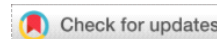




pISSN 2005-8063
eISSN 2586-5854
2026. 9. 30.
Vol.18 No.3
pp. 41-52



Predicting oral reading fluency in Korean EFL learners using speech and reading comprehension measures*

Jimin Suh¹ · Kyeongjin Jo² · Sun-Young Oh³ · Sunhee Kim^{4,**}

¹Department of Foreign Language Education, Seoul National University, Seoul, Korea

²Gaewoon Middle School, Seoul, Korea

³Department of English Language Education, Seoul National University, Seoul, Korea

⁴Department of French Language Education, Learning Sciences Research Institute, Seoul National University, Seoul, Korea

Abstract

Oral reading fluency (ORF) refers to reading a text quickly, accurately, and with appropriate phrasing and expression while comprehending its meaning. However, in schools, ORF assessment often focuses on reading speed and accuracy, despite the importance of prosody and reading comprehension. In addition, ORF is typically evaluated through manual teacher scoring, placing a substantial burden on teachers in large-scale assessments. As a foundational study toward automated ORF assessment, this study proposes and validates predictive models of teacher-rated segmental accuracy and prosody using manually derived speech and reading comprehension measures from Korean middle school English as a Foreign Language (EFL) learners. Students completed a 1-minute oral reading task, 4 passage comprehension items, and 14 silent reading comprehension items. Four English teachers rated the segmental accuracy and prosody on a five-point scale. Inter-rater reliability was examined using Krippendorff's alpha, and Many-facet Rasch Measurement was used to yield the adjusted ORF scores. The analyzed speech measures included words per minute, words correct per minute, syllables correct per minute, syllables correct per second, accuracy, and F0 range. Most speech measures differed significantly across levels, except for the F0 range. Predictive models combining speech and silent reading comprehension measures outperformed other models, supporting their use in future automated ORF assessments.

Keywords: oral reading fluency, EFL, speech measures, reading comprehension, automatic assessment, predictive modeling

1. Introduction

Oral reading fluency (ORF) refers to the ability to read text quickly, accurately, and with appropriate phrasing and expression,

while comprehending its meaning (Rasinski, 2004; Valencia et al., 2010). It comprises three components: accuracy, automaticity, and prosody, which interact complementarily to facilitate learners' reading comprehension (Daane et al., 2005; Rasinski, 2004).

* This study was supported by the Learning Sciences Research Institute at Seoul National University (No. 0767-2025011).

** sunhkim@snu.ac.kr, Corresponding author

Received 8 August 2026; Revised 10 September 2026; Accepted 10 September 2026

© Copyright 2026 Korean Society of Speech Sciences. This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

According to automaticity theory, as word recognition becomes automated, cognitive resources can be allocated to comprehension processes (LaBerge & Samuels, 1974). In this regard, automaticity is closely associated with reading speed (Fuchs et al., 2001). Moreover, oral reading provides opportunities to observe and evaluate learners' reading processes, making it particularly useful in classroom contexts (Hasbrouck & Tindal, 2017; Jeon & Yamashita, 2022).

The assessment of ORF has traditionally relied on words correct per minute (WCPM), a measure developed within the curriculum-based measurement (CBM) system for efficiently monitoring American elementary school learners' reading progress (Deno & Marston, 2006; Kuhn et al., 2010). However, WCPM, which measures word-level accuracy and reading speed, provides limited insight into prosodic aspects related to reading comprehension (Kuhn & Stahl, 2003). This limitation highlights the need to assess prosody separately from accuracy and reading speed (Valencia et al., 2010). One of the most representative measures for prosodic evaluation is the fluency scale developed by the National Assessment of Educational Progress (NAEP), which uses a four-point rating system to evaluate three dimensions of oral reading: phrasing, adherence to the author's syntax, and expressiveness (Pinnell et al., 1995). In addition, Zutell & Rasinski (1991) proposed the Multidimensional Fluency Scale, which assesses fluency based on 4 constructs: expression and volume, phrasing, smoothness, and pace. Although these scales are commonly used, their reliance on human judgment makes them difficult to apply consistently in large-scale assessments.

To address this limitation and reduce teachers' scoring demands, Bolaños et al. (2011) proposed the FLuent Oral Reading Assessment (FLORA) system, which automatically computes WCPM. Building on FLORA, Bolaños et al. (2013) used its lexical and prosodic features to automate expressive reading assessment according to the NAEP scale, thereby extending automated ORF assessment beyond WCPM to include prosodic aspects of oral reading. However, FLORA does not include measures of reading comprehension. Subsequently, Microsoft introduced "Reading Progress" in Microsoft Teams, which automatically reports indicators such as Correct words per minute (WPM), accuracy, and expression, and enables the assessment of reading comprehension (Microsoft, 2026). Utilizing acoustic feature extraction software such as openSMILE, Kim et al. (2021) identified 72 speech features that best predicted students' WCPM, while Wang et al. (2024) determined a set of nine optimal features that most accurately predicted human ratings based on the NAEP scale. Although these automated approaches have successfully predicted WCPM or human-rated NAEP fluency scores, they remain limited in capturing only certain aspects of overall ORF performance.

In the EFL context, early studies on ORF excluded prosodic evaluation due to its subjectivity, focused on the relationship between reading speed, accuracy, and reading comprehension. These studies reported that reading speed and reading comprehension are not entirely independent, but their correlation is weak (Fujita & Yamashita, 2014; Jeon, 2012). More recent research, however, incorporating prosody scores into the analysis, has shown that prosody can serve as a key predictor of learners' reading comprehension performance (Hannah et al., 2025; Hwang, 2025). These findings suggest the importance of including prosody in ORF assessment in the EFL context as well.

Recent studies in the EFL context have aimed to quantify prosodic and fluency-related aspects of ORF using temporal and acoustic measures. Commonly used quantitative indices that capture temporal aspects of foreign language learners' fluency include WPM and syllables per second (SPS); Hieke, 1985; Lennon, 1990). It has been reported that higher proficiency is generally associated with faster WPM (Lennon, 1990), while SPS has been shown to distinguish intermediate and advanced learners more precisely (Hieke, 1985). Examining both temporal and prosodic features in the English oral reading of Korean middle school EFL learners, Kim (2022) reported that SPS and F0 range were identified as the variables that best explained fluency levels, whereas rhythm, measured using the normalized Pairwise Variability Index (nPVI), did not show significant explanatory power. Focusing more specifically on prosody, Kallio (2024) compared data from English learners of different first-language backgrounds who speak typologically similar languages to English and identified nPVI and F0 range as key predictors of prosodic proficiency.

Despite these advancements, commercially available automated ORF assessment systems remain limited in their applicability to diverse EFL contexts, particularly due to a lack of alignment with local curricular standards. To address the limitations identified above, the present study aims to identify speech and reading comprehension measures that can inform future automation. Specifically, as a foundational study toward automated ORF assessment in EFL contexts, we propose and validate predictive models that estimate teacher-rated ORF scores using speech measures from oral reading recordings and comprehension data from Korean middle school EFL learners. The speech measures were manually derived to provide reference values for comparison with future automated extraction methods. Accordingly, this study (1) examines the relationship between ORF scores and reading comprehension measures; (2) investigates differences in speech measures across ORF proficiency levels; and (3) evaluates predictive models for estimating ORF scores using speech and reading comprehension measures.

2. Methods

2.1. Oral Reading Fluency (ORF) data acquisition

The recording and the two sets of reading comprehension tests were conducted using Microsoft Teams. The Reading Progress function automatically provided feedback such as Correct WPM, accuracy, detailed error types, and monotone score, which were collected as reference data. However, they were excluded from the present analysis due to their low reliability as evaluated by the teachers.

Participants were recruited from one middle school in Seoul, Korea. Data collection was conducted over two class sessions, and students with incomplete oral reading or reading comprehension data due to absence from either session were excluded from the final analysis. A total of 237 Korean middle school students participated in a one-minute oral reading recording session and took two reading comprehension tests. The participants' mean age was 14.01 years ($SD=0.87$; range=13–15) including 116 male and 121 female students.

Table 1. Oral reading fluency (ORF) scoring rubric

Criterion	A	B	C	D	E
Segmental accuracy	Pronounces almost all phonemes accurately and self-corrects when errors occur.	Pronounces most phonemes accurately, though occasional errors appear.	Pronounces most phonemes accurately, but some recurring pronunciation errors are present.	Frequently makes errors and omissions when pronouncing phonemes.	Very frequently makes errors and omissions when pronouncing phonemes.
Prosody	Reads primarily in larger meaning units, using appropriate stress, rhythm, and intonation.	Reads mostly in meaning units, but sometimes the grouping is inappropriate; occasionally shows inappropriate stress, rhythm, and intonation.	Reads mainly in two-word phrases or with awkward phrasing; some recurring instances of inappropriate stress, rhythm, and intonation.	Reads mainly word by word, rarely in two-word units; shows little use of appropriate stress, rhythm, and intonation.	Reads mostly word by word, with little to no evidence of appropriate stress, rhythm, or intonation.

2.1.1. Speech data acquisition

A reading text for assessing ORF was selected from passages in first-grade middle school English textbooks of the revised 2015 Korean national curriculum. Some proper nouns and expressions were modified in order to reduce lexical effects and the passage length was adjusted to fit a one-minute oral reading task. The final text was a narrative passage consisting of 127 words and 166 syllables, with a Flesch–Kincaid Grade Level (FKGL) readability index of 3.11.

The recordings were conducted in a designated classroom, where students recorded individually in soundproof booths using iPads. Speech recordings were collected using a headset microphone (Corsair HS65 Surround, Corsair, Milpitas, CA, USA). The recorded audio files were saved in either WEBM or MP4 format, and all files were converted to WAV format (16 kHz, mono, 16-bit PCM) using FFmpeg (version 4.4.2) in the Google Colab environment.

An ORF scoring rubric was devised to rate students' oral reading performance on a five-point scale. It was designed on the basis of existing ORF scoring rubrics (Council of Europe, 2020; Han et al., 2024; Pinnell et al., 1995; Zutell & Rasinski, 1991), taking into account the achievement standards of the national curriculum (Ministry of Education, 2022). Two constructs were evaluated: segmental accuracy and prosody. In this study, segmental accuracy focuses on the detailed assessment of speech segment clarity (Han et al., 2024). Although this construct is not typically included in native-speaker fluency assessments, it is particularly relevant in EFL oral reading contexts. Table 1 provides the ORF scoring rubric (Jo et al., 2025).

Four middle school English teachers participated as raters, and rater training was held prior to scoring. A partially crossed rating design (Eckes, 2015) was applied so that each learner's recording was randomly assigned to two raters, and all ratings were conducted independently.

To examine the reliability of scores assigned by multiple raters, Krippendorff's alpha ($K-\alpha$) was employed. To yield adjusted average measures for use as the ORF scores, many-facet Rasch measurement (MFRM) was employed. A Rasch model including (a) examinee, (b) rater, and (c) criterion as facets was constructed. The MFRM analysis was conducted by Minifac64 (Facets 64-bit Student/Evaluation, version 4.2.2; Linacre, 2024).

2.1.2. Reading comprehension data acquisition

To assess reading comprehension of the given ORF reading text, 4 questions (RC4) were developed, encompassing main idea, detail,

and inference types. Each student answered the questions just after each recording session. The texts used to assess silent reading comprehension (RC14) were also selected from first-grade middle school English textbooks. Following the format of the RC4 question design, a total of 14 items were developed. The reading comprehension items were developed by the research team specifically for this study and subsequently reviewed by middle school English teachers to evaluate their appropriateness for middle school learners. The review focused on text difficulty, the wording of question stems, passages, and answer options, and the identification of potential item errors. The items were revised based on the reviewers' feedback. The genre, number of items, word count, and FKGL readability index for each RC text after revisions, are presented in Table 2. RC4 (maximum score=4) and RC14 (maximum score=14) were collected as reading comprehension measures. Sample items are provided in the Appendix.

Table 2. Characteristics of the reading comprehension (RC) texts

Variables	RC4	RC14 1	RC14 2	RC14 3
genre	narrative	expository	expository	narrative
items	4	5	5	4
word count	127	234	177	255
FKGL	3.11	2.47	2.54	3.38

2.2. Speech Measures

In this study, we use six speech measures, grouped into four categories: one speed measure, three combined accuracy + speed measures, one accuracy measure, and one acoustic measure, as follows.

Speed measure

- WPM: total number of words read, excluding omissions but including errors, within one minute (Valencia et al., 2010).

Accuracy + Speed measures

- WCPM: number of correctly read words, excluding both omitted and error words, within one minute (Valencia et al., 2010).
- Syllables correct per minute (SCPM): the number of correctly read syllables per minute, during the 60-second reading task.
- Syllables correct per second (SCPS): the number of correctly read syllables divided by each learner's actual speaking duration (s). SCPS was calculated as follows:

$$SCPS = \frac{\text{Number of correctly read syllables}}{\text{Actual speaking duration (s)}} \quad (1)$$

Accuracy measure

- Accuracy: ratio of WCPM to WPM (Hwang, 2025)

Acoustic measure

- F0_range: difference between F0_max and F0_min in semitones.

WPM and WCPM were calculated manually by three researchers including an author of this paper through joint listening and word-by-word annotation of errors in learners' oral readings. As word-level units have variations in word length (James, 1956), the current study proposed SCPM and SCPS to refine the WCPM metric with syllable-level precision. Specifically, SCPM was defined by applying the WCPM framework to syllables rather than words. Because the reading task was limited to 60 seconds, the number of correctly read syllables during the task directly represented the number of correct syllables per minute. SCPS was computed as described earlier, based on the concept of SPS (Hieke, 1985), incorporating an accuracy component by calculating the number of correctly produced SPS to more accurately capture learners' fluency patterns. Actual speaking duration was measured using Praat (Boersma & Weenink, 2025) from the manually identified onset to the offset of each learner's oral reading. This duration could vary due to delayed onset or completion before the 60-second limit. Although SCPM and SCPS are both based on the number of correctly read syllables, they differ in how speaking time is incorporated. Learners who correctly produce the same number of syllables could have the same SCPM but different SCPS values if their actual speaking durations differ.

In terms of prosodic features, F0 range was included based on previous studies identifying it as an important acoustic measure associated with L2 fluency and prosodic proficiency (Kallio, 2024; Kim, 2022). F0-related parameters, including F0 mean, F0 minimum, and F0 maximum values extracted across the entire utterance using Praat (Boersma & Weenink, 2025). The pitch floor and ceiling were set to 75 and 500 Hz, respectively, and the time step was set to the automatic default. No additional trimming or exclusion of anomalous F0 values was applied. To account for anatomical differences between male and female speakers, semitone conversion was applied (Zhang, 2018). F0 values were converted to semitones using the following equation, where REF represents each speaker's mean F0:

$$\text{Semitone} = 12 \times \log^2 \left(\frac{F0}{REF} \right) \quad (2)$$

2.3. Statistical Analysis

All statistical analyses were conducted using RStudio (version 2025.09.2+418; RStudio Team, 2025) running on R (version 4.5.1) (R Core Team, 2025). To examine associations between ORF scores and reading comprehension measures, we conducted Pearson correlation analysis. To support the use of speech measures as predictors of ORF scores, we investigated whether they demonstrated significant differences across levels using the Kruskal–Wallis test. Effect sizes (ϵ^2) and Bonferroni-adjusted

Dunn's post-hoc tests were also reported.

To predict ORF scores, we used Elastic Net regularized regression (Zou & Hastie, 2005). Models were fitted using the glmnet package (Friedman et al., 2010), and all predictor variables were standardized (centered and scaled) to eliminate differences in measurement scale. Hyperparameter tuning for the Elastic Net model was performed via repeated 10-fold cross-validation (10 folds $\times$ 3 repeats) implemented in the caret package (Kuhn, 2008). The tuning grid included α values from 0 to 1 in increments of 0.1 and 50 λ values logarithmically spaced from 0.001 to 10. The optimal combination of α and λ was selected based on the lowest cross-validated root mean squared error (RMSE). The same cross-validation folds were used across all models to ensure fair comparison. Differences in RMSE across models were examined using Friedman tests (Friedman, 1937), followed by Nemenyi post-hoc tests (Nemenyi, 1963).

3. Results

3.1. Oral Reading Fluency (ORF) data

Table 3 presents descriptive statistics for ORF scores and RC4, RC14 measures across levels, reported as means (M) and SD. Segmental accuracy (SEG_avg) and prosody (PRO_avg) scores are represented as the average of the raw ratings.

Table 3. Descriptive statistics of ORF-related measures by levels

Level	ORF score	SEG_avg	PRO_avg	RC4	RC14
A	4.6±0.3	4.8±0.4	4.5±0.5	3.3±0.9	12.6±1.8
B	3.6±0.3	3.7±0.5	3.4±0.4	2.7±1.1	10.8±2.6
C	2.4±0.3	2.4±0.4	2.4±0.4	2.3±1.0	8.8±2.9
D	1.3±0.3	1.3±0.3	1.3±0.3	2.1±0.9	7.1±2.6

Values are presented as mean±SD.

ORF, oral reading fluency; PRO_avg, prosody; RC, reading comprehension; SEG_avg, segmental accuracy.

The K- α values for the segmental accuracy and prosody categories were 0.85 and 0.71, respectively, indicating high and moderate levels of reliability (Hayes & Krippendorff, 2007). Before interpreting the MFRM results, the assumptions of unidimensionality and overall model fit were examined, and both were found to be acceptable. Infit mean-square (MnSq) and Outfit MnSq values were examined to examine rater consistency within the MFRM framework. The resulting MFRM fit statistics for each rater are presented in Table 4. A criterion range of 0.50 to 1.50 was adopted, and all values were found to fall within this acceptable range (Linacre, 2020).

Table 4. Many-facet Rasch measurement (MFRM) fit statistics for raters

Rater	Measure	Infit		Outfit	
		MnSq	Zstd	Mnsq	Zstd
R2	0.86	0.85	−1.6	0.85	−1.5
R4	0.47	1.05	0.4	1.13	1.1
R1	−0.62	0.81	−1.7	0.81	−1.4
R3	−0.71	1.18	1.7	1.15	1.2

Figure 1 presents the MFRM category probability curves, which illustrate the functioning of the five-point rating scale. The peak of each curve represents the ability level at which a particular score category is most likely to be assigned. The five score categories were sequentially ordered along the ability continuum, with higher score categories becoming more probable as ability increased. Furthermore, no category disordering was observed, indicating that the five-point rating scale functioned in an ordered manner.

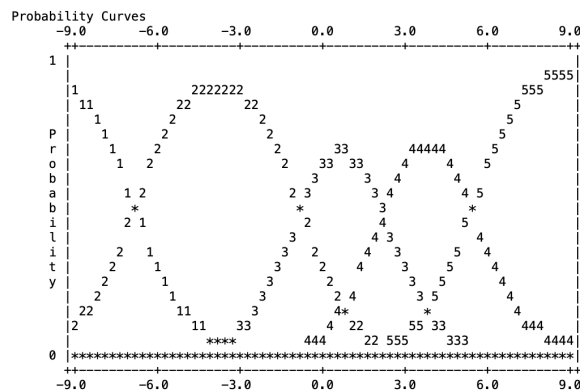


Figure 1. Many-facet Rasch measurement (MFRM) category probability curves for the five-point rating scale.

The mean adjusted ORF score was 3.61 ($SD=1.19$; range=1.05–4.95). To classify learners into proficiency levels, the one-point interval structure of the 1–5 scoring rubric used in this study was maintained. This approach also took into account that the rubric was developed based on the middle school national curriculum and that the lowest score specified in the curriculum is 1. Due to the MFRM score adjustment, however, no student reached the maximum adjusted score of 5, even when full marks were given for both segmental accuracy and prosody in the raw ratings. Accordingly, scores of 4.0 or higher were classified as Level A, scores from 3.0 to below 4.0 as Level B, scores from 2.0 to below 3.0 as Level C, and scores from 1.0 to below 2.0 as Level D.

3.2. Data Analysis

Figure 2 illustrates the correlations among ORF scores, and RC scores. ORF scores showed moderate correlations with RC4 ($r=.47$) and stronger correlations with RC14 ($r=.65$). Both segmental accuracy and prosody showed moderate correlations with RC4 ($r=.46$ and $.47$, respectively) and stronger correlations with RC14 ($r=.62$ and $.65$, respectively). RC4 and RC14 showed a moderate correlation ($r=.45$).

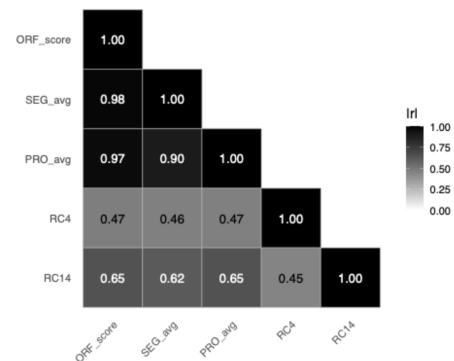


Figure 2. Pearson correlation matrix between the oral reading fluency (ORF) scores and reading comprehension (RC) scores.

Figure 3 illustrates the correlations among ORF scores, and speech measures including WPM, WCPM, SCPM, SCPS, Accuracy, F0_range. ORF scores were most strongly correlated with SCPM ($r=.92$), whereas F0_range showed negligible correlation with ORF scores ($r=.07$). Extremely high intercorrelations were observed among speed related measures ($r=.96-.99$), indicating substantial multicollinearity.

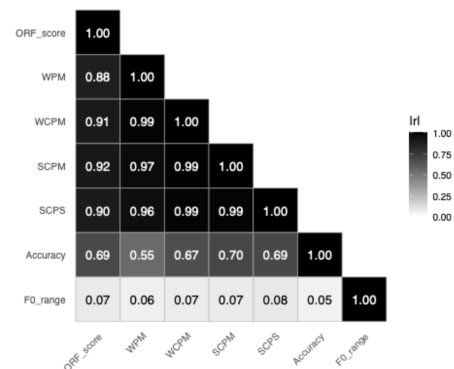


Figure 3. Pearson correlation matrix between the oral reading fluency (ORF) scores and speech measures.

Table 5 presents descriptive statistics for speech measures across ORF levels, including WPM, WCPM, SCPM, SCPS, Accuracy, F0_range, reported as M and SD. Sample sizes for each level are also reported in the table. The Kruskal–Wallis test revealed differences across levels for WPM, WCPM, SCPM, SCPS, and Accuracy (all $p<.05$; $\epsilon^2=.43-.7$). Higher ORF proficiency was consistently associated with higher values across all measures showing significant differences, with large effect sizes indicating substantial differences among proficiency levels. In contrast, F0_range showed no significant differences across levels ($p\geq.05$; $\epsilon^2=.02$). Dunn's post-hoc tests revealed that for

Table 5. Descriptive statistics of speech measures across levels

Level	n	WPM	WCPM	SCPM	SCPS	Accuracy	F0_range
A	115	106.62±16.65	100.97±17.47	132.79±22.90	2.34±0.48	94.52±3.91	29.27±4.93
B	48	80.85±14.52	72.06±12.33	94.63±16.28	1.62±0.28	89.31±4.31	29.40±5.94
C	54	58.70±16.04	48.37±14.18	61.00±18.25	1.05±0.31	82.43±9.24	28.54±5.68
D	20	32.50±14.17	25.65±11.52	30.15±14.75	0.53±0.24	79.73±13.56	27.65±6.00

Values are presented as mean±SD.

SCPM, syllables correct per minute; SCPS, syllables correct per second; WCPM, words correct per minute; WPM, words per minute.

WPM, WCPM, SCPM, and SCPS, most pairwise comparisons were significant after Bonferroni adjustment, except for the comparison between levels C and D (adjusted p -values $\geq$.05).

Among the variables, accuracy showed a distinct post-hoc pattern. For accuracy, no significant differences were observed between proficiency levels B and D or between C and D after Bonferroni adjustment (adjusted p -values $\geq$.05).

3.3. Predictive Model

In this study, four Elastic Net prediction models (Models A–D) were compared according to the composition of the input variables. Each model was defined based on the set of predictors included, as follows.

- Model A: included only speech measures extracted from the oral reading recordings, namely WPM, WCPM, SCPM, SCPS, Accuracy, and F0_range.
- Model B: extended Model A by adding the RC4 score.
- Model C: extended Model A by adding the RC14 score.
- Model D: included all measures, resulting in a total of eight predictors.

Table 6 reports hyperparameters (α and λ) selected from 10-fold cross-validation performance metrics (RMSE $\pm$ SD and R^2) for each model. The RMSE and R^2 values represent the mean performance across 10-fold cross-validation repeated three times at the optimal hyperparameter settings.

Table 6. Optimal hyperparameters and performance metrics for each model

Model	α	λ	RMSE $\pm$ SD	R^2
Model A	0.1	0.001	0.459 $\pm$ 0.075	0.848
Model B	0.1	0.001	0.460 $\pm$ 0.077	0.848
Model C	0.1	0.001	0.445 $\pm$ 0.074	0.858
Model D	0.1	0.001	0.446 $\pm$ 0.074	0.857

Model C exhibited the best predictive performance. The Friedman test revealed a statistically significant difference in RMSE across the four models [$\chi^2(3)$ =27.64, p <.001]. Nemenyi post-hoc comparisons further showed that Model C significantly outperformed Models A, B, and D (all p <.05), whereas no significant differences were found among Models A, B, and D.

Table 7. Selected predictors in the Model C

Selected predictors	Std β
SCPM	1.10
WCPM	−0.47
WPM	0.35
SCPS	−0.23
RC14	0.13
Accuracy	0.12
F0_range	0.02

RC, reading comprehension; SCPM, syllables correct per minute; SCPS, syllables correct per second; WCPM, words correct per minute; WPM, words per minute.

Table 7 presents the standardized regression coefficients (Std β) of the predictors selected in Model C. All candidate predictors were retained in the model. Based on the absolute values of the

standardized coefficients, SCPM showed the largest absolute standardized coefficient (Std β =1.10), followed by WCPM (Std β =−0.47) and WPM (Std β =0.35). However, because several predictors captured overlapping aspects of reading speed and accuracy, the magnitudes and signs of individual coefficients should be interpreted as conditional associations after accounting for the other predictors in the model, rather than as simple relationships with ORF scores.

4. Discussion

This study examined whether ORF scores, defined as an integrated construct of segmental accuracy and prosody, can be predicted using speech and reading comprehension measures from Korean middle school EFL learners. The results showed that ORF scores were significantly correlated with both passage-based (RC4) and silent reading comprehension measures (RC14), indicating that higher ORF proficiency is associated with stronger reading comprehension abilities. The observed correlations between ORF scores and reading comprehension are broadly consistent with previous EFL studies reporting associations between prosodic reading performance and reading comprehension (Hannah et al., 2025; Hwang, 2025).

Significant differences across ORF proficiency levels were observed for most speech measures, including WPM, WCPM, SCPM, SCPS, and accuracy, whereas F0 range did not show consistent level differences. This finding is consistent with previous studies reporting that higher L2 proficiency is associated with higher values of temporal fluency measures (Hieke, 1985; Kim, 2022; Lennon, 1990). In contrast to previous studies highlighting F0 range as an important predictor of fluency (Kallio, 2024; Kim, 2022), F0 range did not significantly differ across ORF proficiency levels in the present study. This may indicate that F0 range calculated across the entire utterance in the present study was not sufficient to distinguish prosodic differences among middle school EFL learners, reflecting a limitation of relying on F0 range alone to capture the broader prosodic characteristics represented in the teacher ratings.

Predictive modeling further revealed that Model C, which combined speech measures with the RC14 score, significantly outperformed Models A, B, and D, although the absolute differences in predictive performance among the models were small. More specifically, Model B, which added RC4 to the speech-only Model A, did not improve predictive performance, whereas Model C significantly outperformed Model B. This difference may reflect the greater discriminative power of the RC14 measure, as its larger number of items may have provided a more stable estimate of reading comprehension.

At the same time, it should be noted that the generally strong predictive performance observed across the models may partly reflect shared variance between the teacher-rated ORF scores and speech measures that incorporate an accuracy component.

In addition to conventional word-based measures, the present study incorporated syllable-based measures, including SCPM and SCPS, to examine their complementary role in predicting ORF performance.

5. Conclusion

This study proposed and validated predictive models as a foundation for future automated ORF assessment using speech and reading comprehension measures from Korean middle school EFL learners. The results showed that ORF scores were associated with reading comprehension, that all speech measures except F0 range differed across ORF proficiency levels, and that the model combining speech and reading comprehension measures achieved the strongest performance in predicting ORF scores. In particular, SCPM showed the largest absolute standardized coefficient in the predictive model, while high correlations were observed among several speech measures.

A key contribution of this study is the development of a curriculum-based ORF scoring rubric and the construction of a speech dataset. Our study extends the traditional ORF framework by providing empirical support for the inclusion of prosodic and comprehension-related features beyond accuracy and speed. Furthermore, for EFL learners, we argue that segmental accuracy should be included as a component of ORF assessment, given the importance of pronunciation in oral reading performance. Our approach incorporates information particularly relevant to EFL contexts while retaining conventional ORF measures. Overall, the findings provide empirical evidence that may inform the future development of automated ORF assessment using speech and reading comprehension data.

The following limitations should be noted. First, the participants were recruited from a single middle school in Seoul, which may limit the generalizability of the findings. Second, although the RC items underwent expert review, internal consistency reliability could not be estimated retrospectively because only total scores were retained for analysis. Finally, the utterance-level F0 range used in this study represents only one aspect of prosody and may not fully capture individual differences in prosodic fluency.

Future research should explore the implementation of speech and phoneme recognition technologies to automatically extract the measures examined in this study. It should also examine learners across a broader range of proficiency levels and conduct external validation using independent samples from different schools and reading passages varying in difficulty and genre. In addition, a broader set of prosodic measures, including percentile-based F0 range, pause measures, rhythm, and phrase-level intonation, should be investigated to better capture prosodic expressiveness.

Acknowledgement

This study was supported by the Learning Sciences Research Institute at Seoul National University (No. 0767-2025011).

References

- Boersma, P., & Weenink, D. (2025). Praat: Doing phonetics by computer (version 6.4.48) [Computer program]. Retrieved from <http://www.praat.org/>
- Bolaños, D., Cole, R. A., Ward, W., Borts, E., & Svirsky, E. (2011). FLORA: Fluent oral reading assessment of children's speech. *ACM Transactions on Speech and Language Processing*, 7(4), 16.
- Bolaños, D., Cole, R. A., Ward, W. H., Tindal, G. A., Hasbrouck, J., & Schwanenflugel, P. J. (2013). Human and automated assessment of oral reading fluency. *Journal of Educational Psychology*, 105(4), 1142-1151.
- Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment - Companion Volume*. Strasbourg, France: Council of Europe Publishing.
- Daane, M. C., Campbell, J. R., Grigg, W. S., Goodman, M. J., & Oranje, A. (2005). *Fourth-grade students reading aloud: NAEP 2002 special study of oral reading. The nation's report card. NCES 2006-469*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.
- Deno, S. L., & Marston, D. (2006). Curriculum-based measurement of oral reading: An indicator of growth in fluency. In S. J. Samuels & A. E. Farstrup (Eds.), *What research has to say about fluency instruction* (pp. 179-203). Newark, DE: International Reading Association.
- Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd rev. upd. ed.). Frankfurt, Germany: Peter Lang Verlag.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200), 675-701.
- Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1-22.
- 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.
- Fujita, K., & Yamashita, J. (2014). The relations and comparisons between reading comprehension and reading rate of Japanese high school EFL learners. *The Reading Matrix*, 14(2), 34-49.
- Han, S., Kim, S., & Chung, M. (2024, May). Constructing Korean learners' L2 speech corpus of seven languages for automatic pronunciation assessment. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 3772-3781). Turin, Italy.
- Hannah, L., Jang, E. E., Lee, M. H., & Russell, B. (2025). Investigating construct representativeness and linguistic equity of automated oral reading fluency assessment with prosody. *Language Testing*, 42(4), 397-422.
- Hasbrouck, J., & Tindal, G. (2017). *An update to compiled ORF norms* (Technical Report # 1702). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1), 77-89.
- Hieke, A. E. (1985). A componential approach to oral fluency evaluation. *The Modern Language Journal*, 69(2), 135-142.
- Hwang, I. (2025). Exploring the relationship between textbook-based oral reading fluency and L2 reading comprehension among Korean EFL learners. *English Language Teaching*, 37(1), 39-61.
- James, T. L. (1956). Standardising shorthand speed tests: I. The

- standard word. *The Vocational Aspect of Secondary and Further Education*, 8(17), 138-148.
- Jeon, E. H. (2012). Oral reading fluency in second language reading. *Reading in a Foreign Language*, 24(2), 186-208.
- Jeon, E. H., & Yamashita, J. (2022). L2 reading comprehension and its correlates. In E. H. Jeon, & Y. In'nami (Eds.), *Bilingual processing and acquisition (BPA): Understanding L2 proficiency* (pp. 29-86). Amsterdam: John Benjamins.
- Jo, K. J., Suh, J. M., Oh, S. Y., Kim, S. H., & Jin, H. Y. (2025, September). Exploring the usefulness of oral reading fluency assessment in the Korean EFL secondary classroom. *Proceedings of the 2025 ALAK International Conference & AILA East Asia Forum*. Seoul, Korea.
- Kallio, H. (2024). The contribution of speech timing, F0 change, and voice quality to perceived prosodic proficiency in L2: A cross-lingual perspective. *Proceedings of Speech Prosody 2024* (pp. 507-511). Grenoble, France: International Speech Communication Association (ISCA).
- Kim, H., Hannah, L., & Jang, E. E. (2021, April). Using acoustic features to predict oral reading fluency of students with diverse language backgrounds. *Collated Papers for the ALTE 7th International Conference* (pp. 198-202). Madrid, Spain.
- Kim, H. Y. (2022). The relationship between fluency levels and suprasegmentals according to the sentence types in the English read speech by Korean middle school English learners. *Phonetics and Speech Sciences*, 14(3), 51-66.
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 1-26.
- Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly*, 45(2), 230-251.
- Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, 95(1), 3-21.
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6(2), 293-323.
- Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40(3), 387-417.
- Linacre, J. M. (2020). Winsteps Rasch measurement computer program: User's guide. Retrieved from <https://www.winsteps.com/a/Winsteps-Manual.pdf>
- Linacre, J. M. (2024). Minifac64: Facets 64-bit student/evaluation computer program for many-facet Rasch measurement (version 4.2.2). Beaverton, OR: Winsteps.com. Retrieved from <https://www.winsteps.com/>
- Microsoft. (2026). Getting started with reading progress in teams. Retrieved from <https://support.microsoft.com/en-us/topic/getting-started-with-reading-progress-in-teams-7617c11c-d685-4cb7-8b75-3917b297c407>
- Ministry of Education. (2022). *2022 Revised national curriculum*. Sejong, Korea: Ministry of Education.
- Nemenyi, P. B. (1963). *Distribution-free multiple comparisons* (Doctoral dissertation). Princeton University, Princeton, NJ.
- Pinnell, G. S., Pikulski, J. J., Wixson, K. K., Campbell, J. R., Gough, P. B., & Beatty, A. S. (1995). *Listening to children read aloud: Oral fluency*. Washington, DC: US Department of Education, National Center for Education Statistics.
- Rasinski, T. V. (2004). *Assessing reading fluency*. Honolulu, HI: Pacific Resources for Education and Learning (PREL).
- R Core Team. (2025). R: A language and environment for statistical computing (version 4.5.1) [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <https://www.R-project.org/>
- RStudio Team. (2025). RStudio: Integrated development environment for R (version 2025.09.2+418) [Computer software]. Boston, MA: RStudio, PBC. Retrieved from <https://www.rstudio.com/>
- 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(3), 270-291.
- Wang, K., Qiao, X., Sammit, G., Larson, E. C., Nese, J., & Kamata, A. (2024). Improving automated scoring of prosody in oral reading fluency using deep learning algorithm. *Frontiers in Education*, 9, 1440760.
- Zhang, J. (2018, December). A comparison of tone normalization methods for language variation research. *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation* (pp. 823-831). Hong Kong.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2), 301-320.
- Zutell, J., & Rasinski, T. V. (1991). Training teachers to attend to their students' oral reading fluency. *Theory Into Practice*, 30(3), 211-217.

• **Jimin Suh**

Ph.D. Candidate, Dept. of Foreign Language Education
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-2-880-7691
Email: gkssk1@snu.ac.kr
Areas of interest: L2 prosody, Oral reading fluency, Second language acquisition, French

• **Kyeongjin Jo**

Teacher, Gaewoon Middle School
935 Bugaksan-ro, Seongbuk-gu,
Seoul 02808, Korea
Tel: +82-2-3291-5492
Email: kdskj0229@snu.ac.kr
Areas of interest: L2 reading, Oral reading fluency

• **Sun-Young Oh**

Professor, Dept. of English Language Education, Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-2-880-7675
Email: sunoh@snu.ac.kr
Areas of interest: Learner corpus studies, The use of corpora and technology in English language teaching

• **Sunhee Kim**, Corresponding author

Professor, Dept. of French Language Education, Research
Affiliate, Learning Sciences Research Institute, Seoul National

University

1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea

Tel: +82-2-880-7693

Email: sunhkim@snu.ac.kr

Areas of interest: French, Phonetics, Phonology, Foreign
language pedagogy, Application of linguistics and language
technologies in language instruction and assessment, Speech
synthesis, Speech and language processing

Appendix

● ORF passage reading comprehension (RC 4)

My name is Tina. I'm a middle school student in South Africa. Last Saturday, my volunteer club went to an elephant home near my school. There were many poor elephants in the elephant home. We helped them.

Our day started early. First, we fed the elephants. Elephants eat a lot, so we brought a lot of food to them. There was one baby elephant. I fed milk to him. He was so cute!

After lunch, we took the elephants to a pond for a mud bath. Elephants' skin is very sensitive, and mud protects it from the sun. Their bath time was my favorite part of the day.

Our work was over around six. My legs felt heavy, but I had a big smile on my face.

1. 윗글의 소재로 가장 적절한 것은? (중심내용)
 - a) 남아프리카로의 여행
 - b) 양로원에서의 봉사활동
 - c) 학교에서의 즐거운 시간
 - d) 코끼리 보호소에서의 하루
2. 다음 중 윗글의 Tina가 지난 토요일에 가장 먼저 한 일은? (세부 내용)
 - a) 코끼리 씻기기
 - b) 코끼리 먹이 주기
 - c) 코끼리 집 청소하기
 - d) 코끼리와 사진 찍기
3. 윗글에서 코끼리가 목욕할 때 사용한 것은? (세부 내용)
 - a) 풀
 - b) 모래
 - c) 비누
 - d) 진흙
4. 윗글의 Tina가 느꼈을 감정으로 가장 적절한 것은? (추론)
 - a) 실망하고 지루함
 - b) 피곤하지만 행복함
 - c) 혼란스럽고 긴장됨
 - d) 긴장됐지만 즐거움

● Silent reading comprehension (RC 14)

Text 1

What do you love to do? Do you love to play the piano? Do you love to watch television? Well, those activities can become your future jobs. Let's read about the following people. They found the perfect jobs for them.

Rahul

Rahul loved to sing. He sang on the way to school. He sang in the shower. He even sang in his sleep! When Rahul was in middle school, he acted in a school play. At that time, he fell in love with acting, too. Rahul is now an actor in musicals. He may not be a good dancer, but he still enjoys being an actor.

Mina

Mina always played with her dolls. She changed their clothes every day. She ate and slept with her dolls. Luckily, Mina found the perfect job. She is a doll dress designer. She makes traditional dresses for dolls around the world. Her favorite traditional dress is hanbok.

Kevin

Kevin wanted to be a doctor. He also loved sports. Now, he works as a doctor for a baseball team. Kevin travels all over the country with his team. He takes care of the players when they get hurt. Kevin loves to help the players. He also loves to watch the games up close!

What do you think? People spend about 30 percent of their lives at their jobs. So, how will you spend this 30 percent of your life?

1. 이 글의 제목으로 가장 적절한 것은? [중심내용]
 - a) 노래를 잘 부르는 법
 - b) 운동 선수가 되는 법
 - c) 다양한 나라의 전통 의상
 - d) 좋아하는 일을 직업으로 삼은 사람들
2. 윗글의 Rahul에 대한 설명으로 적절하지 않은 것은? [세부내용]
 - a) 노래 부르는 것을 좋아했다.
 - b) 연기하는 것을 좋아하지 않았다.
 - c) 중학교 때 학교 연극에 참여했다.
 - d) 현재는 뮤지컬 배우이다.
3. 윗글의 Mina의 직업에 대한 설명으로 가장 적절한 것은? [세부내용]
 - a) 인형을 판매한다.
 - b) 인형을 직접 만든다.
 - c) 인형 옷을 디자인한다.
 - d) 사람을 위한 한복을 디자인한다.
4. 윗글의 Kevin에 대한 설명으로 가장 적절한 것은? [세부내용]
 - a) 운동 선수가 되었다.
 - b) 병원에서 일하는 의사이다.
 - c) 종종 야구 경기에 참여한다.
 - d) 야구팀과 함께 다니며 선수들을 돌본다.

5. 윗글의 밑줄 친 문장을 통해 알 수 있는 것은? [추론]
- a) Rahul은 댄서가 되기 위해 노력하고 있다.
 - b) Rahul은 노래 실력이 부족해서 배우가 되지 못했다.
 - c) Rahul은 춤은 못추지만 자신이 좋아하는 일을 한다.
 - d) Rahul은 춤을 잘 추기 때문에 뮤지컬 배우가 되었다.

Text 2

What's for lunch at your school? Let's read about the different school lunches around the world!

[Polina, Russia]

Today I had kasha and compote for lunch. Russians started to eat kasha, a hot porridge, 1,000 years ago. And we still love to eat it. I also had compote, a kind of drink. I like the sweet and colorful fruits in the glass. I want to drink it every day!

[Lashi, India]

My mom always makes lunch for me. During lunch time, the lunch man brings my lunch box to school. I opened it and found warm chicken curry and soft naan. My friends and I shared our lunch together. Curry is always the best, isn't it?

[Pili, Spain]

I had paella for lunch today. It looks delicious, doesn't it? It is made of seafood and brown rice. People in Spain love to have it for any meal of the day. I also ate gazpacho. It's a cold vegetable soup with tomato sauce. I like it better than a hot soup. It cools down the heat of summer.

6. 윗글의 소재로 가장 적절한 것은? [중심내용]
- a) 가족 소개
 - b) 세계의 전통 놀이
 - c) 세계의 학교 점심
 - d) 가장 좋아하는 수업
7. 윗글의 Polina가 마신 음료로 가장 적절한 것은? [세부 내용]
- a) milk
 - b) water
 - c) kasha
 - d) compote
8. 윗글의 Lashi의 점심에 대한 설명으로 적절하지 않은 것은? [세부내용]

- a) 엄마가 항상 점심을 만들어 주신다.
- b) 엄마가 학교로 점심을 가져다 주신다.
- c) 치킨 커리와 난을 먹었다.
- d) 친구와 함께 점심을 먹었다.

9. 윗글의 gazpacho에 대한 설명으로 적절한 것은? [세부 내용]

- a) 해산물과 갈색 쌀로 만들어졌다.
- b) 차가운 채소 수프이다.
- c) 고기와 감자를 넣어 만든 음식이다.
- d) 주로 겨울에 먹는다.

10. 윗글을 통해 알 수 있는 것으로 가장 적절한 것은? [추론]

- a) 점심시간에는 밥과 국이 함께 나온다.
- b) 모든 나라에서는 학교에서 점심을 제공한다.
- c) 나라에 따라 다양한 점심 문화를 가지고 있다.
- d) 학생들은 점심시간에 혼자 조용히 먹는 것을 선호한다.

Text 3

My name is Anurak, and I live in a small floating village in Thailand. One night, my friends and I were watching the final match of the World Cup on TV. When the game ended, I said, "Let's start our own soccer team. Maybe we can become world champions!" We thought that it was a great idea.

However, the villagers laughed at us. They said, "Look around you. There is no land. Where are you going to play soccer?"

We were sad at first, but we got an idea. We decided to make a soccer field on the sea. First, we collected wood pieces. Then we tied and nailed them together. After we worked hard for months, we finally had our floating soccer field! We were very happy.

Our soccer field was not perfect. However, we loved our soccer field. We played soccer on it every day.

About three months later, we saw a poster about a soccer tournament. I said, "Let's enter it!" We all agreed. Then we started to practice hard for the tournament.

Before our first game started, we were nervous. However, playing soccer on land was easier than playing on our floating field. All the villagers cheered for us. We did very well and took third place! We were proud.

At first, playing soccer was just a dream for us. Now we have our own soccer field and soccer team. Can we really be world champions some day? We believe it is possible. We will

work together and never give up.

11. 윗글의 제목으로 가장 적절한 것은? [중심내용]

- a) 월드컵의 역사
- b) 태국의 전통 놀이
- c) 바다 위의 작은 마을
- d) 바다 위에서 시작된 꿈

12. 윗글의 주인공이 만든 것으로 가장 적절한 것은? [세부내용]

- a) 축구화
- b) 축구장
- c) 축구 골대
- d) 나무 보트

13. 윗글의 마을 사람들이 처음에 주인공의 아이디어에 대해 보인 반응으로 가장 적절한 것은? [세부내용]

- a) 응원해주었다.
- b) 함께 연습했다.
- c) 조언을 해주었다.
- d) 무시하고 웃었다.

14. 윗글의 밑줄 친 문장을 통해 알 수 있는 것으로 가장 적절한 것은? [추론]

- a) 주인공의 팀은 축구를 처음 해 봤다.
- b) 주인공의 경기장이 육지의 경기장보다 안전했다.
- c) 떠 있는 경기장에서 연습한 것이 경기에 도움이 되었다.
- d) 주인공은 육지의 경기장에서 축구하는 것이 더 재밌었다.