Continuous fluency tracking and the challenges of varying text complexity

Beata Beigman Klebanov, Anastassia Loukina, John Sabatini, Tenaha O'Reilly
Educational Testing Service
660, Rosedale Rd, Princeton NJ, USA
08541, NJ, USA
{bbeigmanklebanov, aloukina, jsabatini, toreilly}@ets.org

Abstract

This paper is a preliminary report on using text complexity measurement in the service of a new educational application. We describe a reading intervention where a child takes turns reading a book aloud with a virtual reading partner. Our ultimate goal is to provide meaningful feedback to the parent or the teacher by continuously tracking the child's improvement in reading fluency. We show that this would not be a simple endeavor, due to an intricate relationship between text complexity from the point of view of comprehension and reading rate.

1 Introduction

According to the 2015 report from the National Assessment of Educational Progress on reading achievement, 31% of U.S. 4th graders read below the Basic level.[1] Our goal is to help low-proficiency readers such as these improve their reading skill.

The critical transition from word-by-word reading to fluency, or from learning how to read to reading for learning or enjoyment, requires extended and sustained reading practice. To encourage such practice we propose an educational application which combines (1) an excellent story to achieve engagement (such as "Harry Potter and the Sorcerer's Stone" by J. K. Rowling), and (2) a virtual reading companion, implemented through an audiobook, who would take turns reading aloud with the child – "you read a page, I read the next one". The turn-taking allows the child to alternate between the more effortful reading and the less effortful listening, as well as supplies a model reading of many of the words and phrases the child will encounter during his turn.

In addition to supporting sustained reading by children, the system will also provide the teacher or parent with a detailed picture of the child's developmental trajectory, by continuously tracking the child's reading fluency throughout his reading turns. Oral reading fluency is not only an important indicator of reading skill in itself (Hudson et al., 2008; Fuchs et al., 2001), for students in early elementary grades it is also strongly correlated (r around 0.7) with reading comprehension (Roberts et al., 2005; Good et al., 2001).

The standard measure of oral reading fluency is words correct per minute (henceforth, WCPM) (Wayman et al., 2007), combining aspects of speed and accuracy of oral reading.[2] Several studies (Balogh et al., 2007; Zechner et al., 2009) showed that WCPM can be accurately computed automatically using an automated speech recognizer (ASR) and a string matching algorithm; this approach has already been incorporated into many commercial and research systems for automated oral fluency assessment such as VersaReader (Balogh et al., 2012) or Project LISTEN (Mostow, 2012) (see also Eskenazi (2009) for a review).

Previous studies on reading fluency indicate that WCPM may vary across different texts (Ardoin et al., 2005; Compton et al., 2004). It seems reasonable to assume that variation in text complexity/readability might be one of the sources of variation in oral reading fluency across different passages: Texts that cause comprehension difficulties may also elicit less fluent reading. In fact, this assumption underlies text selection for tests of oral reading fluency such as DIBELS (Good and Kaminski, 2002) that rely on readability to select

[1] https://www.nationsreportcard.gov/reading_math_2015/#reading/acl?grade=4

[2] In some studies, reading rate (words per minute) is used as a separate measure while fluency is defined in terms of expressiveness and adherence to syntax (Danne et al., 2005).

comparable passages (Francis et al., 2008). Since in our application the child will be reading different passages in the book on different days, it is possible that the differences between passages would confound the measurement of the child's progress. In this case, WCPM would need to be adjusted to account for such differences in order to produce interpretable feedback.

Previous work generally focused on text properties and WCPM in short texts that have already been controlled for grade-level appropriate readability. Little is known about the variability of text complexity across a whole book and how this may affect WCPM of a child reading the book. Therefore, the focus of this paper is to see whether an adjustment of WCPM to text is in fact necessary in our context, and, if so, whether it can be done using a state-of-the-art text complexity measure.

We address the following research questions: (1) What is the extent of variation in passage complexity in J. K. Rowling's "Harry Potter and the Sorcerer's Stone" (henceforth, **HP1**)? (2) Does the complexity of the text actually impact reading fluency as measured by WCPM? (3) Do automatically generated estimates of text complexity correspond to the observed fluency patterns?

The rest of the paper is organized as follows. We first introduce previous work related to text complexity measurement and the relationship between text complexity and oral reading fluency. We then present the results of two studies: In the first study we looked at variation in text complexity across passages selected from HP1. In the second study we investigate how text complexity estimates relate to WCPM of children reading selected passages from the book. Our findings are then discussed and implications for research on continuous tracking of fluency are drawn.

2 Related Work

Text Complexity Estimation: While for Dale and Chall (1949) the notion of text readability involved "the extent to which they [readers] understand it [the text], read it at an optimal speed, and find it interesting",[3] most classical (Flesch, 1948; Gunning, 1952; Kincaid et al., 1975; McLaughlin, 1969) and modern (Sheehan et al., 2014; Flor and Beigman Klebanov, 2014; Vajjala and Meurers, 2012; Schwarm and Ostendorf, 2005) measures of text readability/complexity focus on reading comprehension, including special formulas and models designed for special populations, such as young children (Spache, 1953), learners of English as a second language (Beinborn et al., 2014; Heilman et al., 2007), adults with mental disabilities (Feng et al., 2009), among others.

While comprehension-based complexity estimation of relatively short reading passages has been the subject of extensive research for many decades, there is little research on estimating the complexity of long, book-level texts. In early work on readability, Fowler (1978) estimated readability of a novel using the mean of readability estimates of fifteen randomly selected 100-word passages from the novel. Milone (2012) generates book-level complexity estimates by combining complexity estimates for the text in the book with a measure based on the length of the book, following the observation that longer books tend to be more difficult, all else being equal. He decided to base the estimate of text complexity in the book on the analysis of the whole book, as opposed to samples from the book, based on the observation of extensive within-text variability in estimates of text complexity and the concomitant hazard of a large sampling error if only parts of the book are taken into account during complexity estimation (see Appendix E in Milone (2012)). For example, the book *Black Beauty* yields a grade-level estimate of 5.4 based on the text of the whole book; looking at 500-word slices yields estimates anywhere from 2.2 to 9.5 per slice – a range of 7 grade levels. This finding raises the question of a young reader's experience in the face of such variability. To our knowledge, our project is the first study to address variation in within-book reading experiences in general, and variation in oral reading performance specifically.

Relationship between oral reading fluency and text complexity: In Compton et al. (2004), 248 low and average-achieving second graders each read 15 passages of comparable readability levels; their reading performance was recorded in terms of accuracy (proportion of words read correctly) and fluency (WCPM). Analyzing the relationship between textual characteristics and performance, researchers found that Flesch-Kincaid measure, Spache measure, and average sentence length did not significantly correlate with performance. On the other hand, they found that percentage of high frequency words was significantly

[3] Quoted from DuBay (2004).

correlated with both performance measures. Ardoin et al. (2005) examined a number of readability formulas for their ability to predict WCPM and found generally fairly low correlations (r<0.5).

In Petscher and Kim (2011), about 35,000 students in grades 1-3 read three grade-level-appropriate passages (as measured by Spache formula) during each of 4 administrations of an oral reading fluency test throughout the year. The authors estimated the amount of variability in WCPM that was attributable to variation among students vs variability across the text passages. Their results showed that 2%-4% was attributable to variability in passages and/or order of passages for grade 1, with higher proportions for grades 2 (5%-6%) and 3 (3%-9%). Petscher and Kim (2011) also observed an increase in the reading rate from the first to the third administered passage within an assessment, consistently with other studies (Francis et al., 2008; Jenkins et al., 2009), pointing to the existence of practice effects in oral reading performance of consecutively read texts.

To summarize, the related work suggests that (1) some amount of variation in reading fluency is attributable to variation in text passages being read, for early elementary grade children; (2) classical readability formulas are not very effective predictors of oral reading fluency.

We note however that passage readability/complexity variation across texts used in previous studies tends to be limited, since texts selected for assessments are typically controlled for grade-level-appropriate readability. In contrast, we consider a case where children are reading a long novel that is not specifically designed to be grade-level controlled; we therefore expect more variation in complexity across different passages in a book. Larger variation may show better alignment between reading rates and text complexity estimates.

3 Study I: Text complexity in Harry Potter and the Sorcerer's Stone

3.1 Data and methodology

For this first study we considered the variation in text complexity in J. K. Rowling's "Harry Potter and the Sorcerer's Stone". We first split the book into a series of consecutive, non-overlapping 250 word chunks. These should take 2-3 minutes to read for our target population and constitute the approximate amount of text to be read by the child at each turn. For each chunk, after 250 words, we either extended or reduced the chunk to the end of a paragraph, thus ensuring that each passage had a natural break point.

The whole book consists of 79,508 words spread across 17 chapters. We created 318 consecutive passages, with a mean length of 250.0 words (SD=16.9). The shortest passage contained 177 words and the longest passage contained 309 words. Half of the passages (II and III quartiles) fell within 242-259 words range.

We used TextEvaluator,[4] a state-of-the-art measure of comprehension complexity of a text (Napolitano et al., 2015; Sheehan et al., 2014, 2013; Nelson et al., 2012),[5] to conduct text complexity analyses. TextEvaluator extracts a range of linguistic features and uses them to compute a complexity index on the scale of 100-2000, as well as an overall grade equivalent score. TextEvaluator computes three complexity scores based on the models optimized for literary, informational and mixed texts. We used the literary metric as the final complexity score for our passages since all texts were excerpts from a novel.[6]

In addition to the overall score, several dimension scores are provided, including: Syntactic Complexity (using features related to sentence complexity); Academic Vocabulary (the extent to which words in the text are characteristic of academic texts); Word Unfamiliarity (a composite measure of word frequency); Lexical Cohesion (measures the degree of overlap between concepts across adjacent sentences within paragraphs); Level of Argumentation (indexes the ease or difficulty of inferring connections across sentences when the underlying format of a text is argumentative); additional dimensions include Interactive/Conversational Style, Concreteness, Degree of Narrativity.

We note that passage lengths between 177 and

[4] https://textevaluator.ets.org/

[5] TextEvaluator appears in the Nelson et al. (2012) benchmark as SourceRater.

[6] A reviewer of this paper pointed out that Text Evaluator includes an automatic genre classifier which is used to determine the final complexity score (Sheehan et al., 2013), and that it is possible that some passages in a novel could be more on the informational side. In our study, 302 passages (95%) were classified as literary texts by TextEvaluator's genre classifier. Among the remaining 16 passages, 7 passages were classified as informational texts and 9 passages as mixed texts. None of the selected passages (in section 4.1) belong to these 16. Using final instead of literary scores had a negligible effect on statistics reported in section 3.2.

309 words are within scope for TextEvaluator, albeit on the shorter side of the range: Sheehan et al. (2013) report an evaluation with texts ranging in length from 112 to more than 2,000 words.

For this analysis, we treated each chunk from the book as an independent passage. Thus, TextEvaluator had no access to information about other passages. One might contend that there are limitations to such an approach, as some aspects of difficulty of the text may change as the reader accumulates knowledge about the world of the book. For example, words that are initially unfamiliar, such as names of characters, magic creatures and artifacts, spells and curses, would become increasingly familiar as the story progresses. In contrast, other aspects of complexity, such as the syntactic complexity of sentences, are less likely to become more or less challenging as one reads further into the book. In the current study, we have not attempted to capture any such text continuity effects.

3.2 Results

The overall TextEvaluator complexity of passages across the book varied from 160 to 1150 with average complexity 613.4 (SD=163.1). In terms of grade levels this corresponds to variation from second to eleventh grade, with the average around grade six.

The dimension scores also varied across the book although the patterns were different for different dimensions. Figure 1 shows the distributions for different dimension scores. The scale for all scores is 0-100. The score for Academic vocabulary was consistently low across all passages (Mean=27, SD=7.3), while the score for narrativity was consistently high (Mean=83, SD=5.8). The score for the Level of Argumentation showed the largest spread (Mean=53 and SD=19.8). We also note the substantial spread in Syntactic complexity (Mean=46.1 and SD=11.3).

We also considered how the complexity varies as one proceeds through the book (Figure 2). The red line shows values for each passage, the blue line shows a smoothed estimate calculated using lowess (Cleveland, 1979).[7] The plot shows there is a substantial fluctuation from passage to passage as well as potentially longer-range trends that may correspond to the book's narrative structure. Specifically, the peak around 130-140 corresponds to the description-heavy introduction to

[7] as implemented in (Seabold and Perktold, 2010)

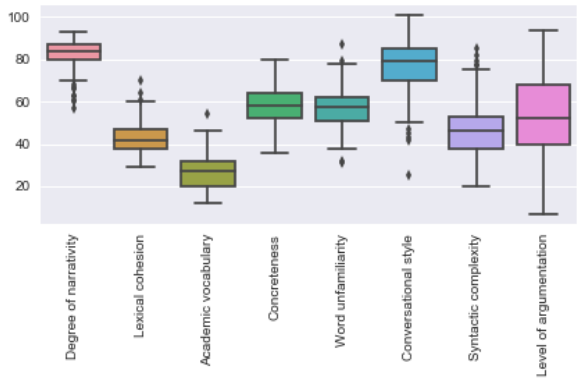

Figure 1: Distributions of scores for various dimensions of text complexity in HP1. The dimensions are ordered on the x-axis by spread (SD).

Hogwarts and Harry's first classes; the valley around 300 corresponds to the fast-moving final stand-off between Harry and Voldemort/Quirrell.

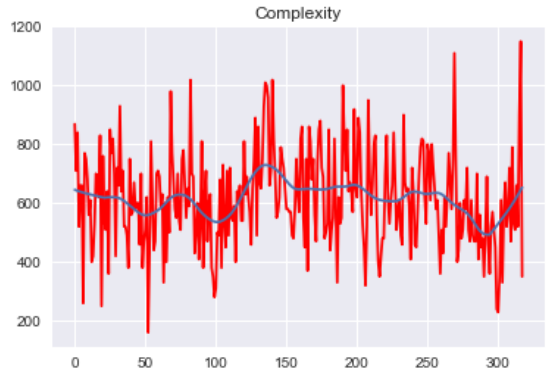

Figure 2: Distribution of holistic text complexity scores as one proceeds through HP1.

The answer to research question 1 is thus: The extent of variation in text complexity across passages in the book is very substantial. If text complexity has any systematic effect on the oral reading performance, the extent of variation in complexity suggests that it is likely to become a major confounding factor in tracking the child's progress in fluency while reading the book.

4 Study II: Text complexity and oral reading fluency

Our second question is: Does the complexity of the passage that is being read significantly impact children's reading fluency for the passage? In order to answer this question, we selected 3 passages with very large differences in text complex-

ity as estimated by TextEvaluator, and collected oral reading fluency estimates for these passages from a sample of 2-4 graders. The details of the procedure and the results are described in this section.

4.1 Passage selection

We ordered all 318 passages by estimated text complexity, and selected passages from the middle of the distribution and from the lowest and highest deciles. In addition to TextEvaluator score, when selecting passages we also took into account whether a passage could be reasonably read as stand-alone text. Table 1 shows the characteristics of these three passages. All passages are from the first chapter of the book.

Passage	# words	TE score	Complexity Percentile
Easy	226	260	1.9%
Medium	282	580	51.5%
Hard	246	800	90.3%

Table 1: The characteristics of the passages used for the data collection: length in words, complexity, complexity relative to the whole book.

4.2 Data collection procedure

The recordings took place in an office with several children recorded simultaneously. The texts were presented on screen and the audio was captured using the head-set with a microphone.

Before reading the experimental passages, the child first listened to the passage that begins the first chapter of HP1 (starting with "Mr. and Mrs. Dursley of number four, ...") as narrated by the professional actor Jim Dale (Rowling and Dale, 2016). Then the child read aloud the passage immediately following the passage read by the narrator. Since all children read this passage first, this passage is used as a reference text to measure baseline WCPM for each child.

The experimental passages were then presented to children in a randomized order, to allow separation between text and order effects in subsequent analyses (Petscher and Kim, 2011; Francis et al., 2008; Jenkins et al., 2009). The children were asked to read at their natural pace.

A total of 30 children took part in this data collection selected via a convenience sample. Table 2 shows the distribution by grade and gender and the average age in each group. All recordings were done in April of 2017.

Grade	Girls	Boys	Mean age
2	7	3	8;3
3	3	7	9;0
4	6	4	10;2

Table 2: The demographic characteristics of participants.

4.3 Computation of oral reading measures

To compute WCPM we used a professional transcription agency to obtain word-by-word transcriptions of each child's reading and aligned them to the passage text using an algorithm based on dynamic programming. We next computed how many words in the original passage matched those in the transcription. This algorithm is similar to that used to compute ASR word error rate, but following the standard practice in reading research we only penalized substitutions and deletions and did not take into account any insertions. Most children's reading closely followed the texts, with the average of 93.8% of all words in each text read correctly (SD=3.7, min=82.7%, max=99.6%).

We manually identified in each recording the time stamps where the child started and finished reading the text. WCPM was computed by dividing the total time it took the child to read the text by the total number of matched words in each text. The average WCPM in the experimental texts in our corpus was 117.1 (SD=27.3, min=57.2, max=196.0). To get an idea where these readers stand with respect to general population of U.S. children of comparable age, we consulted the WCPM norms in Table 1 of Hasbrouck and Tindal (2006), and found that a grade-stratified sample of children from grades 2-4 during spring term is expected to read, on average, at 106 WCPM. The observed rate of 117 WCPM corresponds to 60% percentile – somewhat above average. We note that this is only a rather rough estimate of these children's fluency relative to peers, since the experimental texts differ in complexity substantially from the grade-leveled materials used for oral reading fluency assessments. Still, this estimate accords with our observation during the data collection that these children generally read quite fluently and accurately for their age.

4.4 Results

To evaluate the effect of text on WCPM, we used a mixed effects linear model. These models offer a more powerful way to conduct repeated-measures analyses than a simple repeated-measures ANOVA, because they make it possible to combine both continuous and categorical predictors. We used WCPM as the dependent variable and speaker identity as a random factor. We included the following fixed factors: text identity (categorical), the baseline WCPM on the reference text (continuous), and order in which each text was read (continuous). In addition to the main effects, we also included the interaction between text identity and the baseline WCPM. Table 3 shows the standardized coefficients and significance values for the model. We took WCPM for the Medium text as the reference category.

	Variable	Coeff.	P > |z|
1	Intercept	0.522	<0.001
2	text-easy	-0.814	<0.001
3	text-hard	-1.165	<0.001
4	base_wcpm	0.893	<0.001
5	text-easy:base_wcpm	-0.258	0.001
6	text-hard:base_wcpm	-0.132	0.089
7	order	0.046	0.236

Table 3: The standardized coefficients and their significance values for fixed effects used to predict WCPM on each text (N=90). In addition to the fixed effects, the model also included the random effect for speakers (not shown in the table).

First, we observe that the child's baseline reading fluency estimated from the reference text is a significant factor, as expected. Second, we note that the order in which the experimental texts were presented does not yield a significant effect.

The identity of the passage (Easy, Medium, Hard) has a significant effect on reading fluency. Thus, the Hard text is read 1.2 standard deviations less fluently than the Medium text (row 3); this result accords with expectations. The result in row 2 is surprising: There is a highly significant and large difference in WCPM between the Easy text and the Medium text, but it is in the *opposite* direction – the Medium text is read 0.8 standard deviations *more fluently* than the Easy text. Thus, while the results clearly attest to a substantial effect of the text on WCPM, the estimates of text complexity are in a rather dramatic mis-alignment with the pattern of the oral reading.

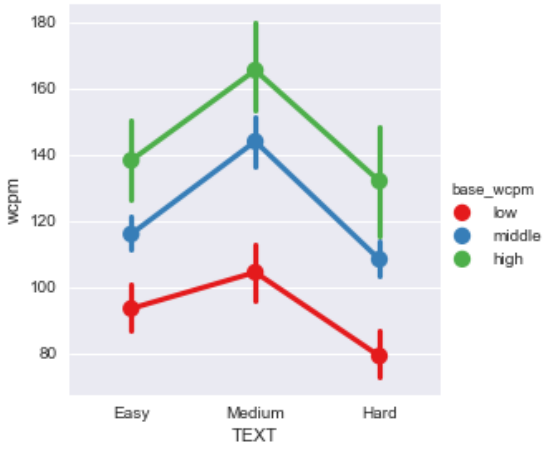

Figure 3: Average WCPM for the three texts in our study. To illustrate the interaction between fluency and text we divided all speakers into three equal bins based on 'base_wcpm'.

Row 5 in Table 3 shows a significant interaction effect between text and base WCPM, for Medium vs Easy texts: The higher the base reading fluency of the child (base_wcpm), the bigger the difference in WCPM between Easy and Medium text. This effect is consistent with the tendency shown in row 6, though it does not reach significance: the more fluent readers also tended to differentiate more between the Medium and Hard texts. This finding suggests that more fluent readers seem to have a tendency to differentiate their oral reading pattern depending on the text they read to a larger extent than the less fluent readers. Indeed, there is a significant, medium-strength correlation between a child's *average* WCPM for the three texts and his or her *variance* in WCPM across these texts: $r = 0.47$, $p < 0.01$.

Figure 3 shows the average of WCPM across the three texts in our study. To illustrate the interaction between text and fluency we divided all children into three equal groups based on their base WCPM on the reference text.

In order to check whether the impact of the text is mostly about the accuracy aspect of the fluency measure (words read *correctly* per minute) or about the reading rate itself (words or syllables per minute), we repeated the analyses above using either words per minute or syllables per minute as the dependent variable instead of WCPM. The results are very similar to those reported in Table 3: Base reading rate has a significant effect; text iden-

tity has a significant effect, with one of the comparisons going in the opposite direction from that predicted; the interaction effect for base reading rate and text for Easy vs Medium is significant; order and the second interaction effects are not significant. This finding suggests that, at least for these readers, the basic speed of reading is systematically affected by the identity of the text.

5 Discussion

The main finding in our study is that while different passages consistently elicit different reading rates, text complexity as estimated by a state-of-the-art measure does not predict the differences correctly – a passage that is rated as 3.2 grade levels more difficult than another is in fact read significantly *faster*, consistently across readers. We consider several possible reasons for this effect:

- TextEvaluator's complexity estimates may not be accurate when applied to passages from a novel.
- *Oral* reading is not only a kind of reading, but also a kind of *speaking*. Reading rate might thus be affected by properties of speech, in a direction that differs, or even contradicts, the impact of text complexity.
- Reading *a story* aloud, or *narration*, is not only a kind of oral reading, but also a kind of *performance* for an audience. While children are not explicitly asked to narrate, the nature of the text might drive them to do so, as well as the model reading provided by the narrator of the audiobook (recall that the children listened to a passage narrated by the actor Jim Dale before reading aloud their own passages). Variation in WCPM across texts could be effected by demands of expressive narration that are unrelated, or at least not directly related, to comprehension complexity of the text.

5.1 Estimation of text complexity in book excerpts

One possible hypothesis for explaining the finding is that TextEvaluator scores may not provide an adequate estimate of complexity for book excerpts, since the engine, like many other complexity/readability measures, has been developed and validated for estimating reading comprehension difficulty of standalone passages meant for use in assessments. In particular, the guidelines for using TextEvaluator specifically exclude drama, yet the Easy text includes an informal conversation with punctuation used to indicate emotions of the interlocutors. The Easy text contains the following excerpts:

> (1) "Well, I just thought ... maybe ... it was something to do with ... you know ... her crowd."
> (2) "Funny stuff on the news," Mr. Dursley mumbled. "Owls ... shooting stars ... and there were a lot of funny-looking people in town today ... "

TextEvaluator treats "..." as if they were sentence-final periods, as in:

> (3) "Well, I just thought. Maybe. It was something to do with. You know. Her crowd."
> (4) "Funny stuff on the news," Mr. Dursley mumbled. "Owls. Shooting stars. And there were a lot of funny-looking people in town today."

This creates multiple very short sentences which in turns lowers the complexity score since average sentence length is one of the indicators of text complexity. However, an alternative interpretation where utterance-internal "..." are more akin to commas is also possible, as in:

> (5) "Well, I just thought, maybe, it was something to do with, you know, her crowd."
> (6) "Funny stuff on the news," Mr. Dursley mumbled. "Owls, shooting stars, and there were a lot of funny-looking people in town today."

After substituting (5) and (6) instead of (1) and (2), respectively, the estimation of the complexity of the text increased from 260 to 300, due to the increase in average sentence length. It is possible that there are other ambiguities that could be resolved in ways with differing levels of complexity, as well as other indicators of complexity that are not picked up or interpreted as such by TextEvaluator. We note that the particular issue pointed out above would not be specific to TextEvaluator, as many complexity indices include average sentence length as a component. Generally, it is possible that measures developed predominantly for

analyzing passages for assessments would not account correctly for stylistic devices used in novels. Indeed, Nelson et al. (2012) observed that various measures of text complexity, including TextEvaluator, generally had better correlations with grade level for informational texts than for narrative texts.

5.2 Text complexity vs general properties of speech prosody

Average sentence length is a text complexity indicator used in both classical (such as Flesh-Kincaid) and modern text complexity measures – longer sentences tend to be more difficult from the point of view of comprehension. From the point of view of speech prosody, however, it is not clear that a long sentence would be uttered slower than a few shorter sentences covering, in total, the same number of words (or syllables). Studies of speech prosody have consistently demonstrated that the duration of segments increases at certain important locations within utterances, sentence boundaries being one such location (see White (2014) for a detailed review of this topic). As a result, the overall time it would take to read a text with many short sentences might in fact be longer than a text with the same number of words split into longer sentences.

We observe that the actor who is narrating the audiobook is unlikely to be influenced by text complexity to the same extent as young readers who are still learning to read. It is hard to imagine that any of the passages in HP1 are genuinely difficult for the narrator, as a reader who is not only proficient but highly skilled,[8] and also very familiar with the text he is narrating. Thus, if we observe substantial variation in reading rates across the three texts for the narrator, it is likely that the reason for the changes is something other than text complexity, as quantified by comprehension-related measures.

To test this hypothesis we compute the reading rate for the narrator following the same approach as described above. We found that the patterns of the reading rate of the narrator closely followed those we observed for children in our study: The Easy text was read slower than the Medium text which in turn was read faster than the Hard text. Figure 4 shows the WCPM for the narrator relative to the children in our study.

[8]The narrator, Jim Dale, has won Grammy awards for his recordings of two of the seven Harry Potter books.

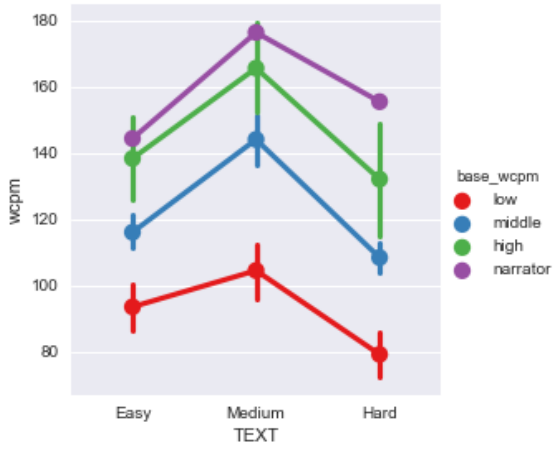

Figure 4: Average WCPM for children in our corpus and the audiobook narrator (purple).

It appears that readers with different levels of reading fluency (young learners and a performing professional), are affected by some aspect of the text *in a similar way*, which makes it less likely that this aspect is directly related to comprehension complexity, since complexity should pose much less of a challenge for a performing professional than for a second grader. General patterns of speech are one potential reason (as also mentioned in section 5.2); another possibility is that in the context of narrating a story, reading rate is affected by "directives" in the text that govern expressive oral reading performance of each passage (cf. Theune et al. (2006)). Such directives could include markers of hesitation, emphasis, surprise, stuttering, etc.; some of these might have a systematic effect on reading rate.

5.3 Interaction between base fluency and impact of text identity

Finally, we also observed an interaction effect between the reader's baseline fluency and the extent to which text identity impacts that reader's fluency. Specifically, for one of the pairs of texts, more fluent readers tend to have significantly larger differences in reading rates between the two texts. This finding is in agreement with the literature – Petscher and Kim (2011) found that the proportion of reading rate variance attributable to variation in passages tends to increase with grade, for grades 1 to 3. This could be due to more proficient readers reading more expressively by attending more closely to the rhetorical and prosodic clues that impact the reading rate. Lower profi-

ciency readers are likely to be focused more on reading words, while better readers also attend to other structures in the text. Indeed, Schwanenflugel et al. (2015) found that more fluent readers communicate linguistic focus while reading aloud by prosodically marking direct quotes, exclamations, and contrastive words. This direction requires further exploration; if the finding is replicated with a larger sample of readers with more variation in reading proficiencies, it would suggest that the extent of adjustment for text effects needs to be moderated by the reader's baseline reading rate.

6 Conclusion

In this paper we discussed the challenges of continuous fluency tracking within an assisted-reading intervention where a child reads a long novel rather than a set of grade-controlled passages. We showed that there is substantial variation in passage difficulty across a single book as estimated by a state-of-the-art measure of text complexity for comprehension and a consistent variation in reading rates between passages. Continuous fluency tracking needs to account for this variability. The results of our small preliminary study suggest not only that a state-of-the-art measure of comprehension complexity does not predict reading rates well, but in fact substantial variation in reading rates may be unrelated to comprehension complexity of the text. Additional research needs to be done to further explore these relationships.

7 Acknowledgements

We would like to thank René Lawless, Kelsey Dreier, and Thomas Florek for their help with data collection; Diane Napolitano for her help with running TextEvaluator; Patrick Lange and Binod Gyawali for their help with preparing the narrator's data. We would also like to thank our many colleagues at ETS who discussed this work with us and provided useful comments.

References

Scott P. Ardoin, Shannon M. Suldo, Joseph Witt, Seth Aldrich, and Erin McDonald. 2005. Accuracy of readability estimates' predictions of CBM performance. *School Psychology Quarterly* 20(1):1–22.

Jennifer Balogh, Jared Bernstein, Jian Cheng, and Bren Townshend. 2007. Automatic evaluation of reading accuracy: Assessing machine scores. *Proceedings of SLaTE ITWR Workshop* .

Jennifer Balogh, Jared Bernstein, Jian Cheng, Alistair Van Moere, Brent Townshend, and Masanori Suzuki. 2012. Validation of automated scoring of oral reading. *Educational and Psychological Measurement* 72(3):435–452.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Readability for foreign language learning: The importance of cognates. *International Journal of Applied Linguistics* 165(2):136–162. https://doi.org/doi:10.1075/itl.165.2.02bei.

William S. Cleveland. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 74(368):829–836.

Donald Compton, Amanda Appleton, and Michelle Hosp. 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* 19(3):176–184.

Edgar Dale and Jeanne Chall. 1949. The concept of readability. *Elementary English* 26(23).

Mary C. Danne, Jay R. Campbell, Wendy S. Grigg, Madeline J. Goodman, and Andreas Oranje. 2005. Fourth-Grade Students Reading Aloud: NAEP 2002 Special Study of Oral Reading. The Nation's Report Card. NCES 2006-469. *National Center for Education Statistics* .

William DuBay. 2004. The principles of readability. http://files.eric.ed.gov/fulltext/ED490073.pdf.

Maxine Eskenazi. 2009. An overview of spoken language technology for education. *Speech Communication* 51(10):832–844.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Association for Computational Linguistics, Athens, Greece, pages 229–237.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology* 32(3):221–233.

Michael Flor and Beata Beigman Klebanov. 2014. Associative lexical cohesion as a factor in text complexity. *International Journal of Applied Linguistics* 165(2):223–258.

Gilbert Fowler. 1978. The comparative readability of newspapers and novels. *Journalism Quarterly* 55(3):589–591.

David J. Francis, Kristi L. Santi, Christopher Barr, Jack M. Fletcher, Al Varisco, and Barbara F. Foorman. 2008. Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology* 46:315–342.

Lynn S. Fuchs, Douglas Fuchs, Michelle K. Hosp, and Joseph R. Jenkins. 2001. Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading* 5(3):239–256.

Roland H. Good, Deborah C. Simmons, and Edward J. Kame'enui. 2001. The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading* 5(3):257–288.

Ronald. H. Good and Ruth. A. Kaminski. 2002. DIBELS oral reading fluency passages for first through third grades. *Technical Report* 10. Eugene, OR: University of Oregon.

Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill.

Jan Hasbrouck and Gerald Tindal. 2006. Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher* 59(7):636–644.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Association for Computational Linguistics, Rochester, New York, pages 460–467.

Roxanne F. Hudson, Paige C. Pullen, Holly B. Lane, and Joseph K. Torgesen. 2008. The complex nature of reading fluency: A multidimensional view. *Reading & Writing Quarterly* 25(1):4–32.

Joseph R. Jenkins, J. Jason Graff, and Diana L. Miglioretti. 2009. Estimating reading growth using intermittent CBM progress monitoring. *Exceptional Children* 46:315–342.

J. Peter Kincaid, Robert P. Fishburne, Richard Rogers, and Brad Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy enlisted personnel. *Institute for Simulation and Training* 56.

G. Harry McLaughlin. 1969. SMOG grading - a new readability formula. *Journal of Reading* 12(8):639–646.

Michael Milone. 2012. The Development of ATOS: The Renaissance Readability Formula. http://mpemc.weebly.com/uploads/5/4/0/7/5407355/development_of_atos.pdf.

Jack Mostow. 2012. Why and how our automated reading tutor listens. *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training* pages 43–52. https://www.cs.cmu.edu/˜listen/pdfs/2012-05-05 ISADEPT2012 keynote final.pdf.

Diane Napolitano, Kathleen M. Sheehan, and Robert Mundkowsky. 2015. Online readability and text complexity analysis with TextEvaluator. In *Proceedings of HLT-NAACL*. pages 96–100.

Jessica Nelson, Charles Perfetti, David Liben, and Meredith Liben. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. In *Technical Report to the Gates Foundation*. http://achievethecore.org/content/upload/nelson_perfetti_liben_measures_of_text_difficulty_research_ela.pdf.

Yaacov Petscher and Young-Suk Kim. 2011. The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology* 49(1):107–129.

Greg Roberts, Roland Good, and Stephanie Corcoran. 2005. Story retell: A fluency-based indicator of reading comprehension. *School Psychology Quarterly* 20(3):304–317.

Joanne K. Rowling and Jim Dale. 2016. *Harry Potter and the sorcerer's stone*. Listening Library/Penguin Random House, New York. https://usd.shop.pottermore.com/collections/audio-books/products/harry_potter_and_the_sorcerers_stone_audio_book1_english.

Paula J. Schwanenflugel, Matthew R. Westmoreland, and Rebekah George Benjamin. 2015. Reading fluency skill and the prosodic marking of linguistic focus. *Reading and Writing* 28(1):9–30.

Sarah E. Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '05, pages 523–530.

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the Python in Science Conference*. pages 57–61.

Kathleen M. Sheehan, Michael Flor, and Diane Napolitano. 2013. A two-stage approach for generating unbiased estimates of text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*. Association for Computational Linguistics, Atlanta, Georgia, pages 49–58.

Kathleen M. Sheehan, Irene Kostin, Diane Napolitano, and Michael Flor. 2014. The TextEvaluator Tool: Helping teachers and test developers select texts for use in instruction and assessment. *Elementary School Journal* 115(2):184–209.

George Spache. 1953. A new readability formula for primary-grade reading materials. *The Elementary School Journal* 53(7):410–413.

Mariët Theune, Koen Meijs, Dirk Heylen, and Roeland Ordelman. 2006. Generating expressive speech for storytelling applications. *IEEE Transactions on Audio, Speech and Language Processing* 14(4):1137–1144.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 163–173.

Miya Miura Wayman, Teri Wallace, Hilda Ives Wiley, Renta Tich, and Christine A. Espin. 2007. Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education* 41(2):85–120.

Laurence White. 2014. Communicative function and prosodic form in speech timing. *Speech Communication* 63-64:38–54.

Klaus Zechner, John Sabatini, and Lei Chen. 2009. Automatic scoring of children's read-aloud text passages and word lists. In *Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, pages 10–18.