.28				−3.48	−3.14**
TTS					3.95	4.90***		5.81	5.91***
Random effects	*SD*	*r*	*SD*	*r*	*SD*	*r*	*SD*		*r*
Site									
Intercept	0.00		0.00		0.00		0.00		
Student									
Intercept	24.40		24.54	1	24.14	1	24.57	1	
TextEvaluator			2.57	0.20			4.02	0.05	1
TTS					1.63	0.74	3.03	0.33	−0.85
Passage									
Intercept	8.99		8.97		8.00		6.96		
Residual	15.93		15.78		15.83		15.67		

Note. Reference category for Data Collection is Data Collection A. TextEvaluator scores and TTS-based estimates were standardized across passages. WCPM = word count per minute; TTS = text-to-speech synthesis.
** $p \leq .01$. *** $p \leq .001$.

results show that some variability in WCPM is associated with differences in the passages, even if much more variability is associated with differences across students. We note that the estimated text-based variability is similar to the 1% to 10% reported in previous work that used standardized test passages to evaluate ORF (Barth et al., 2014; Christ & Ardoin, 2009; Kim et al., 2010; Poncy et al., 2005). There was no difference in WCPM attributable to the data collection procedure (t = 1.07, p = .29), and this held for all the models reported below.

This baseline model was followed by three models. To answer RQs 2 and 3, TextEvaluator scores and TTS-based estimates were each added as a sole predictor to the baseline model (Models 2 and 3) to determine how much text-based variability each feature accounted for by itself. To answer RQ4, both text features were added to the baseline model (Model 4) to determine how much variability they accounted for together but independent of each other. The features were mean-centered (i.e., z score–transformed) across passages. The models were specified with the maximal random effects structure (Barr et al., 2013): Random slopes and associated random-effect correlations (i.e., intercept-slope and slope-slope correlations) were all specified for each additional predictor to maintain nominal false positive rates and to account for individual differences in the predictors' effects. Table 4 summarizes these sets of models, and Table 5 shows the amount and percentage of text-based WCPM variance explained by each predictor as they were added to the baseline model.

Surprisingly, TextEvaluator scores did not significantly predict WCPM (Model 2; $\beta = -0.27$, $t = -0.28$, p = .78) nor accounted for any text-based WCPM variability by itself, contrary to studies based on standardized test passages that report significant relationships between text complexity and ORF (Barth et al., 2014; Betts et al., 2009; Hintze et al., 1998). On the other hand, TTS-based estimates significantly predicted WCPM by itself and accounted for 21% of text-based WCPM variability: passages for which TTS estimated faster reading rates tended to be read faster by students as well (model 3: $\beta = 3.95$, $t = 4.90$, $p < .001$).[6]

Adding TextEvaluator scores to a model that already contains TTS-based estimates (Model 4) resulted in both features being significant predictors of WCPM. As before, passages for which TTS estimated faster reading rates were still associated with higher WCPM ($\beta = 5.81$, $t = 5.91$, $p < .001$), indicating that production-related effects still influenced WCPM after controlling for text complexity, as they did when TTS-based estimates were used as the sole predictor of WCPM. At the same time, passages with higher TextEvaluator scores were now associated with lower WCPM ($\beta = -3.48$, $t = -3.14$, p = .003), which is consistent with the expectation that more complex passages prompt slower oral reading from children. However, this effect is significant only after adjusting for production-related effects.[7] This model explained 40% of text-based WCPM variability, which is 19% more than what is explained by TTS-based estimates alone.[8] This final model

[6] We consider the possibility that this result is due to idiosyncrasies of the TTS model we used. That is, although the TTS model incorporates a comprehensive set of production constraints, it is still an imprecise model because it encounters issues in capturing particularities in human speech, so the result may not be credible. To address this concern, we computed the reading rates of Jim Dale, the actor who narrated the audiobook, for each passage using the approach described in Loukina et al. (2018) and replaced the TTS-based estimates with these measurements. Although Jim Dale's reading rates are faster and more variable than the TTS-based estimates (M = 159.00; SD = 15.40), his reading rates are still strongly correlated with the TTS-based estimates (r = 0.74). We found the same results using Jim Dale's reading rates: On their own, his reading rates significantly predicted WCPM ($\beta = 3.85$, $t = 4.97$, $p < .001$) and explained 17% of text-based WCPM variability. This suggests that the results are not attributable to idiosyncrasies in the TTS model used in this study.

[7] In the multiple regression literature, such effects—where a predictor has no marginal effect on the outcome but has an effect on the outcome conditional on the addition of another predictor—are referred to as "suppressor variable effects" (Cohen et al., 2013).

[8] We found the same results when we replaced the TTS-based estimates with Jim Dale's reading rates.

Table 5

Percentage of Text-Based WCPM Variance Accounted for by Text Features

Model	Remaining Unexplained Text-Based Variance	% Text-Based Variance Accounted by Model
Model 1: Baseline model	80.76	
Model 2: TextEvaluator	80.47	0.36
Model 3: TTS	64.07	20.67
Model 4: TextEvaluator + TTS	**48.50**	**39.95**

Note. WCPM = word count per minute; TTS = text-to-speech synthesis. Model 4 (in boldface) accounts for most text-based variance among the models.

illustrates that production-related effects on ORF are separable and qualitatively different from text complexity effects.

Follow-Up Analyses

In this section, we check the robustness of results from our planned analyses and explore potential explanations for some of them. Namely, we first evaluate whether production-related effects depend on student characteristics. Given that production-related effects explain a large amount of text-based ORF variability, it is important to determine whether these effects are only driven by specific sets of students. Second, we explore why we failed to observe a text complexity effect for HP1 passages. It is surprising that TextEvaluator scores did not predict WCPM on their own, given that text complexity is purported to be an important and independent predictor of ORF, albeit in standardized test passages (Barth et al., 2014; Betts et al., 2009; Hintze et al., 1998). We explore a hypothesis that TextEvaluator's effect was only revealed when we accounted for TTS estimates because marginal text complexity effects reflect a combination of text-complexity-related and production-related effects that may go together or counteract each other. Overall, these follow-up analyses support the results of the planned analyses and show that production-related effects are a robust source of text-based ORF variability.

Exploring Whether Production-Related Effects Are Only Driven by Specific Students

A potential concern about production-related effects being an independent source of text-based ORF variability is that these effects are only present in specific kinds of readers. For example, production-related effects might only be present for students who are already familiar with the text or who understand the text better. The event segmentation theory (Zacks et al., 2009; Zacks & Swallow, 2007), the event indexing model (Zwaan et al., 1995), and the structure building framework (Gernsbacher, 1997) converge on the expectation that students' familiarity and understanding of the book would facilitate the development of their discourse model. Their familiarity provides a head start on initializing the mental representations of characters and settings and their understanding of the story strengthens these representations. This facilitation and strengthening prevents students from struggling to read story-specific elements (e.g., in HP1: reading new character names like McGonagall and Dumbledore, names of spells) and pushes their reading to be more aligned with what is expected based on the timing models in TTS systems. It might also go the other way, where these effects might be absent for students who are familiar with the text or understand the text better because these readers use their familiarity and understanding to infuse their reading with extralinguistic information (e.g., exaggerations based on character persona, events) so that their reading would not be consistent with what is expected based on the timing models in TTS systems. Either way, if production-related effects are only present in specific kinds of readers, we should expect the effect of TTS estimates to be modulated by various student characteristics one way or another.

To explore this possibility, we fit a follow-up model to Model 4 (TextEvaluator + TTS) with several additional terms in the model. We added students' prior familiarity with HP1, accuracy for HP1 questions, age, and the interactions of these factors with TextEvaluator scores and TTS estimates to examine whether our focal effects are modulated by student characteristics. We also added students' turn index within a reading session and its interaction with the data collection procedure as measures of how far students are into a reading session within each data collection procedure. These measures account for the effects of student fatigue within a session and possible differences in the effect of fatigue on WCPM across data collection procedures. We also added students' gender, the number of valid turns they made, and their interest in the HP1 story as additional student characteristics. Table B1 in Appendix B summarizes this model (Model 5).

Whereas Model 5 explains 20% of student-level WCPM variability, it does not explain any additional text-based WCPM variability on top of Model 4. As in our planned models, TextEvaluator scores and TTS estimates still significantly predicted WCPM. Of the added terms, only accuracy on the HP1 comprehension questions significantly predicted WCPM: students who were more accurate had higher WCPM ($\beta = 0.53$, $t = 2.82$, $p = .007$). This effect was qualified by a significant interaction with the effect of TTS estimates: production-related effects were stronger for students who were more accurate on the HP1 comprehension questions ($\beta = 0.11$, $t = 2.08$, $p = .04$). The significant effect of comprehension accuracy on WCPM (see also Table 2) suggests that the more accurate students were more skilled readers, and having certain reading skills in place may have made their oral reading more aligned with expected prosodic patterns based on production constraints. On the same note, the oral reading of less accurate students likely contained more hesitations, pauses, repetitions and other disfluencies that would not be predicted by the TTS timing model, resulting in weaker production-related effects for these students. Because this analysis is exploratory, our findings would need to be confirmed and further tested in subsequent studies that explicitly focus on individual differences in production-related effects. But collectively, these findings suggest that production-related effects are robust and that their strength might differ across different types of students.

Exploring the Failure to Observe a Text Complexity Effect: Text Features Have Complexity-Related Effects and Production-Related Effects

In this section, we explore why we failed to observe a marginal text complexity effect, given that text complexity is purported to be an important and independent predictor of ORF, albeit in standardized test passages (Barth et al., 2014; Betts et al., 2009; Hintze et al., 1998). We begin by noting that most text complexity metrics, including TextEvaluator, are based on a combination of passage features. Some of these features not only affect text complexity but simultaneously also generate production constraints, and both text complexity and production constraints affect the reading rate for a given text. Here, we explore the hypothesis that marginal effects of text complexity on ORF (or lack thereof) may reflect the composite effect of these features on oral reading. To make this exploratory analysis manageable, we consider Flesch-Kincaid (FK) as a case example which only has two features.

Earlier, we described how FK's effect on ORF has been attributed to how its component features—average word length and average sentence length—can affect the complexity of texts. Simultaneously, however, word and sentence length impose prosodic and articulatory constraints that are unrelated to text complexity. Longer words simply take more time to utter, implying that fewer words can be uttered within one minute. As a result, all else being equal, passages with longer words will be read with a lower per-word reading rate as measured by WCPM than passages with shorter words.[9]

The effect of sentence length on reading rate is more complex. Syllables at the end of a sentence are pronounced more slowly—an effect known as sentence-final lengthening (see, for example, White (2014) for a comprehensive review). In longer sentences (that is, a sentence with more words), this "lengthening penalty" will be distributed over a larger number of words, leading to a shorter average duration of each word in a longer sentence than in a shorter one, all else being equal. As a result, longer words are likely to slow reading rate by both increasing text complexity and reading time, whereas longer sentences could have counteracting effects on reading rate: Longer sentences make passages more complex thereby slowing reading rate, but this effect could be offset to some degree by prosodic constraints that make longer sentences faster to utter (per word) than shorter sentences. Furthermore, if the same long sentences contained long words, word and sentence length would interact (likely in a less straightforward fashion) to affect reading rate. Consequently, the observed marginal effect of FK scores on ORF in each given set of passages would depend on a composite of word and sentence length's effects on ORF because of the constraints they impose on both text complexity *and* production.

We hypothesize that breaking down text complexity metrics in this way helps explain both the absence of a marginal text complexity effect in the current study and its presence in previous work. In particular, we argue that the structure of the HP1 passages in the current study substantially differs from the structure of the standardized test passages used in previous studies. Consequently, the interplay between the text-complexity-related and production-related effects of the passages' features on ORF results in different marginal text complexity effects in these two sets of passages.

In turn, we also hypothesize that a marginal effect of text complexity was observed in previous work that used standardized test passages because the aspects of those passages that made them more complex also happen to be those that make them be uttered slower per word. That is, within those passages, the text-complexity–related and production-related effects on ORF operated in the same direction, resulting in the intuitive finding that higher text complexity scores, by themselves, predict lower ORF.

A Case Study: Comparing HP1 and DIBELS Passages. To test these hypotheses, we obtained a sample of 160 passages from the *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS; Good & Kaminski, 2002), a standard ORF assessment,[10] and compared them to the HP1 passages in the current study. We obtained TextEvaluator scores and TTS reading rate estimates for each passage as we did for the HP1 passages.

Next, we compared how FK's features—average word and sentence length—contribute to variability in TTS-based estimates between the two groups of passages (see Table 6). We standardized word and sentence length within each group of passages to assess each component's relative importance in predicting TTS-based estimates.

The models indicate that average word and sentence length are significant predictors of TTS-based estimates and explain 44% and 62% of its variance in HP1 and DIBELS passages, respectively. Moreover, the directions of their effects are identical between the two groups of passages: Longer words slow down the oral reading of passages, whereas longer sentences contribute to faster expected reading rates, as explained above.

However, the components' relative contributions in explaining TTS-based estimates differ between the two groups of passages: whereas sentence length drives reading rate estimates for HP1 passages more than word length ($\beta = 8.72$ vs. -1.85), word length drives reading rate estimates for DIBELS passages more than sentence length ($\beta = -9.77$ vs. 3.84). Thus, we observed that the correlation between TTS-based estimates and FK scores is negative in DIBELS passages ($r = -0.37$) because the dominant feature that characterizes them—average word length—makes passages more complex and slower to read per word when it increases. In contrast, this correlation for HP1 passages is positive ($r = 0.57$), because the dominant feature that characterizes them—average sentence length—makes passages more complex but faster to read per word when it increases.

Why Are the Results for HP1 and DIBELS Passages Different? Why is the feature that drives estimated reading rates different between the two sets of passages? Average word length drives the reading rates of DIBELS passages because word length is a stronger structural feature of DIBELS passages than of

[9] The relationship between the number of syllables and the word duration is not completely linear.

[10] Eighth edition: https://dibels.uoregon.edu/assessment/index/material.

Table 6

Standardized Regression Coefficients in Predicting TTS-Based Estimates for Harry Potter and DIBELS Passages

	Harry Potter (N = 346) $r^2 = 0.44$				DIBELS (N = 160) $r^2 = 0.62$			
Predictor	β	*SE*	*t*	*p*	β	*SE*	*t*	*p*
Average word length	−1.85	0.54	−3.46	.001	−9.77	0.60	−16.30	<.001
Average sentence length	8.72	0.54	16.27	<.001	3.84	0.60	6.40	<.001

Note. TTS = text-to-speech synthesis; DIBELS = Dynamic Indicators of Basic Early Literacy Skills.

HP1 passages. First, we observe that average word and sentence length are more strongly related in DIBELS passages (r = 0.43) than they are in HP1 passages (r = 0.11). That is, in DIBELS, longer sentences tend to be made of longer words, whereas in HP1, words with different lengths are roughly evenly distributed across sentences of different lengths. Second, average word length is also more strongly correlated with TextEvaluator scores in DIBELS passages (r = 0.68) than they are in HP1 passages (r = 0.16). This suggests that average word length characterizes the difficulty of DIBELS passages to a greater extent than it does for HP1 passages.

On the other hand, average sentence length drives the reading rates of HP1 passages because the control we imposed on the total word count of HP1 passages to manage the length of a reader's turn[11] expanded the scope of prosodic effects captured by sentence length. When the word count of a passage is controlled, the number of sentences in the passage largely determines sentence length: When word count is constant, fewer sentences would typically imply longer sentences. Fewer sentences would elicit fewer sentence-final lengthenings and pauses per passage, resulting in faster per-word reading rate, all else being equal. Because fewer sentences imply longer sentences, prosodic effects from sentence count can also be captured by sentence length.

Thus, of the two features we consider here, the feature that dominates the production effects on reading rates for HP1 passages is average sentence length, which also happens to be the feature whose text-complexity-related and production-related effects are *not* aligned with predictions of the overall text complexity effect: longer sentences are more complex but they are faster to read per word due to their prosodic constraints (see Table 6). Because average word and sentence length have opposite production-related effects on the reading rate of these passages, using a text complexity measure that combines these two features additively (as done explicitly in the FK formula) would reveal no marginal text complexity effect on ORF. In fact, replacing TextEvaluator with FK as our metric of text complexity in the mixed-effects models reported earlier produced the same conclusions.

We also arrive at the same conclusions when replacing the overall TextEvaluator score in our models with its eight component dimensions. Of TextEvaluator's eight dimensions, passages' word unfamiliarity and syntactic complexity were the only significant predictors of WCPM. Passages higher in word unfamiliarity—which is correlated with average word length (r = 0.23)—are predicted to be read slower, whereas passages higher in syntactic complexity—which is correlated with average sentence length (r = 0.94)—are predicted to be read *faster*. Importantly, the effect of syntactic complexity is fully captured by TTS estimates when TTS estimates are entered into the model. This model with TTS estimates and TextEvaluator's eight dimensions explains 42% of text-based WCPM variability, which is similar to the 40% explained by our planned model that has TTS estimates and the overall TextEvaluator scores (Model 4; Table 5). These results provide further evidence that we failed to observe a marginal text complexity effect in the HP1 passages because the dimensions that predict WCPM for these passages—word unfamiliarity and syntactic complexity—have opposite text-complexity-related and production-related effects.

Summary. Marginal effects of text complexity metrics on ORF have been interpreted to exclusively indicate how text complexity affects ORF. However, our results suggest that previously observed marginal text complexity effects likely reflect a composite of text-complexity-related and production-related effects because these metrics are built from components that can have simultaneous effects on both text complexity and production.

Consequently, when a marginal text complexity effect is observed without controlling for prosodic constraints, it could be attributable to readers needing longer time to process complex texts (i.e., the traditional, intuitive explanation); or that more complex passages happened to be constructed in a way that makes them longer to utter, independent of the passages' text complexity; or both, where one of them could be driving the effect more than the other. On the other hand, when a text complexity effect is not observed (as in the current study), it could ostensibly suggest that text complexity has no effect on ORF. Our study showed that text complexity effects could still be present, but to reveal them, production constraints imposed by the passages may have to be accounted for.

Discussion

Using data collected through an app that involves interactive oral reading, we estimated and identified sources of text-based variability in children's ORF when reading passages from a novel. We found the following answers to our research questions:

1. Using linear mixed-effects models, we found a small but appreciable amount of text-based variability (9%) in ORF.

[11] Word count was only approximately controlled because we respected paragraph boundaries when segmenting the text into turns, so we would occasionally have turns with long paragraphs, where the number of words is substantially beyond the average for a turn. See the Materials: Reading App section.

2. Contrary to what is typically expected, we found that a comprehensive text complexity metric did not explain text-based ORF variability on its own.

3. In contrast, production-related effects—which we operationalized here as the timing and prosodic factors captured by reading rate estimates from an automated TTS system—explained text-based ORF variability on their own (Model 3: 21% of 9% ≈ 1.8%)

4. Only after controlling for production-related effects do we see a significant effect of text complexity on ORF. The model that has both TextEvaluator scores and TTS-based estimates in the model explained about half of the text-based variability in our data (Model 4: 40% of 9% ≈ 3.6%).

We also tested the robustness of these findings by fitting a follow-up model that accounted for various student characteristics. We found that production-related effects still predicted ORF and that the magnitude of these effects might differ across groups of students. These findings indicate that production-related effects are an independent and robust source of text-based ORF variability.

The robustness of production-related effects in ORF should not be surprising. Before processing effects from text complexity and related features can even affect ORF, the makeup of texts already imposes constraints on how fast they can be fluently read from the outset. These constraints are widely recognized in the speech production and phonetic literatures, which treat *oral* reading as a type of utterance that is subject to the same production constraints as other types of utterances (Hirschberg, 2002; White, 2014). For example, passages with more clauses and sentences would generally prompt more prosodic pauses in fluent oral reading because the presence of more clausal boundaries would license more “valid” pausing locations in the passage. In turn, these boundaries already contribute to how the passage is expected to be read aloud. Processing difficulty and comprehension demands would then interact with these production constraints to result in the oral reading rate that we ultimately observe for a passage.

What was surprising, however, was that production-related effects seem stronger for students who understood the story better. This result can be supported by accounts of discourse processing (e.g., Gernsbacher, 1997; Zacks & Swallow, 2007; Zwaan et al., 1995) which predict that greater understanding of the text reinforces the reader’s discourse model and facilitates oral reading for the text. Given that the students who understood the story more also seemed to be more skilled readers, this result is also consistent with evidence of developmental differences in the use of prosody when reading aloud, where more skilled developing readers have exhibited prosody that is more similar to expert/proficient readers’ than their less skilled peers (e.g., Ardoin et al., 2013; Kim & Wagner, 2015; Miller & Schwanenflugel, 2006). However, we note that our measure of comprehension—accuracy on fact-based questions about the HP1 passages—tapped shallower levels of processing (i.e., mostly at the “textbase” level, with some elements of the “situation-model” level, in the multilevel framework of discourse comprehension in Graesser et al., 2011). Thus, although we have some idea about the students’ comprehension, we are uncertain about how well they actually understood the larger context of the novel. Because our results are preliminary, the link between production-related effects and reading comprehension can be clarified by using measures that tap into deeper levels of comprehension and discourse processing. Clarifying the nature of this link is important because it has implications for measuring the ORF of students who differ in comprehension skills using texts with different production constraints.

Another surprising finding is that the amount of text-based ORF variability we observed in HP1 passages was *similar* to what has been observed in standardized test passages (1–10%; Barth et al., 2014; Christ & Ardoin, 2009; Kim et al., 2010; Poncy et al., 2005). To our knowledge, the range of text-based ORF variability in novels like HP1 is yet to be estimated in the literature. Despite this, an expected limitation of using such texts to study ORF is that they would introduce more variability to students’ oral reading than standardized test passages and other controlled texts. This expectation did not hold for HP1 in this study, but additional studies on HP1 and other fiction books and novels would more clearly reveal the range of text-based variability we could expect from such texts. These studies would also confirm whether concerns about fiction books and novels introducing more text-based ORF variability than standardized test passages are warranted.

The difference we observed from standardized test passages, however, is in *how* text features explained ORF variability. Whereas overall text complexity metrics tend to explain ORF variability on their own in standardized test passages (Ardoin et al., 2005; Barth et al., 2014; Compton et al., 2004; Francis et al., 2008), we found that there was no marginal text complexity effect for HP1 passages and that the effect of text complexity was only revealed when production-related effects were included in the model. We attributed this finding to text complexity metrics consisting of features that have both complexity-related *and production-related effects* on oral reading. We demonstrated that a sample of standardized test passages are structured so that features that drive their complexity also make them slower to read, whereas our passages from HP1 were structured so that the features that drive their complexity make them *faster* to read. Although these results were exploratory and warrant confirming in future work, these findings highlight how texts that exhibit wide variation in their features can offer important insight about text-related effects in ORF.

Theoretical, Educational, and Clinical Implications

The effects of text features on ORF are explained by current theories primarily in terms of processing difficulty or comprehension demands. For example, in Zacks et al. (2009), where event segmentation theory is applied to the oral reading of narrative texts, they claimed that passages containing more event boundaries are expected to be read more slowly because those boundaries mark where situational features (e.g., characters, objects, goals) change and the discourse model is updated. However, we propose that the text making up these event boundaries and situational features have production constraints whose effects on ORF are independent of discourse processing. In fact, in the same Zacks et al. (2009) study, the passages with more event boundaries and changes in situational features also tended to have more syllables and longer clause durations (see their Table 4). Thus, the resulting reading times for these passages may have been attributable to both the production constraints from syllable count and clause duration *and* the discourse effects due to changes in the situational features. Interestingly, when Zacks et al. (2009) included both syllable count and situational features in their analysis, they found that the effect of situational features flipped signs (see their footnote 1), indicating a similar suppression effect to what we observed in the current study for text complexity. Whereas they interpreted this suppression effect as an artifact and decided to remove syllable count from their models, we argue that this is evidence for the need to account for production-related effects when developing models of ORF. To estimate the comprehension-related effects of text features (e.g., text complexity, effects discourse processing effects) independent of their production-related effects and vice versa, both effects would need to be accounted for when predicting ORF.

This theoretical implication has downstream educational and clinical implications. Interpreting differences in ORF between students may not be straightforward even when the texts they read are comparable in features like text complexity. When production-related effects are unaccounted for, differences in WCPM may not directly reflect differences in ORF because there are differences in the production constraints of the texts that students happened to read (see also Francis et al., 2008). Moreover, because the strength of production-related effects could vary across readers, this consequence can have disproportionate effects on different groups of students. For example, this consequence would disproportionately impact students with clinical or subclinical speech motor deficits because their oral reading is especially susceptible to texts' production constraints, and these constraints (e.g., number of syllables, intonation types) can affect clinical groups differently (see, e.g., Kuo & Tjaden, 2016; Patel et al., 2013).

How can production-related effects be accounted for? It would not be practical to create texts that are comparable in their production constraints because there are many such constraints and they interact in complex ways (Hirschberg, 2002; White, 2014). We propose that an alternative is to include reading rate estimates from TTS as a covariate when analyzing differences in ORF. In doing so, differences in ORF would reflect differences beyond the expected reading rates for the passages due to their production constraints.

This alternative is important when measuring the ORF of readers with subclinical/clinical speech motor deficits and comparing them against typically developing readers. When comparing subclinical/clinical samples against typical control samples in experiments or intervention studies, it is common to match the samples in age and ability to ensure that the effect of a manipulation or intervention is not modulated by preexisting differences between the groups. Matching the texts on their expected reading rates across the groups achieves a similar effect, but this process could be as challenging as creating texts that are comparable in production constraints. Entering TTS reading rate estimates as a covariate would serve as a simpler alternative, where the differences between the groups are statistically adjusted for the production constraints in the texts they read. This adjustment would help ensure that the ORF differences between the groups are evaluated independent of the samples' age, ability, and texts read.

Limitations

Our main goal for this study was to test the importance of production-related effects as a source of text-based ORF variability, but we fully acknowledge that there are other important factors that contribute to text-based ORF variability that we did not account for, such as text features related to discourse processing and reading comprehension. For example, students read different HP1 passages aloud and these passages may have differed in the type and amount of discourse features they had. The presence of more events, conversations, plot transitions, and animated characters in a passage is expected to affect oral reading because these features all affect how the reader's discourse model develops and the state of a reader's discourse model affects how quickly information is integrated during reading (Gernsbacher, 1997; Zacks et al., 2009). Given that production-related effects appear to explain a substantial amount of text-based ORF variability on their own (20%), a potential concern about not accounting for discourse processing features is that their effects are conflated with production-related effects.

We note that this conflating is unlikely because production constraints come from fundamental properties of speech, which are in essence unrelated to the meaning of the texts and to the role of particular passages in the bigger narrative. This independence is also reflected in how we operationalized production constraints in this study: As we described, the timing model in the TTS system we used only considers local prosodic effects (segment quality, lexical stress, prosodic phrasing, emphasis based on syntactic structure, pauses between sentences). Thus, the reading rate estimates we obtained from TTS are independent of discourse-related effects. In fact, because modern TTS systems only account for local effects, TTS researchers have tried to modify TTS systems for storytelling contexts to generate digital speech that accounts for emotions, characters, and other discourse-related features (Delmonte & Tripodi, 2015; Doukhan et al., 2011; Ramli et al., 2016; Theune et al., 2006). Nevertheless, investigating discourse-related features as a source of text-based ORF variability in fiction books like HP1 is a fruitful direction for future work. In addition to estimating how much variability these features account for in fiction books, it would be important to examine how production-related features interact with discourse processing-/comprehension-related features to affect the ORF observed for a passage.

Another goal for this study was to analyze text-based ORF variability from the casual reading of a popular fiction book for an extended period of time. Although analyzing text-based ORF variability in this ecologically relevant context is a novel aspect of this study, this context introduced some methodological challenges.

First, because we were interested in students' natural reading behavior, we did not control how students approached the task beyond instructing them to read aloud as they normally would. Consequently, students exhibited an array of on- and off-task behaviors, which led to the exclusion of some participants and many turns (see Appendix A). Although the excluded turns and students were largely similar to those retained for analysis (Tables A1 and A2), we acknowledge that the excluded students' mean accuracy for the HP1 comprehension questions was only 50%, which is substantially lower than the accuracy of the retained students (73%). A potential concern for this exclusion is that we excluded poor readers and only retained highly skilled readers. We note that it is possible for the excluded students to have been poor readers, but given that these students were excluded because all of their turns were unlikely to constitute *bona fide* oral reading, it is also possible that these students were simply not motivated to do the task.

Unfortunately, we cannot distinguish whether students were excluded based on one or both of these reasons. However, we do not think that we only retained highly skilled readers given the variability in their ORF and reading comprehension (see Table 1) and how they compare to national norms (Hasbrouck & Tindal, 2017).

Second, because our study was done outside a formal instructional setting, we were limited in the student characteristics we could obtain: We do not have information about whether students have known speech or language impairments. Because we anticipate that student characteristics, especially speech motor and language clinical diagnoses, would modulate production-related effects, it would be important to conduct similar studies in a setting where such information can be collected and analyzed. In sum, to overcome these methodological challenges, future work in a similar naturalistic study context would benefit from situating the reading sessions in a more structured and instructional setting (e.g., in a school's reading curriculum), where a wider range of student characteristics can be examined.

Conclusions

A principal challenge in assessing readers' ORF is accounting for systematic sources of its variability. The characteristics of passages given to readers are considered one such source, and in particular, text complexity has enjoyed the role of being the most prominent of these characteristics. However, as text complexity only captures some text-based variability, we considered other characteristics to account for more of this variability.

Using oral reading data from upper elementary school children reading passages from a popular novel, we found that articulatory and prosodic constraints related to the oral production of texts should be considered alongside text complexity as another text characteristic to account for when measuring ORF. These production-related features, which we operationalized by using timing models built into TTS systems, explained a substantial amount of text-based ORF variability in the current study (20%). The fundamental role of production-related features in constraining timing patterns has been widely acknowledged in the speech production literature (Hirschberg, 2002; White, 2014), but until the current study these features had yet to receive attention in the ORF literature as constraining reading rate——and therefore as producing systematic text-based ORF variability——independently of text complexity and other factors related to text processing. This finding advances our understanding of factors that could impact ORF measurement. We also note that TTS-based estimates can be readily obtained from current computer operating systems, making it easy for other researchers to expand on our findings.

The current study also illustrates that passages could be read slowly not necessarily or not only because they have higher text complexity but also because they happen to be structured so that they are slower to utter per word regardless of how complex they are. Previous work on ORF has focused on standardized test passages such as DIBELS, where passages with higher text complexity happened to have characteristics that make them slower to read per word. Results on these passages have been consistent with the expected text complexity effect (e.g., Ardoin et al., 2005; Barth et al., 2014; Francis et al., 2008). In contrast, for a set of passages from a novel (HP1) that were controlled for word count, passages with higher text complexity tended to be those that were faster to utter per word, and we suggested that the structure of these passages makes these two factors counteract each other, which may be why we did not observe a marginal text complexity effect on ORF for these passages.

Our results suggest that it is important to consider including a wider range of reading materials beyond standardized test passages when studying the effects of text features on ORF because texts created for different purposes could systematically differ in their structure. In turn, these differences in structure could systematically affect their oral reading. Using a variety of reading materials would cover a wider range of possible variations in text features, thereby allowing a more comprehensive investigation into text-based ORF variability.

References

Ardoin, S. P., Morena, L. S., Binder, K. S., & Foster, T. E. (2013). Examining the impact of feedback and repeated readings on oral reading fluency: Let's not forget prosody. *School Psychology Quarterly*, *28*(4), 391–404. https://doi.org/10.1037/spq0000027

Ardoin, S. P., Suldo, S. M., Witt, J., Aldrich, S., & McDonald, E. (2005). Accuracy of readability estimates' predictions of CBM performance. *School Psychology Quarterly*, *20*(1), 1–22. https://doi.org/10.1521/scpq.20.1.1.64193

Bailly, G., & Gouvernayre, C. (2012). Pauses and respiratory markers of the structure of book reading. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2012* (Vol. 3, pp. 2215–2218).

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barth, A. E., Tolar, T. D., Fletcher, J. M., & Francis, D. (2014). The effects of student and text characteristics on the oral reading fluency of middle-grade students. *Journal of Educational Psychology*, *106*(1), 162–180. doi: https://doi.org/10.1037/a0033826

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beigman Klebanov, B., Loukina, A., Madnani, N., Sabatini, J., & Lentini, J. (2019). Would you? Could you? On a tablet? Analytics of children's eBook reading. *Proceedings of the 9th International Learning Analytics Knowledge Conference* (pp. 106–110). ACM Press. https://doi.org/10.1145/3303772.3303833

Beigman Klebanov, B., Loukina, A., Sabatini, J., & O'Reilly, T. (2017). Continuous fluency tracking and the challenges of varying text complexity. *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 22–32). Association for Computational Linguistics. https://doi.org/10.18653/v1/w17-5003

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111. https://doi.org/10.1016/j.jml.2008.06.003

Bernstein, J., Cheng, J., Balogh, J., & Rosenfeld, E. (2017). Studies of a self-administered oral reading assessment. In O. Engwall, J. Lopes, & I. Leite (Eds.), *Proceedings of the 7th ISCA Workshop on Speech and Language Technology in Education* (pp. 180–184). KTH Royal Institute of Technology.

Betts, J., Pickart, M., & Heistad, D. (2009). An investigation of the psychometric evidence of CBM-R passage equivalence: Utility of readability statistics and equating for alternate forms. *Journal of School Psychology*, *47*(1), 1–17. https://doi.org/10.1016/j.jsp.2008.09.001

Biancarosa, G., Kennedy, P. C., Park, S., Otterstedt, J., Gearin, B., & Yoon, H. (2019). 8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS®): *Administration and Scoring Guide* (Tech. Rep.). University of Oregon. https://dibels.uoregon.edu/docs/materials/d8/dibels_8_admin_and_scoring_guide_09_2019.pdf

Burrows, T., Jackson, P., Knill, K., & Sityaev, D. (2005). Combining models of prosodic phrasing and pausing. *Proceedings of the 9th European Conference on Speech Communication and Technology* (pp. 1829–1832). https://www.isca-speech.org/archive/archive_papers/interspeech_2005/i05_1829.pdf

Capes, T., Coles, P., Conkie, A., Golipour, L., Hadjitarkhani, A., Hu, Q., Huddleston, N., Hunt, M., Li, J., Neeracher, M., Prahallad, K., Raitio, T., Rasipuram, R., Townsend, G., Williamson, B., Winarsky, D., Wu, Z., & Zhang, H. (2017). Siri on-device deep learning-guided unit selection text-to-speech system. *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech 2017* (pp. 4011–4015). https://isca-speech.org/archive/Interspeech_2017/pdfs/1798.PDF

Christ, T. J., & Ardoin, S. P. (2009). Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School Psychology*, *47*(1), 55–75. https://doi.org/10.1016/j.jsp.2008.09.004

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Collins-Thompson, K., & Callan, J. (2004). Information retrieval for language tutoring: An overview of the REAP project. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 544–545). https://doi.org/10.1145/1008992.1009112

Compton, D. L., Appleton, A. C., & Hosp, M. K. (2004). Exploring the relationship between text-leveling systems and reading accuracy and fluency in second-grade students who are average and poor decoders. *Learning Disabilities Research and Practice*, *19*(3), 176–184. https://doi.org/10.1111/j.1540-5826.2004.00102.x

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Current results. *Journal of the Acoustical Society of America*, *83*(4), 1553–1573. https://doi.org/10.1121/1.395911

Daane, M. C., Campbell, J. R., Grigg, W. S., Goodman, M. J., & Oranje, A. (2005). *Fourth-Grade Students Reading Aloud: NAEP 2002 Special Study of Oral Reading. The Nation's Report Card (NCES 2006-469)* (Tech. Rep.). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. https://files.eric.ed.gov/fulltext/ED488962.pdf

Delmonte, R., & Tripodi, R. (2015). Semantics and discourse processing for expressive TTS. *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-Level Semantics* (pp. 32–43). https://www.aclweb.org/anthology/W15-2704.pdf

Doukhan, D., Rilliard, A., Rosset, S., Adda-Decker, M., & D'Alessandro, C. (2011). Prosodic analysis of a corpus of tales. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2011* (pp. 3129–3132). https://www.isca-speech.org/archive/archive_papers/interspeech_2011/i11_3129.pdf

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology*, *46*(3), 315–342. https://doi.org/10.1016/j.jsp.2007.06.003

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, *5*(3), 239–256. https://doi.org/10.1207/S1532799XSSR0503_3

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education*, *9*(2), 20–28., https://doi.org/10.1177/074193258800900206

Gernsbacher, M. A. (1997). Two decades of structure building. *Discourse Processes*, *23*(3), 265–304. https://doi.org/10.1080/01638539709544994

Good, R. H., & Kaminski, R. A. (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed). Institute for the Development of Educational Achievement. https://dibels.uoregon.edu/docs/materials/admin_and_scoring_6th_ed.pdf

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223–234. https://doi.org/10.3102/0013189X11413260

Hasbrouck, J., & Tindal, G. (2017). *An update to compiled ORF norms* (Tech. Rep.). Behavioral Research and Teaching, University of Oregon. https://www.brtprojects.org/wp-content/uploads/2017/10/TechRpt_1702ORFNorms_Fini.pdf

Hintze, J. M., Daly, E. J., & Shapiro, E. S. (1998). An investigation of the effects of passage difficulty level on outcomes of oral reading fluency progress monitoring. *School Psychology Review*, *27*(3), 433–455. https://doi.org/10.1080/02796015.1998.12085928

Hirschberg, J. (2002). Communication and prosody: Functional aspects of prosody. *Speech Communication*, *36*(1–2), 31–43. https://doi.org/10.1016/S0167-6393(01)00024-3

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America*, *25*(1), 105–113. https://doi.org/10.1121/1.1906982

Kim, Y.-S. G., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology*, *102*(3), 652–667. https://doi.org/10.1037/a0019643

Kim, Y.-S. G., & Wagner, R. K.. (2015). Text (oral) reading fluency as a construct in reading development: An investigation of its mediating role for children from Grades 1 to 4. *Scientific Studies of Reading*, *19*(3), 224–242. https://doi.org/10.1080/10888438.2015.1007375

Kim, Y.-S. G., Wagner, R. K., & Foster, E. (2011). Relations among oral reading fluency, silent reading fluency, and reading comprehension: A latent variable study of first-grade readers. *Scientific Studies of Reading*, *15*(4), 338–362. https://doi.org/10.1080/10888438.2010.493964

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel* (Tech. Rep.) Institute for Simulation and Training. https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182. https://doi.org/10.1037/0033-295x.95.2.163

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, *59*(5), 1208–1221. http://scitation.aip.org/content/asa/journal/jasa/59/5/10.1121/1.380986

Kuo, C., & Tjaden, K. (2016). Acoustic variation during passage reading for speakers with dysarthria and healthy controls. *Journal of Communication Disorders*, *62*, 30–44. https://doi.org/10.1016/j.jcomdis.2016.05.003

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests inlinear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word maturity: A new metric for word knowledge. *Scientific Studies of Reading*, *15*(1), 92–108. https://doi.org/10.1080/10888438.2011.536130

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419. https://doi.org/10.1207/s15516709cog0000_25

Loukina, A., Beigman Klebanov, B., Lange, P., Qian, Y., Gyawali, B., Madnani, N., Misra, A., Zechner, K., Wang, Z., & Sabatini, J. (2019). Automated estimation of oral reading fluence during summer camp e-Book reading with MyTurnToRead. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2019* (pp. 21–25). https://doi.org/10.21437/Interspeech.2019-2889

Loukina, A., Liceralde, V. R. T., & Beigman Klebanov, B. (2018). Towards understanding text factors in oral reading. In *Proceedings of the 2018 Conference of the North American chap. of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2143–2154). Association for Computational Linguistics. http://aclweb.org/anthology/N18-1195

Madnani, N., Beigman Klebanov, B., Loukina, A., Gyawali, B., Sabatini, J., Lange, P. L., & Flor, M. (2019). MyTurnToRead: An interleaved e-book reading tool for developing and struggling readers. *Proceedings of the 57th Conference of the Association for Computational Linguistics: System Demonstrations* (pp. 141–146). Association for Computational Linguistics. https://www.aclweb.org/anthology/P19-3024

Marschark, M., Cornoldi, C., Huffman, C. J., Pé, G., & Garzari, F. (1994). Why are there sometimes concreteness effects in memory for prose? *Memory*, *2*(1), 75–96. https://doi.org/10.1080/09658219408251493

McDaniel, M. A., Blischak, D., & Einstein, G. O. (1995). Understanding the special mnemonic characteristics of fairy tales. In C. A. Weaver, S. Mannes, C. R. Fletcher (Ed.), *Discourse Comprehension: Essays in Honor of Walter Kintsch* (pp. 157–176). Routledge.

Miller, J., & Schwanenflugel, P. J. (2006). Prosody of syntactically complex sentences in the oral reading of young children. *Journal of Educational Psychology*, *98*(4), 839–843. https://doi.org/10.1037/0022-0663.98.4.839

Milone, M., (2014). *Development of the ATOS Readability Formula (Tech. Rep.).* http://doc.renlearn.com/KMNet/R004250827 GJ11C4.pdf

Napolitano, D., Sheehan, K., Mundkowsky, R., (2015). Online readability and text complexity analysis with TextEvaluator. *Proceedings of the 2015 Conference of the North American chap. of the Association for Computational Linguistics: Demonstrations* (pp. 96–100). Association for Computational Linguistics. https://doi.org/10.3115/v1/n15-3020

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance. report submitted to the gates foundation*. https://achievethecore.org/content/upload/nelson_perfetti_liben_measures_of_text_difficulty_research_ela.pdf

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, *13*(1), 45–52. https://doi.org/10.3758/bf03193811

O'Brien, E. J., Albrecht, J. E., Hakala, C. M., & Rizzella, M. L. (1995). Activation and suppression of antecedents during reinstatement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 626–634. https://doi.org/10.1037/0278-7393.21.3.626

Patel, R., Connaghan, K., Franco, D., Edsall, E., Forgit, D., Olsen, L., Ramage, L., Tyler, E., & Russell, S. (2013). "The Caterpillar": A novel reading passage for assessment of motor speech disorders. *American Journal of Speech-Language Pathology*, *22*(1), 1–9. https://doi.org/10.1044/1058-0360(2012/11-0134)

Perfetti, C. A. (1994). Psycholinguistics and reading ability. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 849–894). Academic Press.

Peterson, G. E., & Lehiste, I., &. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America*, *32*(6), 693–703. https://doi.org/10.1121/1.1908183

Pfitzinger, H. R., & Reichel, U. D. (2006). Text-based and signal-based prediction of break indices and pause durations. *Proceedings of the 3rd International Conference on Speech Prosody 2006*. International Speech Communication Association. https://doi.org/10.5282/ubm/epub.13159

Pikulski, J. J., & Chard, D. J. (2005). Fluency: Bridge between decoding and reading comprehension. *The Reading Teacher*, *58*(6), 510–519. https://doi.org/10.1598/RT.58.6.2

Poncy, B. C., Skinner, C. H., & Axtell, P. K. (2005). An investigation of the reliability and standard error of measurement of words read correctly per minute using curriculum-based measurement. *Journal of Psychoeducational Assessment*, *23*(4), 326–338. https://doi.org/10.1177/073428290502300403

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. https://www.r-project.org/

Ramli, I., Seman, N., Ardi, N., & Jamil, N. (2016). Rule-based storytelling text-to-speech (TTS) synthesis. *In Matec Web of Conferences*, *77*, 04003. https://doi.org/10.1051/matecconf/20167704003

Rowling, J. K. (1997). *Harry Potter and the philosopher's stone* (1st ed.). Bloomsbury Publishing.

Rowling, J. K., & Dale, J. (2016). *Harry Potter and the sorcerer's stone*. Penguin Random House.

Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal*, *115*(2), 184–209. https://doi.org/10.1086/678294

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2006). How accurate are Lexile text measures? *Journal of Applied Measurement*, *7*(3), 307–322.

Suh, S., & Trabasso, T. (1993). Inferences during reading: Converging evidence from discourse analysis, talk-aloud protocols, and recognition priming. *Journal of Memory and Language*, *32*(3), 279–300. https://doi.org/10.1006/jmla.1993.1015

Theune, M., Meijs, K., Heylen, D., & Ordelman, R. (2006). Generating expressive speech for storytelling applications. *IEEE Transactions on Audio, Speech and Language Processing*, *14*(4), 1137–1144. https://doi.org/10.1109/TASL.2006.876129

Tokuda, K., Hashimoto, K., Oura, K., & Nankaku, Y. (2016). Temporal modeling in neural network based statistical parametric speech synthesis. *9th ISCA Speech Synthesis Workshop*, 2016(2), 106–111. https://doi.org/10.21437/ssw.2016-18

Torgesen, J., Rashotte, C., & Alexander, A. (2001). Principles of fluency instruction in reading: Relationships with established empirical outcomes. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 333–355). York Press.

Toyama, Y., Hiebert, E. H., & Pearson, P. D. (2017). An analysis of the text complexity of leveled passages in four popular classroom reading assessments. *Educational Assessment*, *22*(3), 139–170. https://doi.org/10.1080/10627197.2017.1344091

Turk, A. E., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics*, *28*(4), 397–440. https://doi.org/10.1006/jpho.2000.0123

Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, *35*(4), 445–472. https://doi.org/10.1016/j.wocn.2006.12.001

van Santen, J. P. (1992). Contextual effects on vowel duration. *Speech Communication*, *11*(6), 513–546. https://doi.org/10.1016/0167-6393(92)90027-5

van Santen, J. P. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer Speech Language*, *8*(2), 95–128. https://doi.org/10.1006/csla.1994.1005

Wang, M. D. (1970). The role of syntactic complexity as a determiner of comprehensibility. *Journal of Verbal Learning and Verbal Behavior*, *9*(4), 398–404.

White, L. (2014). Communicative function and prosodic form in speech timing. *Speech Communication*, *63*, 38–54. https://doi.org/10.1016/j.specom.2014.04.003

White, L., & Turk, A. E. (2010). English words on the Procrustean bed: Polysyllabic shortening reconsidered. *Journal of Phonetics*, *38*(3), 459–471. https://doi.org/10.1016/j.wocn.2010.05.002

Yoshimura, T., Tokuda, K., Kobayashi, T., Masuko, T., & Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. *EUROSPEECH*, 2347–2350. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.2007

Zacks, J. M., Speer, N. K., & Reynolds, J. R. (2009). Segmentation in reading and film comprehension. *Journal of Experimental Psychology: General*, *138*(2), 307–327. https://doi.org/10.1037/a0015305

Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science*, *16*(2), 80–84. https://doi.org/10.1111/j.1467-8721.2007.00480.x

Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, *51*(11), 1039–1064. https://doi.org/10.1016/j.specom.2009.04.004

Zeno, S., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Touchstone Applied Science Associates.

Zhao, Y., & Jurafsky, D. (2009). The effect of lexical frequency and Lombard reflex on tone hyperarticulation. *Journal of Phonetics*, *37*(2), 231–247. https://doi.org/10.1016/j.wocn.2009.03.002

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, *6*(5), 292–297. https://doi.org/10.1111/j.1467-9280.1995.tb00513.x

Appendix A

Data Preprocessing Procedure

The original dataset consisted of 2,682 reading turns that were collected from 71 students (33 females [46%]; no gender information for one student; $M_{\text{age}} = 9.35$; $SD = 1.17$; no age information for seven students) across the two data collections. This dataset contained 606 different passages across the book because each student proceeded at their own pace, and students did not always read the same passages from the book.

However, only a subset of this dataset is eligible for analysis because measuring ORF in natural educational settings—where young children read on their own with minimal researcher involvement—introduces a variety of issues that all interact to challenge how well ORF could be measured. Some of the previously identified issues include students not attempting to read during their turn (Beigman Klebanov et al., 2019), recordings with almost no audible speech due to high background noise, silent reading, or a student being distracted by other activities while their speech is being recorded (Loukina et al., 2019). To make valid inferences about text-based variability in children's ORF based on WCPM, the WCPM measure to be analyzed should reasonably reflect how children read the passages in the book out loud. This requires that the data to be analyzed are those reading turns in which students were actually reading the passages they were assigned.

To determine the subset of the data that meet these criteria, the whole dataset was preprocessed using the following procedure. First, following Beigman Klebanov et al. (2019), we identified reasonable turn durations for the passages given the number of words in each passage, the expected reading rates for the students based on ORF norms (Hasbrouck & Tindal, 2017), and an estimate of within-person variation in reading rates across passages. Reading turns that were shorter (longer) than the estimated minimum (maximum) of these durations (translating to reading rates of 50 WCPM and 219 WCPM) were excluded, because these turns were likely too short to contain complete reading data for the passage or too long as to contain substantial nonreading time (e.g., daydreaming, getting distracted). We excluded 1,406 (52%) such turns, and 1,052 of these turns were shorter than ten seconds, indicating that students did not really attempt to read during most of these excluded turns.

Second, we also screened the reading turns for accuracy because reasonable turn durations by themselves are not enough to capture whether students were actually reading the passage *aloud* with sufficient audio quality to allow for transcriptions. Reading turns where the student read fewer than 70% of the words in the passage accurately were excluded. Hasbrouck and Tindal (2017) note that reading accuracy below 90% for grade-level texts suggests frustration from having difficulty in reading or that the text is beyond the instructional level for a reader. Given this standard, we consider our criterion to be reasonably tolerant, allowing for more flexibility in accuracy due to poor reading skill, frustration levels, and other sources of difficulty from reading a different type of text. We excluded 238 more turns with this criterion.

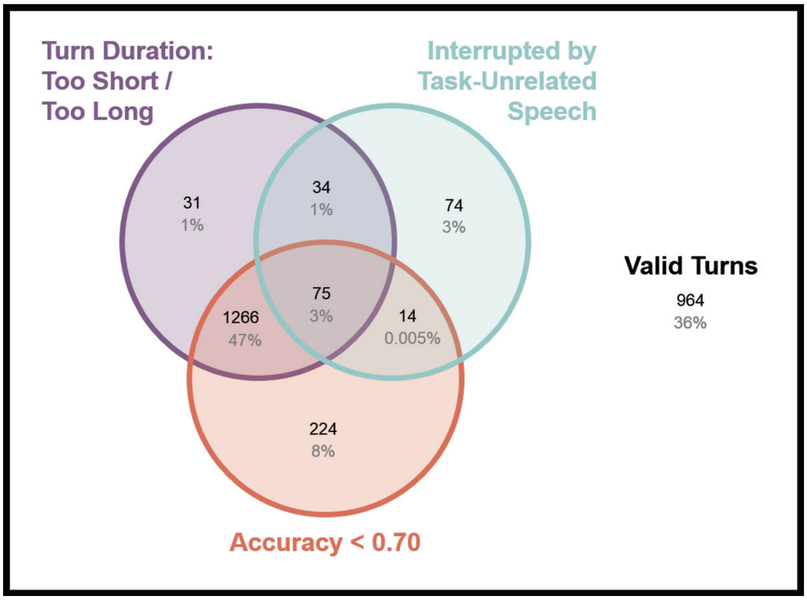
Finally, we excluded turns that were interrupted by task-unrelated speech while the student is reading a passage (e.g., when a student talks to an instructor or another student in the middle of their turn). We excluded 74 more turns with this criterion.

Figure A1 breaks down the membership of the excluded turns in terms of the three preprocessing criteria because many excluded turns satisfy more than one exclusion criterion. Because a substantial amount of data had been excluded, we verified that the exclusion did not introduce selection bias toward specific groups of students or turns: retained students and turns were generally similar to excluded students and turns (see Tables A1 and A2).

(*Appendices continue*)

Figure A1

Venn Diagram Breaking Down the Membership of Excluded and Valid Turns Based on the Preprocessing Criteria


Turn Duration: Too Short / Too Long
Interrupted by Task-Unrelated Speech
31
1%
34
1%
74
3%
Valid Turns
964
36%
75
3%
1266
47%
14
0.005%
224
8%
Accuracy < 0.70

Note. The colored circles show turns that were excluded based on one or more of the three criteria. The turns not in the circles were the turns selected for analysis. See the online article for the color version of this figure.

Table A1

Characteristics of Retained and Excluded Students

Measure	Retained students ($N = 56$)	Excluded students ($N = 15$)	Difference
Proportion female	0.46	0.47	0.01
Proportion English L1 at home	0.82	0.73	0.09
Proportion familiar with HP1	0.36	0.47	0.11
Mean age	9.61	9.09	0.52
Mean HP1 comprehension accuracy	0.73	0.50	0.23

Note. HP1 = Harry Potter and the Sorcerer's Stone.

(*Appendices continue*)

Table A2
Passage Characteristics of Retained and Excluded Turns

Measure	Retained turns (*N* = 964)	Excluded turns (*N* = 1718)	Difference
TextEvaluator	571.78 (178.47)	571.16 (184.47)	0.62
Flesch-Kincaid	5.31 (2.31)	5.39 (2.50)	0.08
TTS	156.12 (13.20)	155.72 (13.78)	0.40
Academic vocabulary	25.39 (7.70)	25.13 (8.57)	0.26
Concreteness	60.73 (8.80)	60.33 (9.54)	0.41
Degree of narrativity	82.07 (6.82)	81.70 (7.90)	0.37
Interactive/conversational style	73.16 (14.76)	71.00 (17.54)	2.16
Level of argumentation	49.23 (23.04)	47.13 (23.55)	2.09
Lexical cohesion	44.75 (7.52)	44.93 (7.81)	0.19
Syntactic complexity	47.12 (14.12)	46.80 (15.03)	0.31
Word unfamiliarity	53.27 (10.42)	54.64 (11.15)	1.37

Note. TTS = text-to-speech synthesis. Standard deviations in parentheses. TextEvaluator component dimension scores are scaled from 1 to 100, with higher numbers indicating higher levels of the dimension.

Appendix B

Follow-Up Analysis Results

Because many students had missing values for the student-level variables, adding these variables in the follow-up model while ensuring that these models are comparable with the planned models was an issue. To address this missingness issue and retain all observations from the planned models in the follow-up model, we built the missingness of these variables into the model specification, which is one way of accounting for missingness in data. We added missingness indicator variables that marked students for whom student-level variables were missing. The addition of these variables allowed the direct comparison of the follow-up model to the planned models because all models were fit to the exact same dataset. Table B1 reports the follow-up model to Model 4 that accounted for various student-level variables and interactions to explore whether production-related effects were driven by specific students (Model 5).

In fitting Model 5, the predictors were reparameterized as follows: TextEvaluator scores and TTS estimates were z-score transformed across passages as in Model 4. Students' age and HP1 comprehension accuracy (in percent, from 0% to 100%) were mean-centered across students, whereas the number of valid turns they made was centered around the median. Lastly, the within-session turn index was shifted so that 0 meant the first turn a student made within a session.

(*Appendices continue*)

Table B1

Follow-Up Model Accounting for Potential Confounds (Model 5)

Fixed effect	β	*t*
Intercept	91.88	5.85***
Data collection[a]	−2.40	−0.19
Focal measures		
TextEvaluator	−3.22	−2.51***
TTS	5.46	4.78***
Student characteristics		
Female	0.41	0.06
Age	4.35	1.31
n_{turns}	0.37	1.76
HP1 Comprehension Accuracy	0.53	2.82**
HP1 Familiarity[b]	−0.96	−0.12
The Harry Potter book was boring.[c]		
Disagree	−3.59	−0.34
Agree	9.97	0.96
Strongly agree	−1.97	−0.15
Session variables		
Within-session Turn Index	0.40	0.58
Interactions		
TextEvaluator × Age	1.50	1.36
TextEvaluator × HP Comprehension Accuracy	−0.08	−1.45
TextEvaluator × HP1 Familiarity	−0.33	−0.16
TTS × Age	−0.28	−0.27
TTS × HP1 Comprehension Accuracy	0.11	2.08***
TTS × HP1 Familiarity	1.01	0.53
Data Collection × Turn Index	−0.38	−0.52
Missing indicator variables[d]		
Age	1.41	0.14
HP1 Familiarity	0.70	0.06
The Harry Potter book was boring. (missing)	9.05	0.63

Random effect	*SD*		*r*	
Site				
Intercept	0.00			
Student				
Intercept	21.79	1		
TextEvaluator	3.60	0.16	1	
TTS	2.74	0.26	−0.78	1
Passage				
Intercept	6.96			
Residual	15.72			

Note. TTS = text-to-speech synthesis; HP1 = Harry Potter and the Sorcerer's Stone.
[a] Reference category is Data Collection A. [b] Reference category is Unfamiliar. [c] Reference category is Strongly disagree. [d] Reference category is Not missing.
** $p \leq .01$. *** $p \leq .001$.

Received April 23, 2020
Revision received January 28, 2021
Accepted January 29, 2021 ■