Detecting learning in noisy data: The case of oral reading fluency

Beata Beigman Klebanov
Anastassia Loukina
John Lockwood
{bbeigmanklebanov,aloukina,jrlockwood}@ets.org
Educational Testing Service
Princeton, New Jersey

Van Rynald T. Liceralde
University of North Carolina and Educational Testing Service
Chapel Hill, North Carolina
vliceralde@ets.org

John Sabatini
University of Memphis
Memphis, Tennessee
jpsbtini@memphis.edu

Nitin Madnani, Binod Gyawali
Zuowei Wang, Jennifer Lentini
{nmadnani,bgyawali,zwang,jlentini}@ets.org
Educational Testing Service
Princeton, New Jersey

ABSTRACT

In a school context, learning is usually detected by repeated measurements of the skill of interest through a sequence of specially designed tests; in particular, this is the case with tracking improvement in oral reading fluency in elementary school children in the U.S. Results presented in this paper suggest that it is possible and feasible to detect improvement in oral reading fluency using data collected during children's independent reading of a book using the Relay Reader™ app. We are thus a step closer to the vision of having a child read for the story, not for a test, yet being able to unobtrusively assess their progress in oral reading fluency.

CCS CONCEPTS

• **Applied computing** → **Education**; **Interactive learning environments**; • **Information systems** → **Data analytics**; • **Human-centered computing** → Empirical studies in HCI.

KEYWORDS

children's reading ; fluency ; oral reading fluency ; reading app; book reading ; reading analytics;

ACM Reference Format:
Beata Beigman Klebanov, Anastassia Loukina, John Lockwood, Van Rynald T. Liceralde, John Sabatini, Nitin Madnani, Binod Gyawali, and Zuowei Wang, Jennifer Lentini. 2020. Detecting learning in noisy data: The case of oral reading fluency. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge (LAK '20), March 23–27, 2020, Frankfurt, Germany*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3375462.3375490

LAK '20, March 23–27, 2020, Frankfurt, Germany

ACM ISBN 978-1-4503-7712-6/20/03…$15.00
https://doi.org/10.1145/3375462.3375490

1 INTRODUCTION

In institutional educational contexts, learning has traditionally been and largely still is tracked through repeated measurements of the skill of interest through a sequence of specially designed tests [22, 23]. In particular, this is the case for tracking the development of oral reading fluency (**ORF**) in elementary school students [6–8, 10, 19].

Is it possible to track improvement in ORF using data collected during the child's independent reading of a book using a reading app? If possible, this would allow the child and the teacher to focus on the target activity, while the tracking would be happening unobtrusively, in the background of the actual reading for learning and enjoyment. It would also potentially allow tracking at much finer granularity than feasible with repeated administration of tests.

We describe Relay Reader™, an application designed to facilitate a developing reader's access to a full-length book by providing the story in a turn-taking format, where the child and a pre-recorded skilled adult narrator take turns reading out loud consecutive passages of the book. The child's oral reading is recorded and stored for further analysis. For the analyses reported here, we use oral reading data collected from upper elementary school children reading *Harry Potter and the Sorcerer's Stone* (**HP1**) on an approximate schedule of 20 minutes a day three times a week during the Drop Everything And Read (DEAR) curriculum slot.[1]

There are three types of challenges in tracking ORF using the continuous measurements from Relay Reader. First, there is evidence that reading rates depend on the text one is reading [28], hence the standardized materials in ORF tests. In contrast, passages in a book may vary widely in textual features. Therefore, there is a need to disentangle growth from fluctuations due to the variability in the passages, for a given child. Second, different participants may be reading different passages of the book, as the assignment of narrator and child turns is dynamically re-calculated at the beginning of each reading session. Third, there is very substantial technical and behavioral noise affecting the recordings – for each passage, the child might have read all or part of the passage, read more

[1] See the report of the National Reading Panel on Teaching Children to Read [20] pages 3-21 - 3-33 for a detailed exposition and analysis of practices related to encouraging independent silent reading in U.S. school curriculum.

or less clearly or loudly, with equipment more or less intact, and recording conditions more or less favorable. Our question therefore is – can such data support growth analysis?

2 RELAY READER APP

Relay Reader™ is a reading and listening app designed to help developing readers successfully engage with long and potentially challenging books with the twin goals of helping them (1) enjoy the experience and therefore improve their attitude towards reading, and (2) improve their reading skill, particularly reading fluency, through sustained oral reading and listening. With this app, the child takes turns reading out loud from a high interest book with a pre-recorded skilled adult narrator (audiobook). The book text is split into consecutive passages of pre-defined average length where passage boundaries correspond to paragraph breaks. In this study, the narrator turns were set to 200 words on average, the child's turns were set to 150 words on average.[2]

When the child logs into the app, the reading session always starts with a narrator turn (even if they finished with listening to the narrator the day before).[3] During a narrator turn, the text is highlighted at phrase level so that the child could follow the narrator on the screen. Then the app prompts the child to read his or her passage. When the child indicates that he or she finished reading, the app plays the part of the audiobook that corresponds to the subsequent passage for narrator turn, followed by another child's turn, and so on throughout the book. Implementation details of the app, such as processing of ebook and audiobook to align text to audio, identification of phrases for the moving highlight, as well as additional features of the app such as reading comprehension questions are described in [16].

We implemented the reading program with the app in a 4th and a 5th grade classes in an elementary school in New Jersey. In total, 42 students participated in the program: 17 were 4th graders; 21 were males. Each child was provided with an Amazon Fire 7 device on which the app was installed, and a pair of consumer-level in-ear headphones with a built-in microphone. The total cost of equipment per child was $55. Before the start of the program, the teachers were trained by the project staff to use the app. The reading program itself was conducted entirely under teacher supervision and monitored remotely by the project staff through the analysis of the app logs. Children read in their regular classroom, with some care taken to disperse them as much as possible in the classroom; teachers also sent some students to read in the corridor adjacent to the classroom. Children read during approximately the same time slots 3-4 times a week, for about 20 minutes a day, until they finished the book. Since children's reading speeds differed, different children finished the book at different times; when a child finished reading HP1 with the app, he or she was given a paperback copy of the next book in the series which they read silently. Recordings were collected between October 2018 and February 2019. 95% of the students finished the book within 15 weeks.[4] Children reported strong enjoyment and engagement: In the exit survey, 93% responded with *agree* or *strongly agree* to "I liked reading with the app" and 83% responded with *disagree* or *strongly disagree* to "The book was boring."

[2]The turn length for child and narrator can be altered by the child in the current version of the app, but not in the version used in this study. By teachers' request, we could assign specific students to another schedule, instead of the default 200/150.

[3]The calculation of reading turns for child and narrator is dynamic.

[4]The first three students to finish the book did so within 6 weeks; one reader did not finish the book by the end of the program in February 2019.

3 MEASUREMENT OF ORF

ORF is typically measured as words read correctly per minute (**WCPM**). Students read aloud from a passage for 1 min and the number of words read correctly is counted [6]; alternatively, the total number of correctly read words in the passage is divided by the total duration of reading in minutes [4]. Omissions, substitutions, mispronunciations are marked as errors [28]; insertions and long hesitations are penalized indirectly through duration. For the popular DIBELS test [6], students read three passages and the median score is reported as the score for the test;[5] passages are grade-level controlled for comprehension difficulty. A similar practice is reported for the Moby.Read computer-based test of ORF [4].

In standardized assessments such as DIBELS, an examiner, usually a teacher, times the reading, marks the errors and then computes WCPM. To scale the scoring up, speech technologies have been previously successfully used for automated scoring of oral reading fluency, usually under quiet, monitored conditions [1, 4, 17]. The feasibility of using speech technologies for automated computation of WCPM for classroom-recorded readings was explored in [13]. They reported, as expected, that recordings contained a substantial amount of behavioral and acoustic noise; still, the correlation between human and automated estimates was $r = 0.8$ at the level of an individual reading passage, increasing to $r = 0.93$ when multiple measurements were aggregated for the same reader.

In this study, our research question is whether the data collected through Relay Reader can be used to track the improvement in fluency despite variation in passages and the quality of the recordings. To minimize the effect of other factors such as errors of the automated speech recognition, we first perform the analysis on manual transcriptions of the data. We then evaluate whether the results hold for the automated measurements.

4 DATA FOR GROWTH ANALYSIS

4.1 Tracking time

The first question for growth analysis is how to track time. Due to absences and difference in reading rates and time management, different kids could read a very different number and set of passages on any given day, and since better readers would tend to finish the book faster, a confounding relationship between reader skill and time is likely, especially towards the end of the reading program. However, we note that all readers, weak and strong, had to read the book's chapters in the sequence written by the author, since Relay Reader does not allow jumping ahead. Furthermore, all participants included in our study but one have reached the end of the book which consists of 17 chapters.[6] A chapter, therefore, would be a unit of analysis that is both guaranteed to align with the progression of time and will not be confounded by reader skill. We will model progression of time by chapter.

[5]The newest version of DIBELS published in July of 2019 uses a single passage. https://dibels.uoregon.edu/docs/materials/d8/dibels_8_admin_and_scoring_guide_07_2019.pdf

[6]On average HP1 chapters are 4,500 words long, ranging from 3,055 to 6,570.

4.2 Selection of readers

In total, 42 children took part in the program and 41 of them finished the book. For 7 of these children (including the one who didn't finish the book), the duration of the student turn was adjusted in consultation with the teacher to be 70 words rather than the default 150 words. A consistent difference in average passage length could have a systematic effect on WCPM and text-based measurements used in this study. Since the reduction in passage length was done to accommodate some of the weaker readers, including their recordings would confound passage and reader characteristics. Finally, since children on the reduced schedule almost always read different passages from the rest of the group, recordings from these readers exacerbate the sparsity of the passage-by-reader matrix (more on this in section 4.4). To avoid complicating the model and introducing confounds, we retained only 35 children who read on the default schedule for further analysis.

4.3 Filtering of recordings

During the 19 weeks of the data collection, we logged 7,849 reading turns from 35 students (12 in Grade 4 and 23 in Grade 5). Of these, the recordings for 344 turns (4.4%) were unusable for technical reasons: The audio was missing or corrupted, or, if present, the audio was completely quiet, suggesting microphone failure. For additional 669 turns (8.5%), the duration of the recording was too short to contain meaningful reading; usually this happened when the student skipped most of the reading [3]. Finally, we excluded 357 recordings (4.5%) where the total number of words in the passage to be read was less than 100. Such shorter passages sometimes happened at the end of the chapter. After applying these filters, we were left with 6,479 recordings, 82.5% of the original data.

4.4 Selection of passages

Transcription of all 6,479 recordings would be a substantial investment, beyond the financial scope of this project. Our next step is therefore to select a smaller sample of recordings while still retaining sufficient data for a growth analysis.

The common practice in ORF tests is to use 3 passages per test to obtain a reliable estimate [4, 6]. We used this as a guideline to select sufficient number of passages from every chapter. We picked 68 passages – 4 in every one of the book's 17 chapters – as follows. To allow for a balanced coverage of the chapters, we split each chapter into beginning (first 30%), middle (30-70%) and end (last 30%) using word count as a measure. Then, for each chapter, we selected 1 passage from the beginning of the chapter that was read by the largest number of students;[7] 2 passages read by most students from the middle of the chapter; and 1 passage read by most students from the end of the chapter. Average passage length was 150 words (min = 100, max = 210, sd = 17.8).

The final transcribed corpus consisted of all recordings that we had for the 68 passages that passed the filters described in section 4.3, for the total of 1,556 recordings (20% of all data). There were on average 2.6 recordings per chapter per reader. Each reader contributed 13-65 readings (mean = 44.4, sd = 13.3) and each passage was read by 15-33 children (mean = 22.9, sd = 4.9). The reader by passage matrix used in the subsequent analysis was 65% full.

[7] We omitted the first passage assigned to the students in chapter 1 even though it was read by all students to account for the fact that students were still getting used to the app when reading this passage.

5 MODELING GROWTH IN ORF

5.1 Data preparation

All recordings selected as described in 4.4 were transcribed by a professional transcription 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.[8] The transcribers were not aware of the goals of this study and were not explicitly given any chapter information. We then used the transcriptions to compute WCPM, the total number of correctly read words divided by the total time it took the child to read the passage based on the timestamps for the beginning and end of on-task speech.

5.2 Controlling for text effects

In the context of standardized testing of children's ORF, it is a strongly held belief that comprehension complexity of the text impacts the fluency of its reading by the child, so that texts that are more difficult to comprehend tend to be read more slowly. For example, the Flesch-Kincaid readability formula is used for estimating difficulty (grade-level) during selection of passages for the DIBELS test [21, p27]. It is also known, however, that controlling for readability alone does not eliminate passage-related variance in children's oral reading fluency; hence, field testing of the items is recommended to pick passages with close average WCPM from the grade-level-controlled pool [24]. Since reading aloud requires actually pronouncing the text, the reading rate is also likely to be subject to prosodic and articulatory constraints imposed by the text [9, 29]. In a recent study of reading rate in adults, it was found that such durational effects were a strong predictor of adults' reading rates [14]. We use scores from TextEvaluator (**TE**) [18], a state-of-the-art measure of comprehension complexity of a text, to estimate difficulty. TextEvaluator assigns each text a score between 100 and 2000, with more complex texts getting a higher score. To capture phonetic constraints, we follow [14] and use state-of-the-art text-to-speech (TTS) synthesis which includes a model for estimating phone durations given the prosodic and articulatory constraints imposed by the text.[9] We then use these to compute the expected reading rate (**TTS**).

5.3 Growth estimation

To evaluate whether there is a consistent increase in WCPM as students progressed through the book, we use a statistical approach described in [26] to test whether the chapter to which a passage belongs is a significant predictor of WCPM for that passage.

Our data has a hierarchical cross-classified structure where multiple recordings are grouped both by student and by passage. We used mixed linear models as implemented in `lme4` [2] package in

[8] Recordings sometimes include conversations with other children before the child settled down to reading the turn, or teacher's instructions to finish the session given to the class when the child has just finished the reading but hasn't pressed the button yet, or re-reading out loud of some of the narrator's turn that happens before the child starts reading his or her own turn. All these would be marked as off-task speech.

[9] We use Apple's built-in TTS engine (OS X 10.14.5) and male Alex voice.

R [25] to control for such. The dependent variable was WCPM for each recording. To account for variation between students and passages, we entered student and passage as random factors. `lmerTest` [11] package was used to estimate significance.

We first fit the null model with random factors only. The `lme4` syntax for the model is shown in equation 1.

$$wcpm \sim (1|passage) + (1|student) \quad (1)$$

The variance partitioning analysis [27] showed that the model attributed 56% of the variance to differences between students and 6% to differences between passages. This is consistent with previous studies which showed that most variance in WCPM in elementary school chidlren comes from the difference in skill rather than between-text variability. For example, [24] used Generalizability theory for variance partitioning and found that 10% of variance was attributed to text and 81% to student.

We next considered whether chapter is a significant predictor of WCPM. To do this, we entered chapter as both a fixed factor and random slope to allow for different effect for different students. We also entered `grade` as fixed variable to account for differences between the two grades represented in our data, and `TTS` and `TE` as fixed variables to account for passage-related variance.

The full model equation is given in Eq. 3 in the Appendix. Eq. 2 shows the equation in `lme4` syntax. The model was fitted using Nelder-Mead optimizer.

$$\begin{aligned} wcpm \sim (1|passage) + (chapter|student) + \\ + grade + TE + TTS + chapter \end{aligned} \quad (2)$$

The model estimates are shown in Table 1. They showed significant positive linear effect of chapter: With each chapter average WCPM would be expected to increase by about 1.3 words per minute (after controlling for other factors). One of the possible alternative (that is, a non-growth) explanations for this result might be that the textual properties of passages change as the child progresses through the book, so that the faster-to-read passages tend to appear later in the book. To test for this, we computed the correlation between the two passage-based measure and the order in which the passage occurs in the book. We found that the correlations were r = -0.17 for TTS and r = 0.04 for TE. This is further illustrated in Figure 1 which shows the average values by chapter for the two text measure and WCPM across all students. The figure demonstrates a steady increase in average WCPM. While there is a lot of variability in the two text measures across the chapters, there is no consistent pattern related to progress through the book. In other words, the growth in WCPM that we observed cannot be attributed to systematic change in text complexity or prosodic properties of the text as one moves along.

The analysis of deviance (likelihood ratio test) showed that this model was a better fit to the data (LogLik -7055 vs. -7027, $p < 0.0001$) than the model that did not allow for different growth rate across students; that is, the rate of improvement differed across students. Further analysis showed that while the estimated slopes were negatively correlated with fluency estimates (r = -0.3), this correlation was not statistically significant. In other words, based on this data we do not have sufficient evidence that the rate of improvement is higher for less proficient students.

Table 1: Model estimates for the model in Eq. 2. The numbers in parenthesis show standard error. The values for TTS and TE were standardized to μ =0 and σ =1, then entered into the model.

	Dependent variable:
	WCPM
Grade5	-0.733 (8.909)
TTS	4.841*** (0.932)
TE	-3.036*** (0.911)
Chapter	1.278*** (0.257)
Constant	99.959*** (7.510)
Observations	1,556
Note:	*p<0.1; **p<0.05; ***p<0.01

5.4 Growth estimation using automated transcriptions

Human scoring of the recorded readings used to compute WCPM can be costly and time-consuming. Automated measurement of reading fluency offers an attractive alternative since the reading can be scored almost instantaneously. However, unlike human transcribers, automated systems can be particularly sensitive to noisy and unclear recordings, such as those often collected under classroom condition [5, 13, 15]. Would we still be able to observe statistically significant growth if WCPM measurements came from an automated system which introduced additional noise due to technical issues?

We used the system described in [13] to obtain automatic fluency measures. Our system uses automated speech recognizer (**ASR**) to convert the signal to text with timestamps indicating the beginning and end of each word; it then uses string matching to identify the part of the transcription where the student is attempting to read the passage; finally it uses the text of the passage, the transcription, and the timestamps to compute WCPM. The system was trained on data from external corpora or previous data collections: none of the data used in this study was used for system training or fine-tuning.

We used this automated system to analyze the 6,479 recordings in our corpus. In agreement with previous studies, we also found that ASR often produced a very short hypothesis[10] due to background noise, mumbling or otherwise unclear speech, and other factors. As it would not be meaningful to use such hypotheses for fluency measurement, we first ran an automated system for removal of leading and trailing off-task material [12], and then only used the recordings where the ASR hypothesis had at least 70% of the target word count for the passage.[11] 1,297 responses (20%) did not meet the minimal on-task length condition and were removed. For the passages where we had both an ASR estimate and a transcription-based estimate of WCPM (N=1,281) the correlation between these two estimates was r=0.75. The final data set contained 5,182 responses from the same 35 students as in 5.3 but this time for 614

[10] The transcription returned by ASR is commonly called a hypothesis.
[11] Note that we only considered the total number of words, we did not further evaluate whether the reading contained errors.

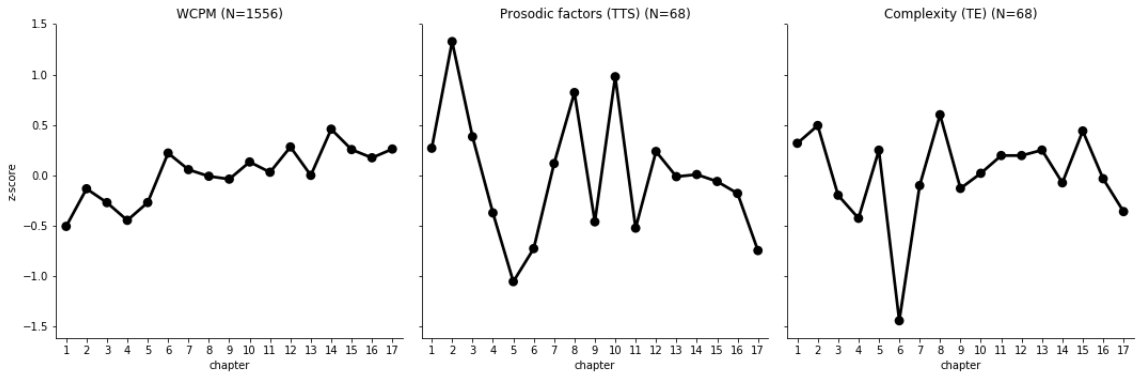

Figure 1: Standardized average values of WCPM, TTS and TE by chapter (see 5.2 for further explanation of TTS ans TE.)

unique passages. The student by passage matrix was only 24% full. We fitted the model in Eq. 2 to this data. Table 2 shows the results.

Table 2: Model estimates for the model in Eq. 2 based on ASR for all recordings from the subset of 35 students and for all 42 students. The values for TTS and TE were standardized to μ =0 and σ =1 and then entered into the model.

	Dependent variable:	
	WCPM	
	35 students	42 students
Grade5	-0.40 (9.70)	1.47 (8.71)
TTS	5.60*** (0.50)	5.78*** (0.40)
TE	-2.43*** (0.49)	-2.88*** (0.41)
Chapter	1.21*** (0.26)	1.12*** (0.25)
Constant	62.08*** (7.91)	60.13*** (6.77)
Observations	5,182	6,814
Note:	*p<0.05; **p<0.01; ***p<0.001	

We found that results based on automated measurements were comparable to those based on transcriptions. Passage-level predictors, TTS and TE, still had a significant effect on student WCPM with the coefficient sign consistent with expectation. Most importantly, chapter remained a significant predictor of student WCPM with growth rate estimated at 1.2 correctly read words per chapter (vs. 1.3 based on transcription).

When selecting the data for the original analysis, we excluded 7 students who read on a different schedule (see 4.2). Since we found that the passage effect on WCPM is estimated to be relatively small (6%), and since the models were robust to a dramatic increase in data sparsity (65% vs 24%), we tested whether adding the data for these students to the dataset would change the results. After applying the same selection criteria as described in 4.3 and earlier in this section (apart from the restriction on number of words in the passage), we had 1,632 recordings from these students, bringing the total number of readings in our corpus to 6,814 (149 hours of audio recordings). There are 1,207 distinct passages now, with the passage by reader matrix only 13% full. We then fitted the model in Eq. 2 to these 6,814 recordings. The results are presented in the last column of Table 2. They remained consistent with what we observed before: The chapter was still a significant predictor of WCPM, confirming that the previous finding still holds when data from all students is taken into account.

6 DISCUSSION

The analysis presented in section 5.3 estimates the rate of growth in oral reading fluency at 1.3 words per minute per chapter. How does this square with growth estimates in the literature? We observe that the October through January administration of the reading program roughly[12] coincides with the time between Fall and Winter administrations of the periodic oral reading fluency checks reflected in the norms [8]. For an estimate of the expected gain in oral reading fluency, we use the 50th percentile norms for 4th and 5th grades. Children in 4th grade are expected to gain 26 WCPM between Fall and Winter; 5th graders are expected to gain 12 WCPM. Our estimate of the cumulative gain throughout the 17 chapters of the book at the rate of 1.3 correctly read words per chapter is 22 WCPM. Our estimate is thus broadly consistent with the expectation.

A series of analyses using automated transcription challenged the robustness of the positive growth result obtained using transcribed data by using much noisier data, albeit in a much larger quantity. Our findings indicate that the significant effect of chapter on WCPM remains intact, as do the passage-related effects observed in the original model.

7 CONCLUSION

In this study, extensive but noisy longitudinal oral reading data was collected from upper elementary school children during multi-week book reading using Relay Reader, a reading app where the child is taking turns reading a book out loud with a model audiobook narrator. In this paper, we present a sequence of modeling and data selection steps that allowed us to zero in on a subset of data small enough to be feasibly transcribed yet large enough to support estimation of growth in oral reading fluency. Using the transcribed

[12]Most of the readers were actually done before January, some as early as November.

data, we found that the time variable yielded a statistically significant contribution with a positive coefficient in a mixed linear model that also controlled for random variation associated with readers and passages, as well as for some known text effects in oral reading. Moreover, the estimated rate of growth in oral reading is consistent with published norms of oral reading fluency based on repeated administration of a dedicated test of oral reading fluency. We also show that growth can be detected using both human transcriptions and automated transcription, suggesting that speech analysis technology is up to the task and can be used to score a large number of recordings without incurring the cost of human transcription in order to detect learning.

In all, our results suggest that unobtrusive tracking of improvement in oral reading fluency in the background of book reading with the Relay Reader app is feasible due to a large number of observations collected for each reader, in spite of lack of control over texts, substantial environmental and behavioral noise, and imperfections of automated speech recognition technology. We are thus a step closer to the vision of having children read for the story, not for the test, yet being able to assess their progress. Next steps include replication of the finding with other books, measurement of learning transfer across books, and investigation of relationships between oral reading fluency as measured through the app with other estimates of fluency and with other reading-related skills, such as comprehension.

REFERENCES

[1] 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 (2012), 435–452.

[2] D. Bates, M. Mächler, B. Bolker, and S. Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67, 1 (2015), 1–48.

[3] Beata Beigman Klebanov, Anastassia Loukina, Nitin Madnani, John Sabatini, and Jennifer Lentini. 2019. Would you? Could you? On a tablet? Analytics of Children's eBook Reading. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge. ACM, Tempe, AZ, USA, 106–110.

[4] J. Bernstein, J.Cheng, J. Balog, and R. Downey. 2018. Artificial intelligence for scoring oral reading fluency. In Applications of artificial intelligence to assessment, H. Jiao and R. Lissitz (Eds.). Charlotte, NC: Information Age Publisher.

[5] Lei Chen. 2009. Audio Quality Issue for Automatic Speech Assessment. In Proceedings of SLaTE Workshop. Wroxall, UK, 97–100.

[6] R.H. Good, R.A. Kaminski, and S. Dill. 2002. DIBELS oral reading fluency and retell fluency. In Dynamic indicators of basic early literacy skills (6th ed.), R.H. Good and R.A. Kaminski (Eds.). Eugene, OR: Institute for the Development of Educational Achievement.

[7] J. Hasbrouck and G. Tindal. 2006. Oral Reading Fluency Norms: A Valuable Assessment Tool for Reading Teachers. The Reading Teacher 59 (2006), 636–644.

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

[9] Julia Hirschberg. 2002. Communication and prosody: Functional aspects of prosody. Speech Communication 36, 1-2 (2002), 31–43.

[10] Roland H. Good III, 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 (2001), 257–288.

[11] Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. Journal of Statistical Software 82, 13 (2017), 1–26. https://doi.org/10.18637/jss.v082.i13

[12] A. Loukina, B. Beigman Klebanov, P. Lange, B. Gyawali, and Y. Qian. 2017. Developing speech processing technologies for shared book reading with a computer. In 6th International Workshop on Child Computer Interaction. 46–51.

[13] A. Loukina, B. Beigman Klebanov, P. Lange, Y. Qian, B. Gyawali, N. Madnani, A. Misra, K. Zechner, Z. Wang, and J. Sabatini. 2019. Automated Estimation of Oral Reading Fluency During Summer Camp e-Book Reading with MyTurnToRead. In Proceedings of the Annual Conference of the International Speech Communication Association. Gratz, Austria, 21–25.

[14] A. Loukina, V. Liceralde, and B. Beigman Klebanov. 2018. Towards Understanding Text Factors in Oral Reading. In Proceedings of Annual 2018 Conference of the North American Chapter of the Association for Computational Linguistics. New Orleans, Louisiana, 2143–2154.

[15] Yi Luan, Masayuki Suzuki, Yutaka Yamauchi, Nobuaki Minematsu, Shuhei Kato, and Keikichi Hirose. 2012. Performance improvement of automatic pronunciation assessment in a noisy classroom. In Proceedings of 2012 IEEE Workshop on Spoken Language Technology (SLT). Miami, FL, 428–431.

[16] 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. Florence, Italy, 141–146.

[17] Jack Mostow. 2012. Why and how our automated reading tutor listens. In Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training. Stockholm, Sweden, 43–52.

[18] 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. Denver, Colorado, 96–100.

[19] Joseph F. T. Nese, Gina Biancarosa, Daniel Anderson, Cheng-Fei Lai, Julie Alonzo, and Gerald Tindal. 2012. Within-year oral reading fluency with CBM: a comparison of models. Reading and Writing 25, 4 (2012), 887–915.

[20] Eunice Kennedy Shriver National Institute of Child Health and DHH Human Development, NIH. 2000. Report of the National Reading Panel: Teaching Children to Read: Reports of the Subgroups (00-4754). https://www.nichd.nih.gov/sites/default/files/publications/pubs/nrp/Documents/report.pdf

[21] University of Oregon. 2019. 8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Administration and Scoring Guide. https://dibels.uoregon.edu/docs/materials/d8/dibels_8_admin_and_scoring_guide_07_2019.pdf

[22] Lynn Olsen. 2005. Benchmark Assessments Offer Regular Checkups on Student Achievement. Education Week 25, 13 (2005), 13–14.

[23] Marianne Perie, Scott Marion, and Brian Gong. 2009. Moving Toward a Comprehensive Assessment System: A Framework for Considering Interim Assessments. Educational Measurement: Issues and Practice 28, 3 (2009), 5–13.

[24] Brian Poncy, Christopher Skinner, and Philip Axtell. 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 (2005), 326–338.

[25] R Core Team. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/

[26] Stephen Raudenbush and Anthony Bryk. 2002. Hierarchical Linear Models: Applications and Data Analysis Methods (second ed.). Sage, Newbury Park, CA.

[27] Tom Snijders and Roel Bosker. 2012. Multilevel Analysis (2 ed.). Sage, London.

[28] Miya Miura Wayman, Teri Wallace, Hilda Ives Wiley, Renata Ticha, and Christine A. Espin. 2007. Literature synthesis on curriculum-based measurement in reading. Journal of Special Education 41, 2 (2007), 85–120.

[29] Laurence White. 2014. Communicative function and prosodic form in speech timing. Speech Communication 63-64 (2014), 38–54.

A MODEL EQUATION

Let $i = 1, \ldots, N$ index students and let $p = 1, \ldots, P$ index passages. Let $W_{i,p}$ denote the measured WCPM for student i when reading passage p. Let $G_i = 0$ if student i is in grade 4, and 1 if student i is in grade 5. Define $\mathbf{X}_p$ to be the vector of passage-level predictors (TTS and TE) for passage p. Define C_p to be the chapter $1, \ldots, 17$ in which passage p is situated. We model $W_{i,p}$ with the following linear mixed-effects model:

$$W_{i,p} = \mu + \gamma G_i + \boldsymbol{\beta}'\mathbf{X}_p + \theta_p + \alpha_{0i} + (\delta + \alpha_{1i})C_p + \epsilon_{i,p}, \quad (3)$$

where μ is a constant, γ is the coefficient corresponding to grade level, $\boldsymbol{\beta}$ is a vector of coefficients associated with the passage-level predictors, and δ is the main effect of chapter on ORF. The model also includes the following random effects, which are assumed to be mutually independent: passage random effects $\theta_1, \ldots, \theta_P$ assumed to be normally distributed with mean zero and variance τ^2, person-level random intercepts and slopes $\{(\alpha_{0i}, \alpha_{1i})'\}_{i=1}^{N}$ assumed to be normally distributed with mean vector $\mathbf{0}$ and (2×2) variance-covariance $\boldsymbol{\Psi}$, and residual errors $\epsilon_{i,p}$ assumed to be normally distributed with mean zero and variance σ^2.