Psychological Test and Assessment Modeling, Volume 65, 2023 (1), 36-54

Predicting Oral Reading Fluency Scores by Between-Word Silence Times Using Natural Language Processing and Random Forest Algorithm

Yusuf Kara[1], Akihito Kamata[1], Emrah Emre Ozkeskin[2], Xin Qiao[1], and Joseph F. T. Nese[3]

[1] Southern Methodist University
[2] Heilbronn University of Applied Sciences
[3] University of Oregon

Abstract

The measurement of oral reading fluency (ORF) is an important part of screening assessments for identifying students at risk of poor reading outcomes. ORF is a complex construct that involves speed, accuracy, and coherent reading abilities including prosody. This study aimed at using between-word-level silence times collected through a computer-based reading assessment system to predict words read correctly per minute (WCPM) scores of young readers as the measure of their ORF levels. Natural language processing (NLP) was utilized to analyze reading passages to inform the locations of syntactically dependent words, namely, meaningful word chunks. Then, silence times before and after the NLP-informed word chunks were used to predict WCPM scores via a random forest algorithm. The results revealed that students' average relative silence times before and after specific word chunks were good predictors of WCPM scores. Also, the model was able to explain more than half of the variation in WCPM scores by using the derived silence times and students' grade-levels as predictors.

Keywords: oral reading fluency, machine learning, natural language processing, random forests

Author Note

The research reported here was partially supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D200038 to the Southern Methodist University and through Grants R305A140203 and R305A200018 to the University of Oregon. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Correspondence concerning this article should be addressed to Yusuf Kara, Center on Research and Evaluation, Southern Methodist University, 6116 N. Central Expressway, Suite 400, Dallas, TX 75206. E-mail: ykara@smu.edu

Oral reading fluency (ORF), generally defined as reading quickly, accurately, and with prosody, is an essential part of reading proficiency (Fuchs et al., 2001). According to the definition provided by the National Reading Panel (2000), fluency is one of the five factors deemed central to reading instruction, and fluency instructions have promoted reading growth. Measures of ORF are robust indicators of comprehension and overall reading achievement.

ORF has mainly been assessed in classrooms as part of curriculum-based measurement of reading (CBM-R) applications. In a typical CBM-R assessment of ORF, students read aloud a given grade-level text for a one-minute session, while an assessor (such as a classroom teacher) tracks students' reading performance and records the words read incorrectly or omitted. At the end of the reading session, the assessor computes the total number of words read correctly per minute (WCPM), which is a widely used metric for ORF assessments. WCPM scores have been shown to have good concurrent and predictive validity in terms of measuring ORF (Francis et al., 2008; Fuchs et al., 2001).

ORF assessment through CBM-R has several benefits including the ease of administration, which in turn helps classroom teachers provide interventions to students at risk of poor reading. Nevertheless, there are also key limitations associated with CBM-R applications, including the inaccurate measurement of reading time and other sources of construct-irrelevant variance such as the errors made by human assessors. More importantly, the traditional approaches to ORF assessment cannot collect other types of reading data such as the time taken to read each word, the duration of silence between adjacent words, and recorded voice data among others. With the advancements and affordability of computer technologies, ORF assessments can be delivered via computer following the current trends in digitally based assessments, which have been adopted in large-scale assessments such as National Assessment of Educational Progress (NAEP) and Programme for International Student Assessment (PISA).

In this paper, we demonstrate the use of rich time data in search of advancing the assessment of ORF beyond using total reading times and words read correctly at the passage level. Our aim is in parallel with approaches in psychometric research that use process data to better understand the underlying mechanisms of solving problems or simply, answering an item correctly (e.g., He & von Davier, 2016; Qiao & Jiao, 2018; Shu et al., 2017; Ulitzsch , et al., 2022; Zhu et al., 2016). In the context of ORF assessment, we specifically aim to explore empirical evidence to support the claim that fluent readers make meaningful pauses in particular locations of reading passages. To achieve this goal, we evaluate whether between-word-level silence times have sufficient predictive power in estimation of ORF scores measured in the scale of WCPM. We utilize natural language processing (NLP) algorithm for the extraction of text features, which are used for the identification of meaningful pausing locations in passages at the word level. NLP-informed silence times are then aggregated to student-level and used as predictors of WCPM scores in a random forest regression (RF) model. Also, we aim to examine an example tree from the RF regression model to evaluate if longer pauses for some meaningful word chunks are associated with higher WCPM scores.

The structure of the paper is as follows. We first provide a brief background for the computer-based ORF assessment system followed by the rationale of using silence times as reading process data to be analyzed for exploration of fluent/non-fluent reading characteristics. In the methods section, we explain our sample and measures, the derivation of NLP-informed aggregated silence times, and procedure of RF regression analyses. Then we present our results followed by a discussion and a limitations and future research section.

A Computer-Based ORF Assessment System

Nese and Kamata[1] developed a computer-based ORF assessment system called computerized oral reading evaluation (CORE) (Nese & Kamata, 2021). With CORE, reading passages are delivered in a computer-based environment, which facilitates the collection of reading accuracy and time data at the word level. CORE employs an automated speech recognition (ASR) engine to decode the reading of students. Then, ASR scores accuracy of reading each word as correct or incorrect.

In addition to more efficient scoring of word-level reading, CORE measures the reading time of each word in the scale of centiseconds (i.e., 1/100 second). Such word-level precise measurement of reading times cannot be achieved by the traditional CBM-R, where human raters keep track of time. Moreover, CORE records the reading of students in a voice data format. Such data can also be retroactively used for more detailed analyses to inform the assessment of ORF and related reading competencies (e.g., Sammit et al., 2022).

Use of Between-Word Silence Times

Word-level time data collected by CORE is a rich source of information in addition to the indicators of passage-level speed (total reading time) and accuracy (total number of the words read correctly) that are used in traditional ORF assessments. The duration of silence or pause, while reading two adjacent words in a passage, can easily be derived from the word-level reading times recorded by the computerized system.

The use of silence times to investigate fluent/non-fluent reading behaviors is not a novel approach in reading research. Pausing has been found to be linked to decoding behaviors in dysfluent readers (Miller & Schwanenflugel, 2008). Lower skilled

[1] Nese, J. F. T. & Kamata, A. (2014-2018). Measuring Oral Reading Fluency: Computerized Oral Reading Evaluation (Project No. R305A140203) [Grant]. Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/funding/grantsearch/details.asp?ID=1492

readers have been found to produce more pausing as they read more complex text (Benjamin & Schwanenflugel, 2010) and as they progress through texts (Cowie et al., 2002). Research has repeatedly shown that students with low fluency scores made proportionally more random and grammatically irrelevant pauses than those with higher fluency levels, such as pauses due to difficulties in decoding and parsing, or physiologically running out of air while reading (Benjamin & Schwanenflugel, 2010). They are also reported to make more and longer intra-sentential pauses (Miller & Schwanenflugel, 2006, 2008; Schwanenflugel et al., 2004), which can be interpreted as excessive and extended pauses made by non-fluent readers.

On the other hand, reading with minimal pauses would not make a reader fluent as it is known that pausing behaviors for fluent readers are associated with meaningful word chunks, such as after a phrase-final commas and other grammatically justified word chunks (Chafe, 1988). In other words, fluent readers are expected to pause at particular locations of a passage, specifically, before and after meaningful word chunks such as a complete sentence of a passage. The relationship between meaningful pauses and fluent reading can also be explained through prosodic reading behavior, which is reported to be strongly associated with ORF (Valencia et al., 2010).

In summary, pausing behaviors of readers while reading aloud a text can be used to evaluate their fluent reading characteristics. It is also important to note that the location and duration of such pauses are all important for such an evaluation. Moreover, pausing behaviors should be explored in connection with the characteristics of the text read, namely, passages. Below, we present how we approached using the pausing behaviors of readers defined by silence times in explaining their relationship with fluent reading, which is measured by WCPM scores.

Methods

Sample

Reading accuracy and time data were originally collected as part of a larger study (Nese & Kamata, 2021). The original sample included 2,094 unique students in Grades 2, 3, and 4 in two school-districts in Pacific Northwest in 2017-18 and 2018-19 academic years. The original data were collected with a total of 150 passages, where a student read 10 passages on average. Around 79.5% students were assigned in one session of reading assessment, around 20% students were assessed in two sessions, and the remaining small proportion 0.5% were assessed in three sessions. Multiple sessions of data from the same student were treated as distinct observations, which led to a total of 2,543 observations. We discarded inaccurate/outlier observations from the data, such as the ones with lower than 50% of words read correctly counts per passage and with missing end-of-reading time stamps. It is noted that the former exclusion was performed to remove the observations for which the WCPM value was not computed correctly, mostly due to a problematic session of reading. For the current study, we selected data for 90 passages and a total of 2,278 observations of students from the three grade-levels.

Measures

CORE passages are original works of fiction (Nese & Kamata, 2021), with ±5 words of the target word length (medium = approximately 50, long = approximately 85). Start and end times were recorded by CORE for each word. In the word-level raw data, a student had as many rows of observations as the number of words that he/she read in all passages. Each row of observation per student contained the actual word of a passage (punctuations were omitted), a 0/1 score for the accuracy of reading that word, and time stamps for the beginning and end of the reading for that word. Reading time of a word was then computed through the mentioned start- and end-time stamps.

We processed the raw reading time data to compute silence times between adjacent words in each passage. Specifically, we used the end time of a word and the beginning time of the adjacent word to compute the silence time between two adjacent words in the scale of centiseconds. Between-word silence times were then converted into relative silence times (RSTs). The RSTs were computed by dividing the between-word silence times by the total silence time for each student for the passage that he/she read. We excluded the silence time prior to starting to read a passage (i.e., silence time before the first word of a passage) from the total silence time. This exclusion was for the purpose of preventing confounding effects of initial wait time, or the time taken for a student to start reading, as we observed that some students had much longer initial silence time.

The reason we use the RST was that the same actual silence times would not necessarily mean the same magnitude without this transformation, because each reader reads in different speed. Thus, silence times relative to a reader’s speed, or more specifically, to the total of silence time while reading a passage would be a more meaningful measure. Further, the premise was that the silence time in a particular location of the passage would be longer than silence times in other parts of the passage when a student makes a meaningful pause. As a result of the transformation to the RST, the silence time measure was comparable both within and between students. Without this transformation, this would have led to a dominant time-fluency relation, such that only non-fluent readers would be identified if actual silence times were used as predictors, as non-fluent readers had longer actual between-word silence times than fluent readers.

The outcome variable for ORF was the WCPM score computed from all passages read in a session by a student. The derivation of the WCPM scores was as follows: (1) the total number of the words read correctly for each student in each session was computed by summing up the word-level 0/1 scores (0 = incorrect, 1 = correct) for all passages, (2) the total reading (in seconds) time was computed by summing up all word-level reading times and between-word silence times, and (3) a WCPM score was computed by the ratio between (1) and (2) in the scale of per minute.

Text Analysis and Feature Extraction

Our intention for using the RSTs as predictors of ORF was tied to our hypothesis about meaningful pauses made by fluent readers at specific locations in a passage. Thus, we did not consider all computed RST measures between any adjacent words of all passages that a student read. We employed an NLP algorithm to analyze the passages read by students to inform our selection of specific word locations in passages. In other words, we aimed at identifying sequences of words for which RSTs (before and/or after these sequences of words) can be used as predictors. This effort can be associated with the use of n-grams in psychometric research (e.g., He & von Davier, 2016; Ulitzsch et al., 2022), where subsequences of actions are identified as part of feature extraction. In our case, sequences would be the groups of adjacent words within a sentence that bear a meaningful word group and/or refer to a grammatical structure such as relative clauses. Note that this step of the analyses was performed solely on the text data, where we used the actual reading passages as texts. Detailed information about how we linked the NLP-based analyses with the RST data at the word level is provided in the following paragraphs.

The NLP analysis of the passages was performed by *spaCy 2.0* package (Honnibal & Montani, 2017) in Python 3.7 (Python Software Foundation, 2018). We followed an iterative approach during the NLP analyses by starting with basic text mining, including identification of stop words and tokenization, where the latter produced tokens that helped us associate the meaning of the text to word-level reading data. Then, we performed parts of speech tagging to the tokenized words and identified the syntactic dependencies associated with tokens. In other words, this last step identified meaningful dependencies between words (i.e., word sequences) in a sentence within each passage. Note that we use the term “meaningful word chunks” to refer to instances of “text features”, which are the types of syntactic dependencies between multiple adjacent words in a sentence. Thus, meaningful word chunks would be instances of generic text features.

Table 1
Extracted Text Features with Natural Language Processing

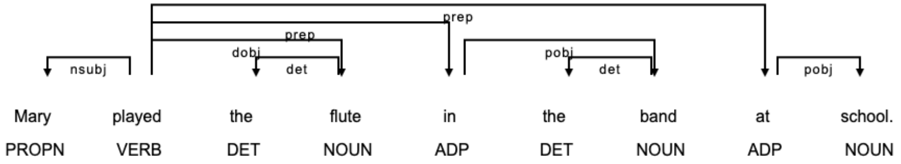
Feature	Description	Obs. %	Feature	Description	Obs. %
acl	clausal modifier of noun (adjectival clause)	16.7	expl	expletive	13.3
acomp	adjectival complement	66.7	intj	interjection	1.1
advcl	adverbial clause modifier	74.4	mark	marker	55.6
advmod	adverbial modifier	98.9	neg	negation modifier	30
agent	agent	1.1	nmod	modifier of nominal	1.1
amod	adjectival modifier	92.2	npadvmod	noun phrase as adverbial modifier	45.6
appos	appositional modifier	13.3	nsubj	nominal subject	100
attr	attribute	53.3	nsubjpass	nominal subject (passive)	20
aux	auxiliary	95.6	nummod	numeric modifier	31.1
auxpass	auxiliary (passive)	21.1	oprd	object predicate	8.9
case	case marking	28.9	pcomp	complement of preposition	15.6
cc	coordinating conjunction	87.8	pobj	object of preposition	98.9
ccomp	clausal complement	68.9	poss	possession modifier	92.2
compound	compound	72.2	preconj	pre-correlative conjunction	2.2
conj	conjunct	86.7	predet	predeterminer	16.7
csubj	clausal subject	1.1	prep	prepositional modifier	100
dative	dative	15.6	prt	particle	38.9
dep	unclassified dependent	2.2	quantmod	modifier of quantifier	1.1
det	determiner	100	relcl	relative clause modifier	50
dobj	direct object	100	xcomp	open clausal complement	77.8

Note. Obs. % is the percent of passages in which the specific feature was observed at least once.

Figure 1 presents an example graphical output of dependencies in one sentence that was extracted from one of the passages. The NLP algorithm identified the meaningful word chunks within this sentence and visualized the beginning and end of each chunk. For example, “Mary played” is referred to as a meaningful word chunk and it was identified as a text feature called “nominal subject” (abbreviated as *nsubj*) by the algorithm. Another word chunk, “in the band”, is identified as an instance of text feature called “object of preposition” (abbreviated as *pobj*). As can be observed in this example sentence, identified word chunks may not be mutually exclusive. In other words, there were several word chunks starting with the same word but ending with different words. Also, some instances of text features such as *pobj* was observed multiple times within a sentence. A total of 40 unique text features were identified in 90 passages (see Table 1).

Figure 1

An Example of Extracted Text Features from a Sentence



Note. Descriptions of the text-feature labels are provided in Table 1.

We flagged the beginning (i.e., first word) and end (i.e., last word) of all meaningful word chunks in reading passages based on the extracted text features via NLP. These flags were then associated with student-level reading data to inform our selection of word-level RST measures. For example, we flagged beginning and end of all nominal subjects read by a student in all passages he/she administered. As a following step, we averaged the before and after RSTs for each type of text feature within a passage per student. For example, a student would have two passage-level average RST measures (before and after) computed from all instances of nominal subjects that appeared in a passage he/she read. This passage-specific averaging allowed us to generalize the RST values related to a particular type of text feature from many instances observed within a passage. Then, passage-level RSTs before and after each text feature were further averaged across all passages for each observation of students' reading. The second averaging at the student-observation level allowed us to generalize the feature-specific average RST values over many passages that a student read. As a result, we derived a total of 80 student-level average RST variables (2 variables x 40 text features) based on all passages they read. Note that this version of the RST dataset was no longer linked to specific passages. In other words, the variables of this dataset were the average RSTs before and after the text features extracted from all passages. Finally, the student-level average RST data were merged with the outcome data, namely, student-level WCPM scores.

Treatment of Missing Data

Average RST data associated with the extracted text features had missing observations due to three major reasons: (1) not all text features appeared in all passages, (2) not all passages read by all students, and (3) unmeasured/wrong word-level reading and/or silence times existed in the raw data. The latter missingness can be considered as the noise in word-level data caused by the computerized data collection system. For example, there were some instances in the word-level data, where ASR was not able to assign a correct time stamp either for the beginning or end of reading a single word. We reported occurrence percentages of the features in Table 1 considering all 90 passages. Specifically, these values represent the percent of passages that included at least one instance of the relevant feature. Thus, a value of 100%, for example, indicates that the relevant feature observed in all 90 passages at least once[2]. Also, features that observed more frequently across passages are expected to have lower percentages of missing observations in the final analysis dataset, which comprised of average RST values per student observation. Although not reported here, the percentage of missing observations in the final dataset varied between 0% - 94.5% among the 80 average RST variables. We excluded average RST variables with more than 30% of missingness. After this exclusion, 49 derived average RST variables remained in the final dataset.

Missing observations for the remaining 49 average RST variables were imputed by adopting a random forest (RF) algorithm approach that is implemented in the *MissForest* (Stekhoven, 2022) package for R (R Core Team, 2023). The imputation with the RF in the *MissForest* R package follows an iterative approach, where the RF algorithm runs on the observed part to predict the missing observations per variable. These RF iterations stop until a pre-defined criterion is met, or the user-defined maximum number of iterations is reached. Also, during the imputation of the test data, imputed training data can be used as an additional source of information. More details about the implementation of this iterative imputation approach can be found in Stekhoven and Buhlmann (2012).

Predictive Analysis

We used the RF regression algorithm (Breiman, 2001) to predict the ORF outcomes measured by the WCPM scores with the average RST measures for extracted text features. The RF is a tree-based method, where multiple regression/classification trees are built based on randomly sampled subsets of data. Then, individual regression/classification trees are combined to produce the final predictions or classifications. The RF method is referred to as an ensemble machine learning (ML) algorithm, because of the combination of multiple prediction models. Also, it is known to produce more accurate prediction and classification results compared to single-tree approaches (e.g., Breiman, 2001; Hayes et al., 2015).

Prior to the RF analyses, we created training and test datasets by randomly splitting the main imputed dataset by the 75/25 ratio. As a result, sample sizes were N=1,708 and N=570 for the training and test datasets, respectively. In addition to the 49 derived average RST measures, we also used the grade level of students as a predictor. Note that the grade level was intended to function as a control variable, since the expected values of the WCPM scores would be higher in higher grades. Also, the regression

[2] We did not report the frequency of occurrence per feature since the final data were comprised of the average relative silence times within and between passages. Thus, if a text feature occurred at least once in a passage, related final data observation was not missing.

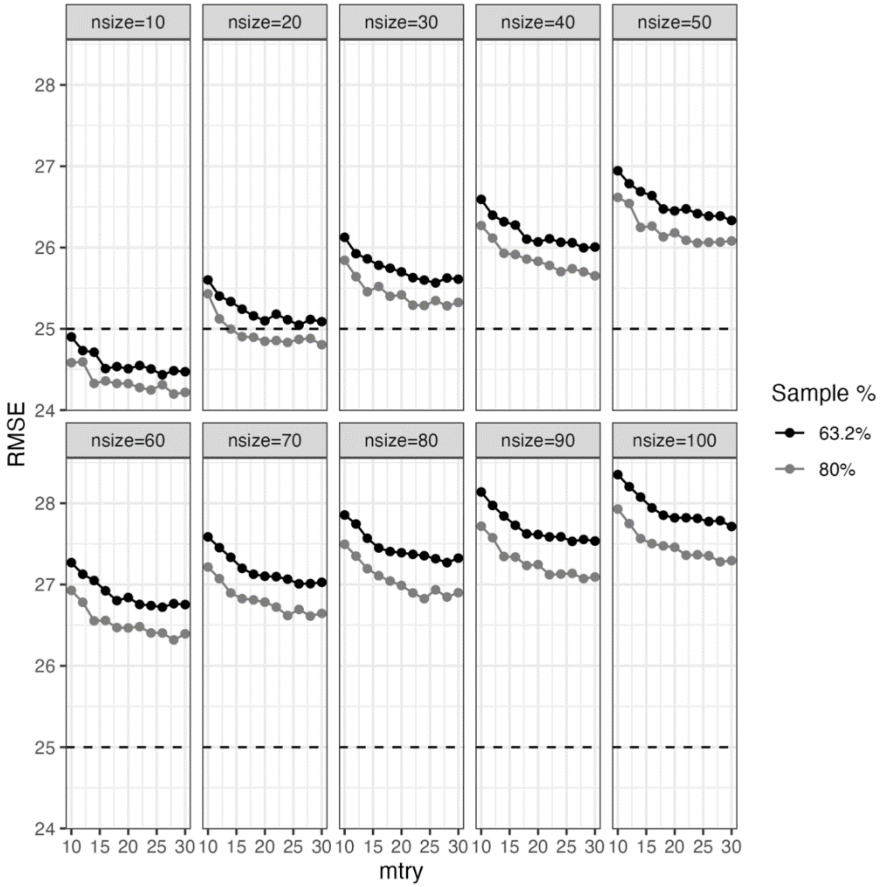
trees may be different between grade levels. Thus, we aimed to capture any differences between grade levels, in terms of the relations between silence times and the WCPM scores. As reported in the result section later, the inclusion of the grade level as a predictor did not inflate the R-square values drastically, compared to the model without the grade level.

We used the *randomForestSRC* package (Ishwaran & Kogalur, 2023) for R (R Core Team, 2023), which implements the RF algorithm (Breiman, 2001) for the RF regression analyses. The default RF algorithm parameters as implemented in the package were as follows: the number of trees (*ntree*)=500, the number of variables for splitting at each node (*mtry*)=1/3 of the number of the predictors, minimum terminal-node size (*nsize*)=5, the number of random splits (*nsplit*)=10, bootstrap sample size (*bsize*)=63.2% of the original sample size, bootstrap sampling type=sampling without replacement, and splitting rule=weighted mean-squared error. As described in Sinharay (2016), a total number of *ntree* decision trees are built based on the *bsize* percentage of bootstrap sampling from the training data, either with replacement or not. For each tree, the algorithm selects a random subset of predictors as the size of *mtry* at each split with a total of desired splits as *nsplit* and the minimum sample size of a node as *nsize*. Then, prediction outputs from each regression tree are averaged over *ntree* trees as the final output of the RF model.

Following the commonly adopted approach in fitting RF models (e.g., James et al, 2013; Qiao & Jiao, 2018), we explored the prediction performance of the model on training data by manipulating values of the three key parameters: *mtry* (10 to 30 with increments of 2), *nsize* (10 to 100 with increments of 10), and ratio for *bsize* (63.2% and 80%). We did not manipulate the number of *ntrees* (used the default value as 500), since it has been known to be less effective on performance of the model compared to other parameters (e.g., James et al., 2013; Sinharay, 2016). Combinations of various values from these parameter factors created a grid of 220 unique conditions. Then, we fitted the RF model to the training data with each of these parameter conditions to identify the optimal RF parameters for model tuning. The RMSE values from the analyses with 220 parameter conditions are graphically summarized in Figure 2. The horizontal dashed line is placed on RMSE=25, with the aim of targeting a maximum prediction error of 25 WCPM scores in the RF regression model, while seeking optimum parameter values for tuning.

Figure 2

RMSE Change Based on Model Tuning Parameters



Note. RMSE: Root mean squared error. mtry: number of the variables for splitting at each node. nsize: minimum terminal-node size

Results summarized in Figure 2 showed that the effect of *mtry* on RMSE values was not as prominent as the effect of *nsize* (the minimum terminal node size). Also, the effect of *bsize* (the % of the bootstrap sample size) showed a small yet recognizable effect. It is generally known that, as the *nsize* gets lower, the precision of the RF model gets better with the expense of deeper trees, which makes the RF model results hard to interpret (Sinharay, 2016). Thus, we intentionally decided not to automatically select the parameters that resulted in the lowest RMSE value. Following our aim for RMSE < 25, we selected the optimal parameters as *mtry*=25, *nsize*=20, and *bsize*=63.3%. The RMSE and R-squared values for the tuned model with these parameters were 25.053 and 0.473, respectively. We also fitted a model with no grade-level as a predictor using the same RF model parameters. The RMSE and R-squared values from this model fitted to the training data were 26.62 and 0.405, respectively, which still demonstrated reasonable fit.

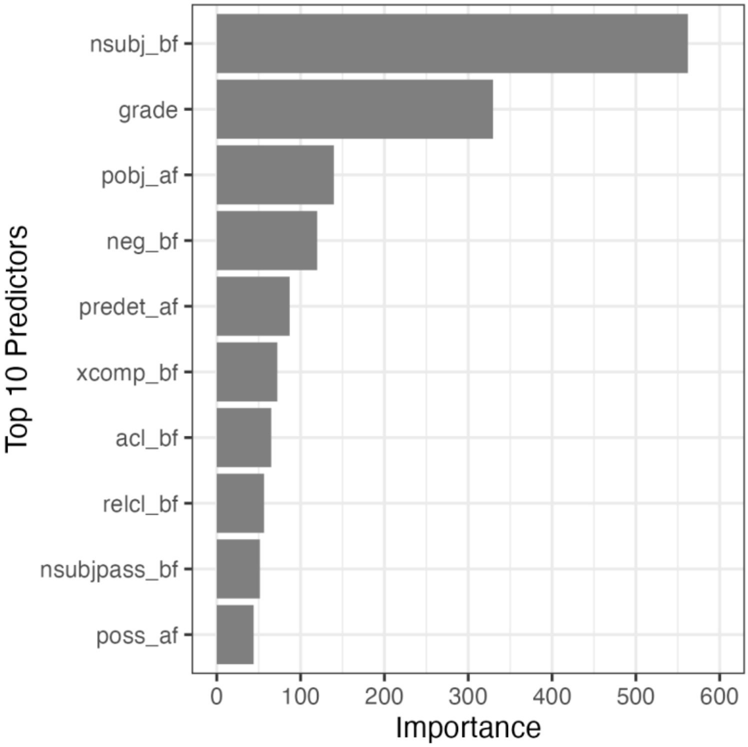
Results

Ten largest variable importance values based on the final tuned model (fitted to the training data) are graphically summarized in Figure 3. Note that these are the permutation importance values, also referred to as the Breiman-Cutler importance (Breiman, 2001; Ishwaran & Kogalur, 2007). The largest importance value was observed for the average RST before the *nsubj* feature, which is “nominal subject”. Also, the importance of the average RST before *nsubj* was almost as twice much as the next largest value, which was the grade level, although our intention was to include the grade level as a control variable due to known differences between grade-specific average WCPM values. The order of the remaining 8 most important average RST measures was as follows: after object of prepositions (*pobj_af*), before negation modifiers (*neg_bf*), after predeterminers (*predet_af*), before open clausal complements (*xcomp_bf*), before adjectival clauses (*acl_bf*), before relative clause modifiers (*relcl_bf*), before passive nominal subjects (*nsubjpass_bf*), and after possession modifiers (*poss_af*).

An example of a tree from the grown RF model is shown in Figure 4. We selected the first tree out of 500 total trees grown in the entire model. In this regression tree, the first node was created based on the grade level, where second graders (indicated with a value of 1) were separated from third and fourth graders. The first quartile of WCPM scores for second graders was 48.4 in the original data (training + test data). By tracking the average WCPM scores lower than the first quartile in the example tree, we can see a node of n=24 observations with the average WCPM of 41.99. This group of second graders were identified based on having higher average RSTs before *ccomp* (> .838), *case* (> .425), and *advcl* (> 1.741). The lowest performing group of second graders (n=13) had the average WCPM of 37.37 and was identified through the same path of the former group plus through having a higher average RST before *dative* (> 1.44) and lower average RST after *prt* (<= .517). The third quartile value of WCPM scores in the original sample was 95.23 for second graders. The highest performing second-grade student group in the tree (n=39) had the average WCPM of 107.2 and identified as having lower average RSTs before *ccomp* (<= .838), after *prt* (<= 1.486), and before *attr* (<= .484); but a higher average RST value before *nsubj* (> 2.887).

Figure 3

Variable Importance for Top Ten Predictors

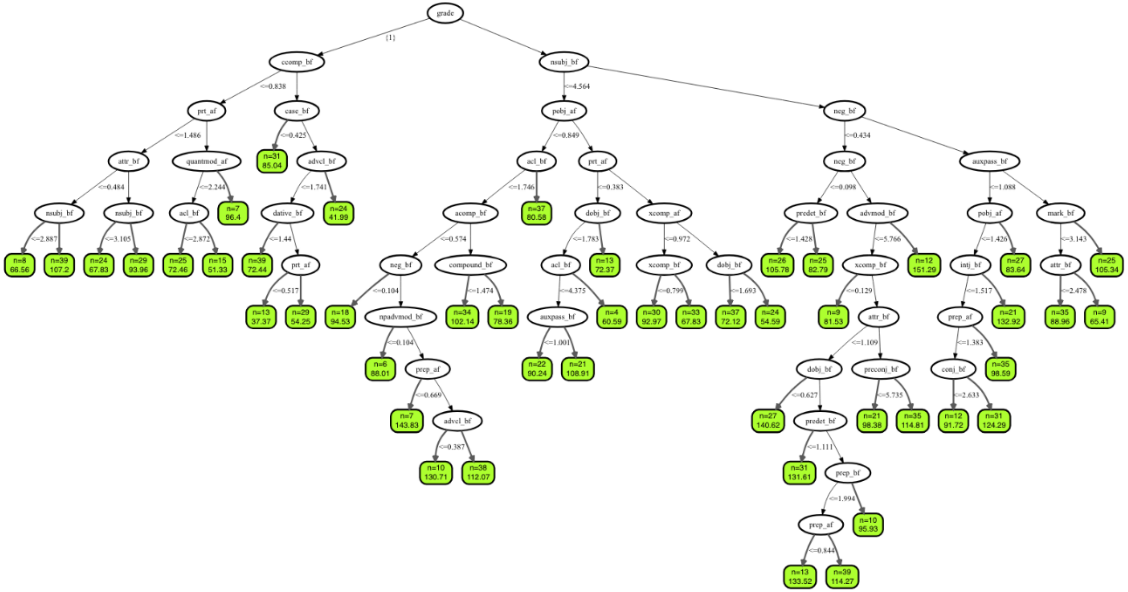


Note. "bf" and "af" stands for "before" and "after" that specific text feature, respectively. Explanations of the text features can be found in Table 1.

The first quartile of WCPM scores for the combined group of third and fourth graders in the original sample was 73.37. There were several nodes in the tree with average WCPM scores varied between 55-72, which were lower than the first quartile. The lowest performing group (n=24, average WCPM=54.59) had an average RST before *nsubj* lower than 4.564 and higher average RST values after *pobj* (>.849), after *prt* (>.383), and after *xcomp* (> .972), and before *dobj* (> 1.693). The third quartile of WCPM scores for the combined group of third and fourth graders was 120.39. There were also several nodes in the tree with average WCPM values larger than this value, specifically with a range of 124-151. The best performing group in the tree (n=12) had the average WCPM score of 151.29 and identified as having an average RST larger than 4.564 before *nsubj*, and lower than .434 average RST before *neg*, followed by a larger average RST before *neg* (> .098) and before *advmod* (> 5.766).

Figure 4

An Example Tree from the Random Forest Model



Note. A value of {1} for grade refers to second graders. *Note.* “bf” and “af” stands for “before” and “after” that specific text feature, respectively. Explanations of the text features are provided in Table 1.

Cross-validation analysis conducted on the test data demonstrated promising results related to prediction power of average RST measures. Specifically, the correlation between the predicted and observed WCPM scores was .764, which corresponds to an R-squared value of .584. This implies that the RF model with the grade-level variable and average RSTs associated with the extracted text features explained approximately 58.4% of the variance in the WCPM scores as the measures of ORF. The R-squared value for the model with test data without the grade level as a predictor was 0.527, which corresponds to 52.7% of WCPM score variance explained with average RST measures. This highlights that the average RST measures were the dominant predictors that explain the majority of the variance in the WCPM scores, compared to grade-level, which was incorporated into the model as a covariate/control variable.

Discussion

In this study, we demonstrated the use of rich time data collected through a digitally administered ORF assessment. Aiming to find empirical evidence for a claim that fluent readers make meaningful pauses, we attempted to predict the WCPM scores as the measures of ORF, by focusing on readers' silence (i.e., pausing) tendencies while reading aloud a text. We performed NLP analyses to extract text features from reading passages and derived RST measures associated with these extracted text features. Then, we used the RF regression, an ensemble ML algorithm, to examine the prediction power of the RSTs occurred before and after the extracted text features in estimating the WCPM scores as the measures of ORF.

Our analyses revealed promising results. First, more than 50% of the variance in WCPM scores was explained by the trained model. More importantly, it was revealed that average RSTs before and/or after specific text features, such as nominal subjects (e.g., "she was" or "Mary played"), were important predictors of WCPM scores as the measures of ORF. Moreover, the inspection of an example tree from the grown forest provided more specific insights, especially for the identification of at-risk students in reading performance. For example, the lowest performing second graders were identified mainly with larger average RSTs before/after specific text features. Larger RSTs mean relatively longer pauses while reading. Such consistently large RST before/after specific text features can be considered as indicators of poor reading performance. The tree also revealed the silence time tendencies of high-performing students in terms of reading fluency. The highest preforming second graders were explored to have mainly lower average RSTs before/after specific text features. Compared to former interpretation for low performers in second grade, the top performers are identified with relatively shorter silence times before/after specific text features. Nevertheless, for third and fourth graders, the top performer group was identified through some larger RSTs before/after specific text features. Thus, it was not always true that fluent readers have shorter average RSTs before/after specific text features. This also means that fluent readers at higher grades have longer pauses relative to their entire silence time, which supports our hypothesis that fluent readers are expected to perform meaningful pauses at specific locations of a passage.

In conclusion, our results showed that student-level average RST measures before or after specific word chunks identified via NLP algorithm are associated with overall ORF scores. Moreover, examination of the regression trees can be used for a more detailed exploration of how RST can be used to identify more/less fluent readers by looking at specific type of text feature and the associated RSTs. These two conclusions from the current study support the same major aim: searching for the usability of rich silence time data for a better understanding and measurement of ORF. Even though we approached this aim by showing the prediction power of such derived silence times, another strategy could be incorporating them into a measurement model, which can improve the current WCPM-based scoring approach for ORF assessment.

Limitations and Future Research

There are several limitations for the current study. First, we interpreted a single tree from the entire RF model. However, we acknowledge this was a limitation of the study since the pattern of the silence tendencies may be different from one tree to another. Further, the generalizability of these single-tree observations would be limited considering the low sample sizes we observed for low/high performing groups of students. Thus, the RF model should be used at its entirety for the identification of fluent vs non-fluent readers via their predicted WCPM scores in a future study.

Second, we did not include any other predictors such as word-by-word reading times, actual/relative reading times of meaningful word chunks, or word-level reading accuracy data (i.e., 0/1 score for reading each word correctly). We believe that incorporating such predictors has the potential to improve the prediction power of the model, as well as better explanations of the association between silence time and the performance on ORF assessment. In addition, different ML algorithms can be applied to the same data in search of a better predictive model.

Third, we aggregated the RST measures within each of the 40 unique text feature types. By aggregating the RSTs, it helped us understand the association between the silence time associated with the extracted text features and the ORF performance. However, we acknowledge the loss of information by this approach. In a future study, attempts to retain more fine-grained information will be worthwhile by using the RST measures differently. For example, observations of RST can be aggregated only at the passage level per text feature to retain potential between-passage differences in the associations between text features and the ORF performance.

Fourth, text features extracted by the NLP algorithm were limited to a total of 40 types. There may be other meaningful word-chunks or single-word locations in passages, where readers are expected to pause while reading. Examples for such word-chunks or locations that were not identified by the NLP algorithm include quotes (i.e., phrases indicated with quotation marks), complete sentences, and article/stop words, among others. Future studies can consider extracting more text features. Further, studies with more passages would contribute to an increased representation of different types of text features, which may lead to a better prediction of fluency scores.

Lastly, we used WCPM scores as the measures of ORF as they were the only available measures for the current study. Other measures such as Dynamic Indicators of Basic Early Literacy Skills (DIBELS) scores can also be used as the measures of ORF ability of young readers. Use of different scores for ORF is considered to provide another layer of cross-validation to the usability of silence times in exploring the pausing tendencies of fluent and non-fluent readers.

There is no doubt that digitally based assessments will gain further interest in the field of psychometrics for better measurement of various abilities/constructs. With the availability and affordability of computer technology, assessments are adopted not only in large-scale assessments, but also in classroom-based assessments. In addition to key advantages, such as decreasing the burden caused by the administration of paper-pencil tests, digitally based assessments provide new opportunities for the collection and analysis of richer and new types of data. ORF assessments can also get benefit from the availability of rich process data including reading/silence times, text specifications of passages, and even recorded voice data among others. Thus, there is potential for further research to be conducted on the effectiveness of using process reading data collected by digitally based ORF assessment systems. Moreover, the RST before and after meaningful word chunks identified in this study may be useful in improving an automated scoring algorithm for the prosody (Sammit et al., 2022), as the meaningful pauses are characteristics of the prosody (Hirschberg, 2002).

References

Benjamin, R. G., & Schwanenflugel, P. J. (2010). Text complexity and oral reading prosody in young readers. *Reading Research Quarterly, 45,* 388-404.

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5–32. https://doi.org/10.1023/A:1010933404324

Chafe, W. (1988). Punctuation and the prosody of written language. *Written Communication, 5*(4), 395–426. https://doi.org/10.1177/0741088388005004001

Cowie, R., Douglas-Cowie, E., & Wichmann, A. (2002). Prosodic Characteristics of Skilled Reading: Fluency and Expressiveness in 8—10-year-old Readers. *Language and Speech, 45,* 47-82.

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

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

Hayes, T., Usami, S., Jacobucci, R., & McArdle, J. J. (2015). Using classification and regression trees (CART) and random forests to analyze attrition: Results from two simulations. *Psychology and Aging, 30*(4), 911–929. https://doi.org/10.1037/pag0000046

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.) *Handbook of research on technology tools for real-world skill development* (pp. 749-776). Hershey, PA: Information Science Reference.

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

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Ishwaran, H., & Kogalur, U. B. (2007). Random survival forests for R. *R News, 7*(2), 25–31.

Ishwaran, H., & Kogalur, U. B. (2023). Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC), R package version 3.2.0. https://cran.r-project.org/web/packages/randomForestSRC/index.html

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*, Vol 112. New York, NY: Springer.

Miller, J., & Schwanenflugel, P. J. (2006). Prosody of syntactically complex sentences in the oral reading of young children. *Journal of Educational Psychology, 98*(4), 839–853. https://doi.org/10.1037/0022-0663.98.4.839

Miller, J., & Schwanenflugel, P. J. (2008). A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. *Reading Research Quarterly, 43,* 336-354.

National Reading Panel. (2000). Report of the national reading panel: Teaching children to read (Publication No. 00-4769). https://www.nichd.nih.gov/sites/default/files/publications/pubs/nrp/Documents/report.pdf

Nese, J. F. T., & Kamata, A. (2021). Evidence for automated scoring and shorter passages of CBM-R in early elementary school. *School Psychology, 36(1),* 47–59. https://doi.org/10.1037/spq0000415

Python Software Foundation. (2018). Python language reference (Version 3.7). https://www.python.org

Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: A didactic. *Frontiers in Psychology, 9*, 2231. https://doi.org/10.3389/fpsyg.2018.02231

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Sammit, G., Wu, Z., Wang, Y., Kamata, A., Nese, J. F. T., & Larson, E. (2022). Automated Prosody Classification for Oral Reading Fluency with Quadratic Kappa Loss and Attentive X-Vectors. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2022, pp. 3613-3617. https://doi.org/10.1109/ICASSP43922.2022.9747391

Schwanenflugel, P. J., Hamilton, A. M., Kuhn, M. R., Wisenbaker, J. M., & Stahl, S. A. (2004). Becoming a fluent reader: reading skill and prosodic features in the oral reading of young readers. *Journal of Educational Psychology*, *96*, 119.

Shu, Z., Bergner, Y., Zhu, M., Hao, J., & von Davier, A. A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks. *Psychological Test Assessment Modeling*, *59*(1), 109.

Sinharay, S. (2016). An NCME instructional module on data mining methods for classification and regression. *Educational Measurement: Issues and Practice, 35*, 38-54. https://doi.org/10.1111/emip.1211

Stekhoven D. J. (2022). missForest: Nonparametric Missing Value Imputation using Random Forest. R package version 1.5.

Stekhoven, D. J., & Buhlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Ulitzsch, E., He, Q., & Pohl, S. (2022). Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. *Journal of Educational and Behavioral Statistics, 47*(1), 3–35. https://doi.org/10.3102/10769986211010467

Valencia, S. W., Smith, A. T., Reece, A. M., Li, M., Wixson, K. K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly*, *45,* 270-291.

Zhu, M., Shu, Z., & von Davier, A. A. (2016). Using networks to visualize and analyze process data for educational assessment. *Journal of Educational Measurement,* 53, 190-211. https://doi.org/10.1111/jedm.12107