A dynamic model of lexical experience for tracking of oral reading fluency

Beata Beigman Klebanov, Michael Suhan, Zuowei Wang, Tenaha O'Reilly
Educational Testing Service, Princeton NJ, USA
{bbeigmanklebanov,msuhan,zwang,toreilly}@ets.org

Abstract

We present research aimed at solving a problem in assessment of oral reading fluency using children's oral reading data from our online book reading app. It is known that properties of the passage being read aloud impact fluency estimates; therefore, passage-based measures are used to remove passage-related variance when estimating growth in oral reading fluency. However, passage-based measures reported in the literature tend to treat passages as independent events, without explicitly modeling *accumulation* of lexical experience as one reads through a book. We propose such a model and show that it helps explain additional variance in the measurements of children's fluency as they read through a book, improving over a strong baseline. These results have implications for measuring growth in oral reading fluency.

1 Introduction

Teaching young children the skill of reading is one of the major tasks of an education system. In the U.S., a common solution to monitoring the development of reading skill is the periodic administration of oral reading fluency (**ORF**) tests, as fluency scores can serve as indicators of early literacy skills (Biancarosa et al., 2021; Hasbrouck and Tindal, 2017; Bernstein et al., 2017; Kim and Wagner, 2015; Pikulski and Chard, 2005). For example, the popular DIBELS test is administered three times a year. A specific passage is given in a particular grade at a particular time; e.g., the passage titled *Trees* is administered in the spring of 3rd grade (Biancarosa et al., 2021, p.106). ORF is typically measured as words read correctly per minute of oral reading (**wcpm**), which accounts for both accuracy and speed (Fuchs et al., 2001). Each passage is normed so that a student's performance can be mapped to a percentile score relative to peers.

One of the weaknesses of this system of monitoring is the need to administer specific, pre-set assessment passages. First, time is taken away from reading for learning and pleasure to read for a test. Second, in striving for socio-culturally responsive assessment, one would want to give agency to teachers and students in choosing what to read, as choice and interest could enhance engagement and performance. Our reading app, RELAY READER,[1] addresses this weakness by letting students read different books aloud as a learning-and-pleasure activity and measuring ORF in the background. This solution also allows for continuous monitoring, which means that students receive frequent opportunities to demonstrate their skill.

A fundamental challenge in this endeavor is that since students read from a variety of stories throughout the year, it is not feasible to collect sufficient readings of every passage of every story to create norms. One might wonder why passage-specific norms are needed to begin with – won't students whose reading rate is 90 words per minute read any text at this rate? Alas, readers exhibit a distribution of wcpm across passages (Beigman Klebanov et al., 2020; Barth et al., 2014; Ardoin et al., 2005), as the reader's fluency is not the only factor accounting for some of the variance in the wcpm measurements.

In particular, passage effects are a known source of variance. A variety of measures proposed in the literature control for passage effects, including aspects of text complexity, genre, local discourse structure, and prosody (Beigman Klebanov et al., 2020; Barth et al., 2014). All these measures assume passages are independent – as they typically are in a testing context. However, passages in a book of fiction are not independent; there is continuity of characters and settings in a well-crafted narrative that create an immersive reading experience. This continuity could impact oral reading – while a reader might stumble on *Hogwarts* for the first time due to the word's unfamiliarity, the 50th

[1] https://relayreader.org/

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 567–575

encounter is likely to be less challenging.

It is not only the rarest words that would become less challenging when mentioned many times; repeated encounters in general are known to produce faster readings (Bell et al., 2009). Had it been the case that the first chapter introduced all the word types to be used in the book and subsequent chapters repeated those in various combinations, one would expect a steady increase in the reading pace as the reader moves through the story. Such an increase would be only partially related to the improvement in the general ORF skill of the reader, since the increase relies heavily on repetition of the same limited vocabulary and will likely disappear, at least partially,[2] when unrelated text is read.

To the best of our knowledge, little is known about the relationship between repetition and story location. We hereby pose to the community a novel challenge of modeling the dynamic of a reader's lexical experience. We offer an initial exploration and show empirical results that suggest practical usefulness of further research in this area.

2 Surprisal

The reader does not start reading *Harry Potter* with a blank lexical slate, so-to-speak; the within-the-book experience is a continuation of an ongoing lexical experience that accumulates across prior reading materials (and other language experiences, with more or less direct connection to reading). We therefore model a reader's prior knowledge using a large corpus, with the book experience viewed as an addition to the corpus – dynamically, one word at a time.[3] For every word token in the book, we use a measure of surprisal at seeing this word at this location in the book – namely, surprisal given the starting background knowledge and the within-book experience up until the current location.

In prior research, surprisal is typically defined as log of inverse of probability (Tribus, 1961), that is, for a random variable Y, the surprisal of the value $Y = y$ is given by $log_2 \frac{1}{P(y)}$. In our case, the estimate of the probability $P(y)$ for a word y is updated continuously as the student progresses through the book, token by token. Thus, words that are rare in general but frequent in the book will become less surprising as the reader moves through the book, as their estimated probability will increase. Surprisal will be highest for completely new words appearing near the end of the book – this is the first occurrence in all the experience so far (background + book). In contrast, words that are generally more frequent than in the current book would become gradually more surprising, but the increase will be small, since a frequent word has accumulated a lot of prior occurrences and the impact of any new ones is relatively small. Thus, if a book generally has a lower frequency of the word *the* than the background corpus, *the* will become more surprising as one adds the book to their lexical experience, but since even a long single book is orders of magnitude shorter than a large corpus that models the background knowledge, the book will only have a small impact on the surprisal values of generally frequent words.

3 Experiment 1: Surprisal with respect to book location

3.1 Data sources

For the current study, we use two novels – *Harry Potter and the Sorcerer's Stone* by J. K. Rowling (**HP**) and *The Adventures of Pinocchio* by C. Collodi translated from Italian by Carol Della Chiesa (**Pinocchio**) – and four background corpora, in order to observe consistency (or not) of the patterns in the two books and robustness to variation in background corpora. The background corpora are:

SFI This corpus was compiled to allow estimation of word frequencies a student might have encountered after 12 years of schooling. The corpus covers a variety of text types, including samples from high school and college text books, classical and popular literature, non-fiction, biographies, speeches, periodicals, and encyclopedias (Breland et al., 1994).

TASA3 The TASA corpus is a subset of SFI focused primarily on textbooks and other materials used in the US schools sampled by readability across grade levels (Zeno et al., 1995). Versions of this corpus have been used extensively to induce educationally relevant semantic spaces, e.g., Landauer et al. (1998). We use the cutoff for up to grade 3 readability,[4] in view of the study with 4th anbd 5th graders (Section 4).

[2] It is possible that some of those heavily repeated words will also occur in another story.

[3] If the background corpus has 5,155,569 tokens, the first token in the book will be token number 5,155,570.

[4] http://wordvec.colorado.edu/word_embeddings.html

BNC The British National Corpus (BNC Consortium, 2001) has samples of written and spoken British English from a wide range of sources from the later part of the 20th century.

SUBT This corpus is comprised of subtitles from U.S. films from 1900–2007 and U.S. television series (Brysbaert and New, 2009).

We use pre-existing unigram counts for each of the corpora, either as raw counts (for BNC, SUBT, TASA3) or deriving the probability estimates from the standard frequency indices (SFI) using the reversed estimated-to-standard frequency transformation[5] and the published total corpus sizes to induce estimated counts. Table 1 shows information about the various corpora.

Corpus	# tokens	# types (unique tokens)
TASA3	2,692,335	32,732
SFI	14,418,651	94,563
BNC	100,136,361	537,729
SUBT	49,719,560	73,609

Table 1: Corpora used to model prior lexical experience.

3.2 Data pre-processing

All background corpora were pre-processed to normalize British/American spelling and handle contractions and hyphenation. The tokenization process used for generating the unigram counts differed somewhat across corpora and we generally followed the tokenization practice of the given corpus when tokenizing the book as a continuation of experience following that corpus. For example, *can't* corresponds to two tokens *can n't* in BNC, whereas SFI only retains *can* as a token.

The next step is turning a book into a series of passages. Each book is split into consecutive passages of approximately 250 words (about one page): We add paragraphs to a passage as long as the total word count is under 250 words. Whether to add the next paragraph into the passage depends on whether there is a larger absolute difference from 250 with or without adding it. Thus, passages always contain full paragraphs. Passages do not cross chapter boundaries; if the last passage of a chapter is very short – less than 50 words – we discard it. Four chapter-final passages were discarded for HP and three for Pinocchio. Table 2 shows the descriptive statistics of the book data.

Book	# chapters	# passages	passage length mean (std)
HP	17	315	246.83 (28.82)
Pinocchio	19	162	241.75 (44.04)

Table 2: Descriptive information for the book data.

3.3 Measures

To represent surprisal patterns in a given passage, we experiment with four measures. We use average, median, and standard deviation (stdev) of token-level[6] surprisal estimates per passage and a high-percentile (97%) cut-off that captures the extent of surprisal of a few of the most surprising words in the passage. We expect the 97-percentile to capture invented or rare vocabulary – exactly the kind of words for which we expect the most impact upon multiple within-book encounters. Table 3 shows words above the 97% cut-off for three passages in the beginning, middle, and end of HP and Pinocchio, including surprisal estimates for each word using SFI as the background corpus.

3.4 Research questions

Our research questions are as follows. First, is it the case that the overall dynamic of surprisal within the book tends towards lower surprisal later in the book? Second, do we observe consistent patterns across (a) the two books, and (b) the different background corpora? If the patterns vary dramatically across corpora, this would underscore the need to model the target user's prior reading profile in a more precise and personalized manner.

3.5 Results

Table 4 shows Pearson's correlations between the surprisal measures and the serial number of the passage in the book. Our first research question is answered in the affirmative – it is the case that

[5]We use the formula SFI = 10 ($\log_{10}$ U + 4), where SFI is the standardized frequency index and U is the estimated frequency per millions words using dispersion D = 1, following the definition in Terzopoulos et al. (2017), which differs slightly from that offered in Breland et al. (1994). SFI is the name of the standardized index and also of one of our corpora, since the paper that introduced the corpus was also the one to introduce the index (Breland et al., 1994).

[6]If a word occurs multiple times in a passage, each occurrence will get a slightly different surprisal value – a later mention would incorporate the experience of having seen the word earlier in the passage as well as of not having seen it since that prior mention; see examples in Table 3.

Loc	HP		Pinocchio	
	Word	Surp.	Word	Surp.
early	dursley	21.20	geppetto	22.78
	dursley	20.97	polendina	22.78
	dudley	22.20	antonio	22.78
	dudley	21.78	geppetto	22.20
	dudley	21.46	geppetto	21.78
			antonio	22.20
			polendina	22.20
			polendina	21.78
middle	hermione	18.98	tremble	16.69
	hermione	18.93	dolphin	16.74
	overhearing	19.54	marionette	16.99
	gryffindor	19.20	dolphin	16.73
	seamus	20.33	gait	17.28
	filch	19.62	fro	17.41
	sneering	18.93	idle	16.86
			tramp	17.59
late	wardrobes	20.79	snail	17.0
	greener	19.54	lizard	16.06
	tidier	21.79	bravo	18.62
	bertie	21.79	mischief	16.65
	bott	21.79	deserve	15.96
	muggle	19.40	praise	15.91
	wizened	19.54	models	16.39
	muggles	19.54	obedience	17.64

Table 3: Words in early, middle, late HP and Pinocchio passages that are the top 3% surprisals for the passage. Words are listed in their book order: *Dursley* in row 2 occurs later in the passage than *Dursley* in row 1.

surprisal trends downwards as one moves through the book, for the two books and the four measures.

Measure Corpus	mean HP/P	median HP/P	stdev HP/P	97% HP/P
TASA3	-.21/-.18	-.11/-.11	-.26/-.24	-.22/-.20
SFI	-.14/-.16	-.14/-.06	-.26/-.28	-.32/-.42
BNC	-.13/-.16	-.11/-.06	-.18/-.25	-.21/-.44
SUBT	-.16/-.14	.-16/-.08	-.11/-.23	-.29/-.28

Table 4: Pearson's correlations between book location (serial number of the passage in the story) and surpisal measures. In each cell, the value for Harry Potter is shown first (HP), followed by Pinocchio (P).

To address the question of robustness towards variation in background corpora, Table 4 shows that the trends are generally similar across the four corpora. Figure 1 exemplifies the trends. The corpora are in agreement regarding the general trajectories even if the exact estimates of surprisal are different. Surprisal values are generally higher for the larger corpora, since the occurrence of new words is more surprising with more background experience. Interestingly, for HP, it is not the case that chapter 1 is consistently more surprising than the rest; chapters 5 (*Diagon Alley*) and 7 (*The Sorting Hat*) are more surprising. This makes sense with respect to the story – while some of the "normal" (*muggle*)

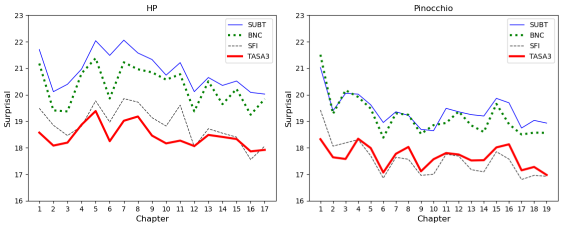

Figure 1: Average 97-percentile surprisal values per chapter across background corpora.

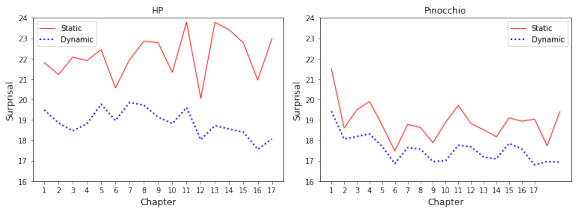

Figure 2: Average 97-percentile surprisals by chapter using static and dynamic SFI-based computations.

characters like the Dursley family are introduced in chapter 1, it is not until chapter 5 that the immersion in a very different, magic world happens, which is accompanied by a lot of rare or invented vocabulary related to magic artifacts (ch 5) and houses, teachers, and classes in a school of magic (ch 7). In contrast, the pattern for Pinocchio does show a drop after chapter 1, with more minor ups and downs later in the story.

To appreciate the difference between the measures discussed here and a 'static' surprisal calculation based on the background corpus only, without the dynamic recalculation following the token-by-token reading experience, Figure 2 plots the 97-percentile measure using the SFI background corpus. Without accounting for the within-book experience, some later chapters in HP have extremely high surprisal scores (chapters 13 and 14). The dynamic index shows, in contrast, that by that point in the story, life in a school of magic is somewhat business-as-usual, with these chapters being part of the general downwards trajectory. The discrepancy between the static and dynamic measures for the later HP chapters is such that the overall correlation with book location is actually *positive* for the static 97-percentile measure for all background corpora – in contrast to the universally negative correlations reported in Table 4 for the dynamic measures.

For the next experiment with 4th and 5th grade students, we used the TASA3 corpus to model background knowledge.

4 Experiment 2: Modeling fluency

4.1 Data

The oral reading data come from 35 students in grades 4 (12) and 5 (23) in an elementary school in New Jersey.[7] Students read on Amazon Fire 7 tablets with the RELAY READER app (previously called MY TURN TO READ, Madnani et al. (2019)) for up to 19 weeks, approximately three times a week for 20 minutes at a time, during the time generally set aside in the curriculum for independent reading. All the 35 students finished HP; those who finished earlier were provided the next book in the series in the paperback format. The students used consumer-level in-ear headphones with a built-in microphone.

When reading with the app, students took turns reading out loud consecutive passages of the book with a pre-recorded audiobook narrator. When splitting the text of a chapter into reading turns for the reader and the narrator, an algorithm as described in section 3.2 is used, with the target of 150 words per student turn and 200 words per narrator turn. The splitting is dynamic in that when the child first logs in on a given day or starts a new chapter, the narrator reads first starting from the current location, no matter who read last in the previous session, to ease the reader into the activity. Since students read at different rates, the daily starting locations varied and so did the passages read.

A set of 1,529 recordings with as many readers as possible per passage that span the beginning, middle, and end of each of the chapters were selected for the analysis, 67 passages in total with 100-170 words per passage (mean = 149.9, std = 17.5). Each reader contributed 13-64 readings (mean = 43.7, std = 13.2) and each passage was read by 15-33 children (mean = 22.8, std = 4.9). There were 60-111 recordings per chapter (mean = 90, std = 13). The recordings were transcribed by a professional agency. The transcribers were provided with the text of the passage and were asked to indicate any deletions, substitutions, and insertions as well as provide timestamps for the beginning and end of on-task speech. We then used the transcriptions to compute wcpm (the number of correctly read words divided by the time in minutes it took the child to read the passage).

[7] See the Ethics Statement for more detail.

4.2 Models

We now move to evaluating whether surprisal explains additional passage-based variance in wcpm, above and beyond baseline predictors. We fit linear mixed models using R's *lmer* function.

As a baseline, we use the model from Beigman Klebanov et al. (2020) where wcpm is modeled as a combination of passage and student random effects and a number of fixed effects: (1) the grade level of the student (to capture any systematic differences between grades); (2) a text complexity score produced by Text Evaluator (TE) (Napolitano et al., 2015); (3) a words-per-minute measurement of a "reading" generated by Apple's text-to-speech synthesizer (the Alex voice) to model variation in duration of different phonemes and reasonable inter- and intra-sentential pausing (TTS), and (4) the number of the chapter the passage is in. In the Beigman Klebanov et al. (2020) analysis, the coefficient of the chapter variable captures the average extent of improvement in oral reading fluency per chapter. Chapter is also used as a random slope to allow for different growth rates across participants. The model is specified using *lmer* syntax in equation 1; the coefficients are shown in the "Baseline" column of Table 5.

$$wcpm \backsim (1|passage) + (chapter|student) + \\ + grade + TE + TTS + chapter \quad (1)$$

We next fit a model that is identical to the Baseline but has an additional fixed effect – the stdev of the surprisal values per passage, using the TASA3 corpus as background. The coefficients are shown in the "+Surprisal" column in Table 5. We show results with stdev index since models with 97% and mean did not converge and the model with median showed a similar pattern of results but worse fit than the model with stdev.

Table 5: Model estimates (with standard error). The values for TTS, TE, and Surprisal were standardized to $\mu = 0$ and $\sigma = 1$ and then entered into the model.

	Dependent variable: wcpm			
	Baseline		+Surprisal	
Grade 5	-0.83	(8.95)	-0.70	(8.95)
TTS	4.72***	(0.94)	3.37***	(0.89)
TE	-3.05**	(0.92)	-1.86*	(0.85)
Chapter	1.27***	(0.26)	1.09***	(0.25)
Surprisal			-3.39***	(0.76)
Constant	99.96***	(7.54)	101.41***	(7.51)
Note:			*p<0.05; **p<0.01; ***p<0.001	

We observe that surprisal is a significant predictor of wcpm, after controlling for complexity and prosody, with higher surprisal corresponding to slower reading. The Baseline model puts the amount of passage-based unexplained variance at 22.4; the number is reduced to 13 in the Baseline+Surprisal model, a reduction of 42%.

We also observe that the estimated rate of growth is somewhat reduced, from 1.27 additional wcpm per chapter to 1.09. This extent of growth is predicted after controlling for the within-book repetition of key book-specific vocabulary, so it might allow for a better estimate of the more generalizable part of the growth in fluency.

5 Related work

There exists a substantial body of work investigating the relationship between stand-alone properties of passages and the speed of reading. Beigman Klebanov et al. (2020) showed evidence that text complexity and prosody explain variance in children's wcpm. Barth et al. (2014) reviewed a variety of indices used to characterize the language in passages and found that text complexity, narrativity (the extent to which a passage is story-like rather than informational), and referential cohesion were predictive of wcpm, with complexity entering with a negative coefficient, while narrativity and cohesion enter the model with positive coefficients. Referential cohesion quantifies "the extent to which words overlap across sentences in the text" and is thus capturing an aspect of local, sentence-to-sentence predictability. A related but even more localized notion of predictability – within sentences rather than across sentences – was found to predict speedup in silent reading in adults; both syntactic and lexical immediate contexts were significant predictors (Monsalve et al., 2012). Given the findings, it is possible to manipulate the difficulty of a story by, for example, susbtituting shorter words instead of longer words or by repeating words across sentences.[8] These would, however, alter the language of the story and could reduce its literary quality and authenticity. In contrast, surprisal can be manipulated without changing the language by sequencing stories – having the first Harry Potter book in your prior reading experience would make a lot of the vocabulary in the second book less surprising.

[8]Indeed, text complexity is an explicit and quantitative design principle when creating texts for ORF assessments: "The Spache readability formula was used in creating and revising passages" (Good and Kaminski, 2002, p.3).

Another related body of literature is the work on modeling word frequency distributions (Piantadosi, 2014; Baayen, 2001; Katz, 1996). In particular, the finding that various types of corpora, including single books, tend to exhibit certain consistent large-scale patterns of keyword burstiness is promising for generalization of findings such as ours across books (Altmann et al., 2009; Sarkar et al., 2005; Montemurro and Zanette, 2002).

The extensive work on language modeling in NLP, including the advances achieved with transformer models, can be brought to bear on modeling surprisal at various granularities (word, sentence, passage) and given various types of prior experience (model pre-training, fine-tuning). Furthermore, the assumption that an encounter results in reduction in surprisal for that word only is an over-simplification, as the literature on associative and semantic priming suggests that related words are also somewhat activated (Pickering and Gambi, 2018; Plaut and Booth, 2000; Masson, 1995). Transformer models were recently shown to exhibit certain priming effects themselves (Lindborg and Rabovsky, 2021; Misra et al., 2020), making them a promising basis for modeling surprisal while accounting for priming effects. Our work with a word-level dynamic surprisal is just a first step.

6 Conclusion

In this paper, we presented a new NLP challenge coming out of the need to estimate the latent skill of oral reading fluency based on measurements of words read correctly per minute as readers move through a book using our electronic book reading app. Since the measurements are known to systematically depend on the properties of the passage, it is important to control for the passage-based variance in order to produce more precise skill estimates.

In particular, work presented here suggests that it is not only stand-alone properties of reading passages that are implicated in explaining slow-downs or speed-ups in oral reading, but also properties of a particular passage that have to do with its specific position in the reader's overall reading experience. As the reader reads through a book, they become more familiar with the special (invented or rare) vocabulary used in the book; this, in turn, could result in a speed-up in the reading. While the reader might be having an experience of increasing flu-

ency, some of the gain might be book-specific and therefore not generalize to the next book the developing reader tackles. Accurate tracking of oral reading fluency – a foundational reading skill that is a robust predictor of other skills such as comprehension – is a practical issue that will be helped by further research into dynamic models of a reader's lexical experience.

Limitations

The limitations of the findings in experiments 1 and 2 have to do with the relatively small scale of the study. We experimented with two books and, while the findings were broadly consistent, it could be that results would not generalize to other books. Experiment 2 was conducted with a specific group of readers in a specific context of implementation; studies with additional groups of readers are needed to evaluate generalization of the findings.

Another limitation of our experiments is that the dynamic model of lexical experience is evaluated only as an aggregate index per passage and not as a predictor for specific words or types of words. In particular, the model predicts a slight increase in surprisal of function words if their density in the story is generally lower than in the background corpus. This assumption may or may not be correct; further experimentation is necessary to evaluate the surprisal model in more detail. We thank a BEA reviewer for pointing out this limitation.

Ethics Statement

RELAY READER, the reading app discussed in this paper, specifies Terms of Use and provides a link to Privacy Policy. In particular, the Terms of Use specify the legitimate uses of the data and commits to keeping the data of users-in-the-wild anonymous.[9]

For the book data, we used a public domain text of *The Adventures of Pinocchio* from Project Gutenberg and the text of *Harry Potter and the Sorcerer's Stone* provided to us by the copyright holder[10] as a part of a license to use the book and the audiobook narration by Jim Dale in the app for a specified limited number of students; the students whose data is analyzed in Experiment 2 are within that cap.

The corpora used in the study are either broadly available for research purposes (BNC, SUBT) or have a more limited research and/or operational availability through contracts (TASA3, SFI).

The study during which oral reading data was collected from grade 4 and 5 students in a school in New Jersey was approved by the Institutional Review Board at our organization. Parental consent was obtained for students' participation in the activity and for use of students' data (including recordings, log data of the reading activity, and demographic information provided by the parents such as the grade data used in this study) for research.

The goal of this research is to improve the quality of assessment of oral reading by identifying factors that could impact fluency measurements that are not entirely due to the students' developing skill and build models that would allow compensating for the impact of such factors. Accurate assessment of oral reading fluency controlling for text effects will benefit teachers and students in that the assessment can be done on a variety of texts, including different passages for different students, instead of using a single pre-set normed passage as in the current practice. This would give both teachers and students more agency in selecting reading materials based on interest and preference and will thus help assessment to be more socio-culturally responsive while still providing the measurement signal necessary to monitor skill progression.

References

Eduardo Altmann, Janet Pierrehumbert, and Adilson Motter. 2009. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLOS one*, 4(11):e7678.

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

Harald Baayen. 2001. *Word frequency distributions*, volume 18. Springer Science & Business Media.

Amy Barth, Tammy Tolar, Jack Fletcher, and David Francis. 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.

Beata Beigman Klebanov, Anastassia Loukina, JR Lockwood, Van Rynald Liceralde, John Sabatini, Nitin Madnani, Binod Gyawali, Zuowei Wang, and Jennifer Lentini. 2020. Detecting learning in noisy data:

[9] https://relayreader.org/terms

[10] We did not alter anything in the HP book. For Pinocchio, we re-chaptered the original 36 short chapters of the story that we downloaded from Project Gutenberg into 19 longer chapters in order to better adjust to the turn-taking setup of RELAY READER.

The case of oral reading fluency. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 490–495.

Alan Bell, Jason Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111.

Jared Bernstein, Jian Cheng, Jennifer Balogh, and Elizabeth Rosenfeld. 2017. Studies of a Self-Administered Oral Reading Assessment. In *Proceedings of SLaTE 2017*, pages 180–184, Stockholm. KTH Royal Institute of Technology.

Gina Biancarosa, Patrick Kennedy, Sunhi Park, and Janet Otterstedt. 2021. 8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS®): Administration and Scoring Guide. Technical report, University of Oregon.

BNC Consortium. 2001. The British National Corpus, version 2 (BNC World).

Hunter Breland, Robert Jones, and Laura Jenkins. 1994. The college board vocabulary study. *College Board Report; Educational Testing Service Research Report*, 94(26).

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Lynn Fuchs, Douglas Fuchs, Michelle Hosp, and Joseph 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 Good and Ruth Kaminski. 2002. DIBELS Oral Reading Fluency Passages for First through Third Grades. Technical report, University of Oregon, Eugene, OR.

Jan Hasbrouck and Gerald Tindal. 2017. An update to compiled ORF norms. Technical report, Behavioral Research and Teaching, University of Oregon.

Slava Katz. 1996. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2(1):15–59.

Young-Suk Kim and Richard Wagner. 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.

Thomas Landauer, Peter Foltz, and Darrell Laham. 1998. An introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2-3):259–284.

Alma Lindborg and Milena Rabovsky. 2021. Meaning in brains and machines: Internal activation update in large-scale language model partially reflects the N400 brain potential. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Nitin Madnani, Beata Beigman Klebanov, Anastassia Loukina, Binod Gyawali, Patrick Lange, John Sabatini, and Michael Flor. 2019. My turn to read: An interleaved E-book reading tool for developing and struggling readers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 141–146, Florence, Italy. Association for Computational Linguistics.

Michael Masson. 1995. A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(1):3.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring bert's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635.

Irene Monsalve, Stefan Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408.

Marcelo Montemurro and Damián Zanette. 2002. Entropic analysis of the role of words in literary texts. *Advances in complex systems*, 5(01):7–17.

Diane Napolitano, Kathleen Sheehan, and Robert Mundkowsky. 2015. Online readability and text complexity analysis with TextEvaluator. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 96–100, Denver, Colorado.

Steven Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130.

Martin Pickering and Chiara Gambi. 2018. Predicting while comprehending language: A theory and review. *Psychological Bulletin*, 144(10):1002.

John Pikulski and David Chard. 2005. Fluency: Bridge between decoding and reading comprehension. *The Reading Teacher*, 58(6):510–519.

David Plaut and James Booth. 2000. Individual and developmental differences in semantic priming: empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review*, 107(4):786.

Avik Sarkar, Paul Garthwaite, and Anne De Roeck. 2005. A Bayesian mixture model for term re-occurrence and burstiness. In *Proceedings of the 9th Conference on Computational Natural Language Learning*, pages 48–55.

Aris Terzopoulos, Lynne Duncan, Mark Wilson, Georgia Niolaki, and Jackie Masterson. 2017. HelexKids: A word frequency database for Greek and Cypriot primary school children. *Behavior Research Methods*, 49(1):83–96.

Myron Tribus. 1961. Information theory as the basis for thermostatics and thermodynamics. *Journal of Applied Mechanics*, 28(1):1–8.

Susan Zeno, Stephen Ivens, Robert Millard, and Raj Duvvuri. 1995. *The educator's word frequency guide*. Touchstone Applied Science Associates.