INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Automated estimation of oral reading fluency during summer camp e-book reading with MyTurnToRead

Anastassia Loukina, Beata Beigman Klebanov,
Patrick Lange, Yao Qian, Binod Gyawali,
Nitin Madnani, Abhinav Misra, Klaus Zechner, Zuowei Wang, John Sabatini

Educational Testing Service, USA
aloukina@ets.org, bbeigmanklebanov@ets.org

Abstract

Use of speech technologies in the classroom is often limited by the inferior acoustic conditions as well as other factors that might affect the quality of the recordings. We describe MyTurnToRead, an e-book-based app designed to support an interleaved listening and reading experience, where the child takes turns reading aloud with a virtual partner. The child's reading turns are recorded, and processed by an automated speech analysis system in order to provide feedback or track improvement in reading skill. We describe the architecture of the speech processing back-end and evaluate system performance on the data collected in several summer camps where children used the app on consumer-grade devices as part of the camp programming. We show that while the quality of the audio recordings varies greatly, our estimates of student oral reading fluency are very good: for example, the correlation between ASR-based and transcription-based estimates of reading fluency at the speaker level is r=0.93. These are also highly correlated with an external measure of reading comprehension.

Index Terms: child-computer interaction, educational applications

1. Introduction

Many existing commercial and research applications already use ASR and other speech processing technologies to assess oral reading fluency and assist with its development [1, 2, 3, 4]. These studies consistently report high agreement between automated estimates and human ratings. Yet research also shows that the reliability of speech technology-based measurements is closely related to the audio quality of the recording [5]. Background noise, mumbling or unclear speech can all interfere with the processing and result in inaccurate measurements. Thus Moby.Read [6] instructs users to "Be in a QUIET place" on the sign-in page of the app while Flora [3] recommends the use of a noise-cancelling microphone.

To ensure good quality recordings under classroom conditions, previous studies of automated reading tutors often arranged students in small groups in a quiet area [7, 3, 8], provided recording equipment [7, 8, 3], and had experimenters or trained facilitators monitor the session [1, 3, 8]. Yet even under these conditions, [8] reported the recordings from some children were silent or completely unintelligible.

In this paper we present a new application, MyTurnToRead, designed to encourage sustained reading while providing continuous and unobtrusive measurement of oral reading fluency. The focus on reading for pleasure and over an extended period of time makes consistently controlling the recording conditions very difficult, while the low-stakes, not-for-test nature of such reading means that there is little extrinsic motivation for students to read clearly and loudly.

Our previous evaluations using data collected during a series of pilot studies showed a high correlation between automated and transcription-based estimates of oral reading fluency, thus confirming previous findings that the technology is sufficiently mature to support such applications. Yet we also found that even the recordings collected during monitored data collection contained substantial acoustic and behavioral noise. In this study, we describe our first experience of having the app used in two summer camps as part of the programming over an extended period of time and without close monitoring by the project researchers.

We evaluate (1) whether the recordings collected under such conditions contain a sufficient amount of intelligible read-aloud speech for measuring student reading fluency; (2) whether the audio quality allows obtaining accurate automatic estimates of reading fluency; and (3) whether such estimates are correlated with external measures of children's reading skill.

2. MyTurnToRead

MyTurnToRead is a cross-platform web and mobile app which uses e-book technology to allow the user to take turns reading a book aloud with a virtual partner, realized through an audiobook narration [9]. The goal of MyTurnToRead is to encourage sustained enjoyment of a full-length novel over a period of several weeks. The premise is that the interest in the story and the quality of the narration would increase enjoyment, while the interleaving of the effortful reading with the more relaxing experience of listening to a skilled narrator should help reduce the perceived difficulty of the task. As the children are reading with MyTurnToRead, the app logs information about their interaction with the app, while the audio of the user's turn is recorded and sent to the server for further processing. See [9] for further information about app functionality beyond audio recording.

2.1. Speech processing pipeline in MyTurnToRead

The speech processing pipeline used for the evaluation of oral reading fluency in MyTurnToRead includes three major components: automated speech recognition, off-task speech identification and computation of oral fluency measures.

Automated speech recognition The automated speech recognition (ASR) used for this project is based on the Kaldi toolkit [10]. We use *nnet2* acoustic models trained on the LibriSpeech corpus [11]. The features used to train the DNN are concatenated MFCC features and i-vector features. The MFCC features have 39 dimensions, while the i-vector features have 100 dimensions extracted from each response and appended to

http://dx.doi.org/10.21437/Interspeech.2019-2889

the frame-level MFCC features. The input features stacked over a 15 frame window (7 frames to either side of the center frame for which the prediction is made) are used as the input layer of the DNN. The DNN has 5 hidden layers, and each layer contains 1,024 nodes. The sigmoid activation function is used for all hidden layers. All the parameters of the DNN are first initialized by layer-wise, back-propagation pretraining, then trained by optimizing the cross-entropy function through back-propagation, and finally refined by sequence-discriminative training, state-level minimum Bayes risk (sMBR).

We trained a custom language model for each chapter of the book currently incorporated into the app. Training and testing of language models is further described in [12].

Identification of off-task speech Many of the recordings collected by the app contain at least some off-task speech, especially before and after the reading. Furthermore, as shown in [12], background speech during silences can often be picked up by ASR and recognized as part of the child's response. In order to make sure our analyses are only based on the child reading the text, we developed a module for identifying such speech when it occurs before or after the on-task reading. The performance of this system is evaluated in [12].

Measures of oral reading fluency The ASR hypothesis and the associated timestamps are then used to compute various measures related to oral reading fluency. In this paper, we focus on one such measure: words correct per minute (WCPM). WCPM is a standard measure of oral reading fluency which combines aspects of speed and accuracy and has been shown to be a good predictor of reading skills [13, 14]. It is computed as the total number of correctly read words divided by the total time it took the child to read the passage based on the timestamps we estimated for the beginning and end of on-task speech.

2.2. Challenges of controlling the audio quality

The accuracy of speech processing components is closely related to the quality of the audio recordings. In the educational space, several approaches have been used to ensure that the recordings of spoken responses allow for valid measurement of the skill of interest. These include controlling the recording conditions and equipment as well as monitoring the sessions. Self-administered applications are often set up so that the user's experience is directly linked to the quality of their recordings. For example, the student who submitted low-quality recordings might be asked to repeat their answer or would not receive any score.

MyTurnToRead is designed to be used over the period of several weeks or months. In fact, one of the reasons we developed a mobile app was to allow the children to read where they are comfortable. This makes consistently controlling the environment very difficult, leading to a high likelihood of background noise. Furthermore, the devices available in school or summer camps for such extended reading might not necessarily be optimal for high-quality sound recording.

The low-stakes nature of reading with MyTurnToRead also means that there is little extrinsic motivation for students to read clearly and loudly. We also know that students often do not read the target passage completely [9], thus producing readings that are both partial *and* noisy.

One possible solution to this problem would be to automatically monitor the quality of the recordings and encourage the user to address any problems. However, the first and foremost goal of MyTurnToRead is to encourage sustained reading. Interrupting student reading whenever the audio quality fails to meet sufficient standards or asking them to re-read their turns would be counterproductive. Our aim is to track fluency unobtrusively, without interfering with the user's flow of reading and the enjoyment of the book. However, the danger of such an approach is that a large share of the recordings captured by the app might not contain a sufficient amount of intelligible, read-aloud speech, making it impossible to use speech processing to provide estimates of user reading fluency.

3. Pilot and field study

Before deploying the app in the extended summer camp trial, we tested various components in a series of pilot studies. These evaluations are described in [12, 15] and summarized in this section.

3.1. Pilot corpus

The first round of evaluations was done on the 'pilot corpus' collected in Spring 2017 with a goal of evaluating various system components. As no app was yet available, the task was not interactive reading but just plain reading of a few passages [16, 12]. It included 63 recordings (2.2 hours) of 22 children reading 3 texts from the same book used in the app. The recordings were captured using a laptop and a professional-grade microphone. The recordings were conducted in an office with 2-3 children reading simultaneously and monitored by proctors. While the age of the children matched that of the target population, the children were selected via a convenience sample and were very fluent and accurate readers in comparison to their peers. The automatic estimates of WCPM showed almost perfect correlation with the transcription-based estimates with r= .98, once again confirming that speech processing can be used to estimate WCPM with a high degree of accuracy [1, 4, 6].

3.2. Field study

The next data collection took place in Summer-Fall 2017 using a research prototype of the actual interactive reading system. The system displayed the book and alternated between playing the recorded narration for some passages and prompting the user to read aloud other passages. The system was deployed on laptops and the reading aloud was captured via the same headsets as for the 'pilot corpus'. The data collection was conducted at several after-school programs and lasted 5 days at each site. 36 children participated in the study. The first session was used to administer various tests. During the remaining four days, the children interacted with the program for 25 minutes a day. All sessions were monitored by several members of the project team. The 'field corpus' collected during this study consisted of 538 recordings (15.9 hours).

The analyses conducted by [15][1] showed that the data collected during this field study could be used to obtain accurate automatic estimates of oral reading accuracy (r=.88 at speaker level). At the same time, about 40% of recordings contained a substantial amount of background or behavioral noise, raising the question of what would happen if the app were to be used for an extended period of time and with no monitoring from the project staff. The rest of this paper addresses this question.

[1] [15] presents evaluations of an ASR system running on-device. The server-based system discussed in this paper uses similar models and achieves similar performance.

4. Extended summer camp trial

The new data considered in this paper was collected in the summer of 2018 in two summer camps in the Greater New York area. Both camps were located in low-income areas.

The summer camp trial reflected real conditions under which the app might be used in an after-school or summer camp setting. In both camps, regular sessions with MyTurnToRead were scheduled as part of the camp programming and continued for around 6 weeks. The recording conditions, including background noise, were outside our control: although we provided a set of recommendations, the camp instructors selected the time and place of the reading sessions. They were also the ones who monitored the activity. The children used a beta version of the app (see [9] for further detail). While previously recordings were done using professional-grade equipment provided by the project, in this study the children used low-cost Android tablets and consumer-grade headsets with built-in microphones to interact with the app.

In total, 32 children participated in the activity (18F, 14M).[2] Their ages ranged from 8 to 11;10 with mean age = 9;7. All children were fluent English speakers although some spoke a second language at home. The children read a popular children's/young adult novel (almost 80,000 words). Since 6 weeks was not sufficient to finish the novel, all children were given a paperback version after the end of the trial.

We logged almost 170 hours of total reading with the app and 2,399 student turns. The average length of the passage read by students at each turn was 139.7 words (sd = 31.1). All turns were transcribed by a professional transcription agency who also annotated the quality of the recording on a three point scale: 'good', 'noisy' or 'bad'.

4.1. Reading efficiency scores

Before the start of the reading sessions, the children took a computer-administered test to obtain an external measure of their reading comprehension skills. We used the efficiency of basic reading comprehension (EBR) sub-test from the Reading Inventory and Student Evaluation (RISE) test [17]. As described in [17], this subtest consists of three passages. In each sentence within a passage, one of the words is replaced with three choices, only one of which makes sense in the sentence. According to [17], this task design, known as the 'maze technique', has been shown to be a good indicator of basic reading efficiency and comprehension. Due to scheduling issues, some of the children were not able to complete the test. The reading efficiency scores were available for 21 out of 32 children in our sample.

4.2. The percentage of 'Bona-fide' readings

We first analyzed the logs captured by the app and the transcriptions of all turns to evaluate whether the app was successful at eliciting read-aloud speech and capturing sufficiently high-quality recordings for human transcribers.

The app logged 2,399 student turns. Of these, the audio was missing for 50 turns (2%) and empty for 31 turns (1.2%). The total number of non-empty audio recordings was 2,318. We also excluded a small number of turns (7 or less than 1%) where the passage to be read was shorter than 20 words, thereby leaving us with 2,311 turns (47.2 hours of audio).

[2]We excluded 4 children who were initially enrolled in the program but logged fewer than 20 turns.

As previously shown by [9], logged turn duration can be a useful source of information about student behavior when interacting with the app. They used the published norms and the length of the prompt to estimate whether the turn duration is 'reasonable' under the assumption of *bona-fide* reading. Out of 2,311 turns, 787 were determined 'too fast' and indeed the transcriptions for these turns contained no evidence of on-task speech (category 3 in Table 1).

Out of the remaining 1,524 turns with 'reasonable' durations, the transcription for 273 contained fewer than 5 words (category 4 in Table 1), even though the passage contained at least 24 words to be read (median passage length for these 273 turns was 142 words). Such relatively long recordings with almost no speech transcribed might be due to high background noise, especially if the speaker was reading softly (92% of such recordings were marked as 'bad' or 'very noisy'); silent reading; or perhaps a student being distracted by other activities while the recording is running. In other words, as we expected, the recordings for a large number of turns were not usable for analyzing student reading due to technical issues or unexpected user behavior.

For the remaining 1,251 turns, the number of words transcribed as on-task reading was on average 135.6 words (sd=46.7). Median reading accuracy (total number of correctly read words/total number of words in the passage) was 95%[3]. Thus about 50% of all 2,399 logged turns contained what we can assume to be 'bona fide' reading of the book recorded with quality generally sufficient for a human to transcribe. Only one out of 32 participants did not have any turns that fell into this category. Furthermore, the percentage of 'bona fide' turns in the data remained relatively constant (around 50%) throughout the 6 weeks of the data collection.

Table 1: *Summary of different types of turns recorded through the MyTurnToRead app during the summer camp trial. The categories are applied from top to bottom such that the count for each subsequent category excludes turns that have already been included in previous categories.*

Category	**N**	**%**
1. Missing/empty audio	81	3.4%
2. Passage $<$ 20 words	7	$<$1%
3. Turn duration 'too short'	787	32.8%
4. $<$ 5 words in transcriptions	273	11.3%
5. 'Bona fide' reading	1,251	52.1%
Total logged turns	**2,399**	

4.3. Automatic and transcription-based estimates of oral reading fluency

We next considered whether the audio quality of the recordings is sufficient for automatic measurement of oral reading fluency. In Table 1, we listed 4 types of responses that we did not consider 'bona fide' reading. Three of these (1-3) can be easily identified automatically and therefore we excluded such responses when selecting the data for evaluation. However, we kept in the sample the 273 turns where the duration of

[3]The distribution of reading accuracy was very skewed. While the majority of responses had accuracy higher than 90%, there was a long tail where the transcription-based accuracy was low either due to student skill or behavior or because the transcribers could not transcribe parts of the recording due to background noise.

the recording was 'reasonable' but the transcribers did not transcribe any reading aloud (category 4 in Table 1), since an additional module would be necessary to identify such recordings automatically. We also excluded 5 turns where our estimates of WCPM were greater than 500. This occurred when the identified on-task reading consisted of a single short word.

This left us with 1,519 recordings from 32 participants. The analysis of quality annotations for this subset showed that 49.2% of these recordings were marked as 'good' quality with the remaining deemed 'bad' or 'noisy'. The ASR word error rate (WER) for on-task speech varied from 0% to substantially above 100% with a median value at 40%. Unsurprisingly, WER was lowest for responses marked as 'good' quality with the median value of 25%. Yet even on 'good' responses, ASR performance was substantially lower than the performance of this system on the 'pilot' corpus (9.3%) or 'field' study (33.4% overall, 14.6% for good responses)

Despite the high percentage of noisy responses and low ASR performance, our estimates of WCPM remained relatively accurate. The correlation at the individual turn level (N=1,519) was Pearson's r = .80[4]. Aggregating the measurements across several turns improved the performance further. The correlation for average WCPM across all turns recorded in a single day for each child (N=302) was r=.87, while the correlation of average WCPM across all turns for each student (N=32) was r=.93.

4.4. Correlation with reading efficiency scores

Our analyses so far showed that automatic estimates of WCPM showed a high correlation with transcription-based estimates. Yet they also revealed a substantial amount of noise in the recordings, which raises the question of whether human transcriptions can be considered a reliable measure of students' reading fluency under such circumstances. To address this, we compared automatic and transcription-based estimates of student WCPM with their performance on the efficiency of basic reading comprehension test (EBR).

Valid EBR scores were only available for 21 out of 32 students who took part in this study. There was no significant difference in average WCPM between the 21 children with EBR scores and the 11 children who did not complete the test (t-test $t(29)$=-0.32, p=0.78). We also confirmed that our previous finding about the correlation between transcription-based and automatic estimates of WCPM still held for this subset of 21 speakers: the correlation for this group was r=.92, similar to what we observed for the whole group (r=.93).

The correlation between EBR scores and the *transcription-based* estimates of speaker WCPM from our app was r=.73 (p=0.0001), which means the estimates of oral reading fluency we obtained from the app recordings were closely aligned with the reading comprehension scores for the same students. This result is consistent with previous studies who reported a similar correlation ($r \sim 0.7$) between WCPM and reading comprehension measures for elementary-grade children [18, 19].

The correlation between the *automatic* estimates and the EBR scores was r=0.69 (p=0.0005). In other words, the automated speaker-level estimates of oral reading fluency were not only consistent with the transcription-based estimates, but they were also correlated with the external measure of their reading skill.

The correlation matrix is summarized in Table 2.

[4]All correlations reported in this section are significant at α=0.0001 unless stated otherwise.

Table 2: *Correlation matrix between speaker-level automatic and transcription-based estimates of* WCPM *and reading efficiency scores* EBR. *N=21*

	WCPM(automatic)	EBR
WCPM (transcription)	.92	.73
WCPM (automatic)	-	.69

5. Discussion and conclusion

In this paper, we presented the setup and evaluations of automated analysis of oral reading by elementary school children collected through MyTurnToRead, an app designed to support sustained reading over an extended period of time while providing unobtrusive measurement of oral reading fluency. The data was collected through an extended trial in two summer camps where reading sessions with the app were offered as part of camp programming under the conditions the app would be used in future real-life applications.

We found that children continued using the app throughout the duration of the summer camp. Furthermore, despite the informal context and the low-stakes nature of reading with the app, about 50% of the children's reading turns were clear enough to transcribe and process automatically. Finally, the oral reading fluency measurements extracted from these readings were meaningfully related to other measures of children's skill (r=0.7).

Our focus on unobtrusive measurement and sustained reading meant that students could continue reading the book even in conditions that were not optimal for sound recording. We also did not penalize students who did not fully cooperate with the task (e.g. did not read aloud the whole text of their turn). As a result, a large number of the recordings did not contain any reading while many more recordings contained background noise, mumbling or unclear speech.

However, the extended reading setup with MyTurnToRead affords lots of opportunities for a child to demonstrate his or her reading skill. Only one child who participated in our data collection did not record any oral reading. All other children read aloud clearly and loudly enough at least some of the time and as a result we were still able to obtain good skill estimates. Even if MyTurnToRead users act on only *some* of the opportunities to demonstrate their true oral reading skill while interacting with the app over the span of a few weeks, as long as they do act on them at least occasionally, we will likely to be able to collect an adequate number of 'bona fide' reading samples that evidence their true skill to enable estimation. Additional data collections in other educational contexts are needed to see whether this state of affairs generalizes, and whether the collected data can also support longitudinal tracking of improvement in reading fluency.

6. Acknowledgements

We would like to thank Jennifer Lentini who took care of all logistic aspects of the trial and was instrumental to its success; Yuriy Donev, Georgi Angelov and the rest of Astea Solutions team for the outstanding app design and development work; Kelsey Dreier for assistance with RISE test; site administrators and instructors in the summer camps for their enthusiasm implementing the reading program. We also thank Keelan Evanini, Michael Flor, Xinhao Wang, Hillary Molloy, and three anonymous reviewers for their comments and suggestions.

7. References

[1] J. Balogh, J. Bernstein, J. Cheng, A. Van Moere, B. Townshend, and M. Suzuki, "Validation of automated scoring of oral reading," *Educational and Psychological Measurement*, vol. 72, no. 3, pp. 435–452, 2012.

[2] J. Mostow, "Why and how our automated reading tutor listens," *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training*, pp. 43–52, 2012.

[3] D. Bolaños, R. A. Cole, W. H. Ward, G. A. Tindal, J. Hasbrouck, and P. J. Schwanenflugel, "Human and automated assessment of oral reading fluency." *Journal of Educational Psychology*, vol. 105, no. 4, pp. 1142–1151, nov 2013. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0031479

[4] J. Bernstein, J. Cheng, J. Balogh, and E. Rosenfeld, "Studies of a Self-Administered Oral Reading Assessment," in *Proceedings of SLaTE 2017*, O. Engwall, J. Lopes, and I. Leite, Eds. Stockholm: KTH Royal Institute of Technology, 2017, pp. 180–184.

[5] S.-Y. Yoon, A. Cahill, A. Loukina, K. Zechner, B. Riordan, and N. Madnani, "Atypical Inputs in Educational Applications," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 60–67. [Online]. Available: http://aclweb.org/anthology/N18-3008

[6] J. Cheng, "Real-time scoring of an oral reading assessment on mobile devices," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-Septe, no. September, pp. 1621–1625, 2018.

[7] J. Mostow, G. Aist, P. Burkhead, A. Corbett, A. Cuneo, S. Eitelman, C. Huang, B. Junker, M. B. Sklar, and B. Tobin, "Evaluation of an Automated Reading Tutor That Listens: Comparison to Human Tutoring and Classroom Instruction," *Journal of Educational Computing Research*, vol. 29, no. 1, pp. 61–117, 2005.

[8] J. Bernstein, J. Cheng, J. Balogh, and R. Downey, "Artifical intelligence for scoring oral reading fluency," in *Applications of artificial intelligence to assessment.*, H. Jiao and R. Lissitz, Eds. Charlotte, NC: Information Age Publisher, to appear, pp. 1–33.

[9] B. Beigman Klebanov, A. Loukina, N. Madnani, J. Sabatini, and J. Lentini, "Would you? Could you? On a tablet? Analytics of Children's eBook reading." in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19*. New York, New York, USA: ACM Press, 2019, pp. 106–110. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3303772.3303833

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[11] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2015, pp. 5206–5210. [Online]. Available: http://ieeexplore.ieee.org/document/7178964/

[12] A. Loukina, B. Beigman Klebanov, P. Lange, B. Gyawali, and Y. Qian, "Developing speech processing technologies for shared book reading with a computer," in *WOCCI 2017: 6th International Workshop on Child Computer Interaction*, no. November. ISCA: ISCA, nov 2017, pp. 46–51.

[13] S. P. Ardoin, T. J. Christ, L. S. Morena, D. C. Cormier, and D. A. Klingbeil, "A systematic review and summarization of the recommendations and research surrounding Curriculum-Based Measurement of oral reading fluency (CBM-R) decision rules," *Journal of School Psychology*, vol. 51, no. 1, pp. 1–18, 2013.

[14] M. M. Wayman, T. Wallace, H. I. Wiley, R. Tich, and C. A. Espin, "Literature synthesis on curriculum-based measurement in reading," *The Journal of Special Education*, vol. 41, no. 2, pp. 85–120, 2007.

[15] A. Loukina, N. Madnani, B. Beigman Klebanov, A. Misra, G. Angelov, and O. Todic, "Evaluating on-device ASR on Field Recordings from an Interactive Reading Companion," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, dec 2018, pp. 964–970. [Online]. Available: https://ieeexplore.ieee.org/document/8639603/

[16] B. Beigman Klebanov, A. Loukina, J. Sabatini, and T. O'Reilly, "Continuous fluency tracking and the challenges of varying text complexity," in *Procceings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. Copenhagen, Denmark.: Association for Computational Linguistics, 2017, pp. 22–32.

[17] J. Sabatini, K. Bruce, and J. Steinberg, "Sara Reading Components Tests, Rise Form: Test Design and Technical Adequacy," *ETS Research Report Series*, vol. 2013, no. 1, pp. i–25, 2014.

[18] R. H. Good, D. C. Simmons, and E. J. Kame'enui, "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*, vol. 5, no. 3, pp. 257–288, 2001.

[19] A. D. Roehrig, Y. Petscher, S. M. Nettles, R. F. Hudson, and J. K. Torgesen, "Accuracy of the DIBELS Oral Reading Fluency Measure for Predicting Third Grade Reading Comprehension Outcomes," *Journal of School Psychology*, vol. 46, no. 3, pp. 343–366, jun 2008.