

Automated Scoring of Oral Reading Fluency: An examination of validity evidence

Walter L. Leite¹ Krishna Sudeep Kumar¹ Jaiden Magnan¹
Stephanie Hammerschmidt-Snidarich¹ Seyedahmad Rahimi² Zoey Liu¹

¹University of Florida ²Georgia Institute of Technology

walter.leite@coe.ufl.edu,
{krishnan.sudeepk, jaiden.magnan, s.snidarich, liu.ying}@ufl.edu,
srahimi@gatech.edu

Abstract

This study provides validity evidence for automated scoring of oral reading fluency (ORF) using off-the-shelf automatic speech recognition (ASR) models. A corpus of 320 audio recordings of 63 children collected with a digital literacy platform was used to estimate words correct per minute (WCPM) with six model variants.

1 Introduction

Children need to develop Oral Reading Fluency (ORF) as a prerequisite for reading comprehension, writing, and success in other academic disciplines. Children typically start learning decoding skills in kindergarten and gradually improve ORF through elementary school by reading increasingly more complex texts and acquiring vocabulary (Maki and Hammerschmidt-Snidarich, 2022). Because of the importance of ORF, elementary school teachers regularly track its development from kindergarten to 2nd or 3rd grade. ORF is a latent variable with three components: rate, accuracy, and prosody (Kim et al., 2021). However, it is most commonly operationalized as words correct per minute (WCPM) (Valencia et al., 2010). Traditional ORF assessments require students to read for one-minute, and are manually scored by counting the number of words read, minus the errors, excluding repetitions and insertions. WCPM includes information about rate and accuracy, but ignores prosody (Kim et al., 2021). In the U.S., teachers and school administrators measure ORF mainly with standardized reading passages presented on paper by a human proctor and provide WCPM scored manually. DIBELS (Shaw and Shaw, 2002) and FASTBridge (Christ, 2018) are popular paper-based standardized assessments of ORF used nationwide. Although each test lasts one minute per student plus setup and scoring time, they are very burdensome for teachers. Assessing an entire classroom

can take substantial time, and many school districts require teachers to measure ORF once or twice per year from kindergarten to 2nd grade.

Digital reading devices are increasingly available in classrooms and homes (e.g., computers, tablets) (Alqahtani, 2020; Hou et al., 2024). These devices create an opportunity for automated administration and scoring of ORF assessments. Researchers have attempted to develop automated ORF assessments since the 90's (Mostow, 2016). However, recent advances in automatic speech recognition (ASR) technology have increased interest in this topic (Bhardwaj et al., 2022). While the U.S. has no widely administered automated ORF assessments, other countries have already deployed national automated ORF testing (Silva et al., 2021).

Beyond replacing manual paper-based testing in schools, automated ORF assessments can be built into digital literacy platforms, such as reading apps and intelligent tutoring systems (ITS) (Xu et al., 2019). This approach is a form of stealth assessment because it can be conducted while children read stories on a digital device without their awareness of testing. Stealth assessments, a type of formative assessment, seek to measure students' abilities without the burden and stress of formal educational testing (Rahimi and Shute, 2024). Digital literacy platforms with built-in ORF assessment can then use the scores to provide specific supports (e.g., decoding drills), track progress, and inform teachers.

This study aims to collect preliminary validity evidence supporting the use of state-of-the-art ASR models for automated ORF scoring for stealth assessment. These models could be integrated into digital literacy platforms used in schools and students' homes. Stealth assessment of ORF would allow teachers to monitor students' ORF growth trajectories and identify at-risk students. Stealth assessment of ORF could also provide proximal outcome measures for educational interventions.

This study addresses the research question: Do off-the-shelf ASR models provide valid and reliable ORF scores for use in stealth assessment?

2 Literature Review

2.1 Stealth Assessment

Stealth assessments [Shute et al., 2022](#) are automated learning assessments embedded in digital learning environments (e.g., reading apps, digital games) to elicit evidence of an ability or skill. It is called stealth not because the learners do not or should not know about the ongoing assessment, but because the assessment is invisible (i.e., there is no typical question in front of the learners to respond to). As learners interact with the digital learning environment, the software records their interactions, the stealth assessment model analyzes them, and the system uses the results to provide continuous ability estimates or categorical skill profiles. Then, the learning environment can use the estimates or skill profiles to adapt to learners' level of competency (e.g., change the level of task difficulty), provide appropriate learning supports (e.g., feedback), and report to the teacher or the learners themselves (e.g., dashboards) so that they can take appropriate actions to address learning challenges. Alternatively, the learning system can adapt the next challenges. The accuracy of stealth assessments depends on the quality and frequency of the signal recorded by the digital learning environment. For ORF scoring, the signals are recordings of students reading passages on digital reading devices. In classroom settings, signal quality may depend on background noise levels and the quality of the microphone and noise-filtering software used.

2.2 Oral Reading Fluency Scoring

Previous studies have used several methods to estimate WCPM from recordings of student readings. These methods typically used ASR models and/or extracted audio features to predict WCPM with machine learning. Project LISTEN scored WCPM through the Reading Tutor system [Mostow, 2016](#), which was an ITS for reading. Reading Tutor collected speech data in real time from students and analyzed it using forced alignment. This system provided real-time feedback, alerting the student whether they spoke a word correctly or incorrectly. By acknowledging self-corrections and restarts, the system avoided penalizing students incorrectly, which would otherwise inflate the system's perfor-

mance, as measured by its Word Error Rate (WER) ([Chen and Mostow, 2011](#)).

Project CORE developed a WCPM scoring model using the Baviaca ASR model ([Nese and Kamata, 2021](#)). The authors evaluated the model's performance on a sample of 13,766 audio files of 2nd- to 4th-grade students reading standardized passages for up to 90 seconds. They estimated standardized effect sizes between ASR scoring and the human criterion, controlling for passage length. They used a criterion of 0.15 standard deviations as the maximum acceptable difference between ASR-estimated WCPM and the human criterion. They found that ASR performance depended on passage length, but most results were within 0.15 SD of the human criterion. The ASR model performed acceptably with short (25 words), medium (50 words), and long (85 words) passages, but degraded to unacceptable levels with a fixed 60-minute reading time and a variable-length passage. Using the same dataset, [Kara et al.](#) predicted WCPM using relative silence times (RSTs), which are between-word silence times divided by the total silence time for the passage. They used NLP to create 40 meaningful word chunks and derived 80 features by averaging RSTs before and after each chunk. Then, they predicted WCPM using random forests with the 80 RST-based features. This approach produced predicted WCPM that correlated strongly with human-rated WCPM ($r = 0.764$).

Project FLORA addressed challenges in children's speech recognition such as higher pitch and inconsistent pronunciation ([Bolaños et al., 2011](#)). The system used the Boulder Language Technologies (BLT) ASR engine, leveraging Hidden Markov Models and Gaussian Mixture Models to improve accuracy. A critical factor of FLORA was its ability to process audio much faster than it is spoken for real-time feedback ([Bolaños et al., 2011](#)). Their system used forced alignment against a reference to calculate the WCPM. They found that the FLORA system produced WCPM scores that differed from human-rated scores by an average of 3.62 words. Cohen's kappa between FLORA's WCPM estimates and human scores ranged from 0.87 to 0.97.

2.3 Automatic Speech Recognition Models for Children's Speech

Older systems that scored WCPM produced acceptable results in some conditions but struggled with the high variability of children's speech. New

ASR systems based on wav2vec2.0 (Baevski et al., 2020) and Whisper (Radford et al., 2023) architectures promise score WCPM accurately for diverse children’s speech under a variety of classroom conditions. However, while there are many recent studies evaluating wav2vec2.0 (Gao et al., 2025; Peng et al., 2021; Getman et al., 2023; Xu et al., 2021) and Whisper (Jain et al., 2023, 2024; Molenaar et al., 2023; Liu et al., 2024; Dietz Smith et al., 2025) for children ASR using word error rate (WER) and sentence error rate (SER) metrics, research using these new ASR models for ORF scoring has been scarce.

Recent research on automated detection of reading errors (i.e., miscue detection) of children wav2vec2.0 and Whisper provide a foundation for scoring WCPM. By fine-tuning tiny, based, small, medium and large Whisper models with children speech, researchers have been able to obtain substantially reduced WER and SER (Jain et al., 2023, 2024; Molenaar et al., 2023; Liu et al., 2024; Dietz Smith et al., 2025). Dietz Smith et al. (2025) extended Whisper for miscue detection of English children’s speech by sending both the audio recording and the target text to the model, and new tokens for end-to-end (E2E) miscue detection. Sending the correct text to the model improved verbatim transcription of the Whisper medium model from 15% to 11% with the CMU Kids corpus (Eskenazi et al., 1997). The authors also expanded Whisper’s tokenizer to include the special miscue tokens <OMIT>, <SUBSTITUTE>, and <INSERT> so that the model could perform E2E by labelling miscues directly during transcription. E2E showed promising results, with macro-F1 >0.50) but consistently showed low recall, especially for substitutions.

3 Methodology

3.1 Data

We collected audio data for the current study using the Storiza digital literacy platform (Leite et al., 2025), which uses generative AI to create decodable stories for elementary school children matched to specific lessons in the UFLI Foundations phonics program (Lane and Contesse, 2022). We recruited parents in the U.S. through the UFLI Foundations Facebook Group and asked them to supervise their children as they read and recorded at least one story per week for six weeks. The stories ranged from 50 to 400 words. The audio data consisted

of 320 recordings of 63 children (37 females, 25 males, 1 not reported; kindergarten: 6; 1st grade: 22; 2nd grade: 24; 3rd grade: 11) totaling about 9 hours of audio. On average, each story is 1min29s, ranging from 7.03s to 5min36s. The geographic distribution of these participants (by ZIP code) is: 14.1% Northeast, 23.3% South, 20.9% Midwest, and 41.7% West. The platform recorded children reading stories as .wav files and stored the story text.

A team of 10 linguistics students, supervised by a linguist, performed detailed annotation of the recordings, including classifying the types of reading errors students made (Rose et al., 2026). Correct word production accounts for 70.17% of all words, consistent with the grade-level findings of Hasbrouck and Tindal (2017) for early readers. Among error categories, disfluency errors are the most frequent (10.93%), followed by phonological (7.42%), orthographic (3.65%), and grammatical errors (2.46%). Figure 1 presents the distribution of error categories in the data. Details about this corpus are presented in Rose et al..

3.2 Analysis

We applied six ASR model variants drawn from two families to estimate WCPM. From the Wav2Vec 2.0 framework (Baevski et al., 2020), we include wav2vec2-base-960h, pretrained and fine-tuned on LibriSpeech (Panayotov et al., 2015), and wav2vec2-large-xlsr-53-english, a cross-lingual model fine-tuned from wav2vec2-large-xlsr-53 on English data from Common Voice (Ardila et al., 2020). Both are encoder-only models that learn speech representations through self-supervised pretraining on unlabeled audio followed by supervised fine-tuning for ASR. From the Whisper family (Radford et al., 2023), we evaluate four model sizes (i.e., tiny, base, small, medium) which vary in parameter count and capacity. Whisper is a sequence-to-sequence model trained in a fully supervised fashion on large-scale weakly labeled web audio, and unlike Wav2Vec 2.0, generates transcriptions autoregressively through a decoder. All models are used off-the-shelf without any fine-tuning on child speech.

The estimation of WCPM involved four steps: 1) Submit the child voice recording to each ASR model variant to obtain the ASR hypothesis, which is a prediction of the words read by the child (i.e., transcription); 2) Perform word-level alignment

of the ASR hypothesis against the reading passage using the `jiwer` Python library (Jitsi, 2024). This alignment is made using the Levenshtein distance with both strings normalized to lowercase and with punctuation removed (except apostrophes, which are phonologically meaningful in contractions); 3) Identify substitutions, deletions, and insertions based on the alignment results. 4) Calculate $WCPM = ((N - S - D)/t) \times 60$, where N is the total reference word count, S and D denote substitutions and deletions, and t is the audio duration in seconds. Following the curriculum-based measurement (CBM) convention (Deno et al., 1982), WCPM is computed against the reference text using the same alignment, with only substitutions and deletions counted as errors. Insertions, such as repetitions, fillers, and extra words, are excluded, as they already penalize the reader by consuming reading time (University of Oregon, 2023; Deno et al., 1982).

We used the validity evidence framework of the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014) to evaluate the WCPM estimates. Specifically, we evaluated validity evidence with respect to internal structure, content, response process, and relationships with other variables. For internal structure evidence, we measured ASR transcription quality by computing WER and Character Error Rate (CER) between the ASR hypothesis and a human-verified transcript of the child’s actual speech, following standard ASR evaluation convention (Morris et al., 2004). Word Error Rate (WER) is defined as $WER = (S + D + I)/N$, where I denotes insertions. CER is calculated similarly to WER but at the character level. We also estimated the relative bias of WCPM, defined as $\text{mean}((WCPM_{\text{estimated}} - WCPM_{\text{annotated}})/WCPM_{\text{annotated}})$, to assess how much the ASR-model WCPM deviates from the manual score. Following Nese and Kamata, we consider the relative bias of WCPM estimates to be acceptable if smaller than 0.15 standard deviations.

For evidence related to content, we use a fixed-effects regression model (Allison, 2009) to estimate the relationships between the model-based WCPM estimates and characteristics of the passages used for ORF assessment, controlling for manual WCPM scores, fixed effects of child and week of participation. For ORF automated assessment, it is desirable that when WCPM variation from the manual WCPM scores and due to child

and time are accounted for, the residuals of WCPM scores are unrelated to the particular characteristics of the reading passage used to assess ORF. In other words, this evaluates whether the passage has no method effects (Podsakoff et al., 2003). This is important validity evidence given that traditional ORF measures such as the DIBELS (Shaw and Shaw, 2002) use expert-written standardized reading passages, while the automated ORF estimation implemented within a digital literacy platform uses a different AI-generated reading passage for every assessment (Leite et al., 2025). Therefore, finding no relationship with passage characteristics would indicate that the ORF assessment is robust to random reading passage variation. The passage characteristics we evaluated include readability, as measured by Spache readability scores (Spache, 1953), syntactic complexity (Szmrecsányi, 2004), and lexical diversity (Jarvis, 2013) as measured by the Moving-Average Type-Token Ratio (Covington and McFall, 2010).

For response process evidence, we report the test-retest reliability of WCPM scores and the relationship of test-retest reliability estimates with grade level and recording duration. We estimated test-retest reliability by identifying all pairs of recordings the same student completed within the same week. We then estimated test-retest reliability as the Pearson correlation between WCPM estimates from two recordings, which equals the intraclass correlation (ICC) with only two measures.

To assess validity with respect to relationships with other variables, we examined the relationship between WCPM and grade level. We stratify all metrics by grade level and report mean WCPM, WER, and CER per grade, following the normative framework of Hasbrouck and Tindal (2017). Grade-level stratification enables detection of systematic ASR degradation at early grades, where child speech is most acoustically distinct from adult speech (Potamianos and Narayanan, 2003).

4 Results

4.1 Internal Structure Evidence

Table 1 presents WER and CER for two Wav2Vec2 variants. Both models substantially underperformed compared to Whisper. Wav2Vec2 Large achieved a WER of 0.727 compared to Whisper medium’s 0.2613. Given their high transcription error rates, we did not use Wav2Vec2 variants for WCPM estimation.

Tables 2 and 3 present WER, CER, WCPM, and relative bias for four Whisper variants at sentence and story level. Larger models consistently yielded lower WER and CER. At both the story and sentence levels, the relative bias of WCPM showed a non-monotonic relationship with model size: Tiny and base models underestimated WCPM, while small and medium models overestimated it. However, using 0.15 standard deviations as a guideline for acceptable relative bias of WCPM (Nese and Kamata, 2021), all Whisper variants produced acceptable relative bias at the story level, and only Whisper tiny exceeded the threshold at the sentence level.

4.2 Evidence Related to Content

The results of the fixed-effects regression model shown in 4 indicate that the WCPM scores estimated with the Whisper variants were not related to passage readability or syntactic complexity. However, WCPM estimates from Whisper tiny and small were significantly and positively related to lexical diversity, indicating that when two students with the same oral reading fluency read two different passages, reading the passage with higher lexical diversity was associated with a higher estimated WCPM. Therefore, we found evidence that these Whisper variants are vulnerable to method bias due to passage. However, WCPM estimates with Whisper base and medium were not related to lexical diversity.

4.3 Response Process Evidence

Table?? presents the test-retest reliability estimates for the WCPM scores from manual annotation and Whisper models of different sizes. Overall estimates ($n = 424$) ranged from 0.693 for manual annotation to 0.756 for Whisper Small, but the confidence intervals indicate that there is no significant difference between them. However, WCPM estimates depended strongly on recording duration and grade. Test-retest reliability estimates with recordings lasting at least 60 seconds ($n = 348$) were between 0.693 and 0.782, while with recordings shorter than 60 seconds ($n = 76$), estimates were between 0.252 and 0.298. Test-retest reliability estimates were higher with 1st grade ($n = 149$) and 3rd grade ($n = 56$) students, ranging from 0.718 to 0.861. Reliability was lower with 2nd grade students ($n = 200$), ranging from 0.593 to 0.661. It was much lower with Kindergarten students ($n = 19$), ranging from 0.252 to 0.298. Within each

subgroup, model-based reliability estimates did not differ significantly from manually annotated estimates, and did not depend on model size.

4.4 Evidence of Relationship with Other Variables

Tables 6 and 7 present grade-level WCPM breakdowns for Whisper medium, which consistently shows the lowest WER and is used for grade-level analysis. Both annotated and predicted scores show consistent grade-level progression, confirming that the pipeline captures developmental fluency growth from Kindergarten through 3rd grade. The positive bias widens at higher grades from +0.043 at Kindergarten to +0.065 at 3rd Grade+ at sentence level, and from +0.056 to +0.111 at story level, indicating that subtle miscues produced by more fluent readers are increasingly harder for general-purpose ASR to detect. Story-level WCPM values are markedly lower than sentence-level for both annotated and predicted scores (54.81 vs. 76.90), as story-level aggregation captures inter-sentence pauses and passage transitions that reduce the overall reading rate.

5 Discussion

We found that the relative bias of WCPM estimates from Whisper variants had a non-monotonic relationship with model size. This non-monotonicity may stem from the interaction of two error sources: smaller models may underestimate WCPM due to high transcription error rates, while larger models may overestimate it by auto-correcting child mispronunciations. Story-level WER is slightly higher than sentence-level for most models, suggesting that error accumulation across longer contexts poses a greater challenge for ASR. The pipeline captures developmental fluency progression across grade levels, though with a positive bias that widens at higher grades as subtle miscues become harder for general-purpose ASR to detect.

Standardized oral reading fluency assessments such as DIBELS Shaw and Shaw, 2002 and FAST-BridgeChrist, 2018 use a fixed 60-second reading time. While we did not evaluate WCPM estimation from fixed 60-second passages, the evaluation of test-retest reliability provides some evidence that 60-second passages may be too short for reliable WCPM estimation using off-the-shelf Whisper variants. These results are consistent with the findings of Nese and Kamata, indicating that automated

WCPM estimation was inadequate with 60-second passages.

6 Conclusion

Off-the-shelf ASR models may offer a practical, scalable solution for automated ORF assessment that can provide substantial time savings for teachers, enable weekly monitoring of student ORF trajectories, and facilitate personalized digital interventions such as digital reading platforms and intelligent tutoring systems (Xu et al., 2019). Automated WCPM scoring can be used to estimate an individualized growth curve for each student, which is helpful to detect at-risk students. Among the models evaluated, we found that Whisper medium performed best, providing valid and reliable WCPM scores under certain conditions, such as with passages longer than 60 seconds and with 1st grade students. However, test-retest reliability was low in some conditions, such as with passages shorter than 60 seconds and with kindergarten students. Therefore, while promising, ASR pipelines need further refinement to support general application for ORF assessment in digital literacy platforms across elementary grades.

One limitation of the WCPM pipeline we used is that it relies solely on word-level ASR alignment, missing disfluency-based errors such as hesitations, self-corrections, and sounded-out words. As shown in Figure 1, disfluency is the most frequent error category in the corpus (10.93%), and Whisper suppresses these during transcription (Radford et al., 2023), directly contributing to the observed positive bias. The current study is also limited in that it operationalizes ORF solely in terms of WCPM, ignoring prosody. Recent studies show that incorporating prosody into ORF assessment improves its ability to predict reading comprehension (Wolters et al., 2022; Kim et al., 2021).

Because all models used in this study were trained on adult speech, future research should fine-tune them on child read-aloud recordings to reduce bias and increase reliability. A more diverse corpus would benefit the fine-tuning process. A limitation of the current corpus (Rose et al., 2026) is that it was recruited via a Facebook Group of teachers and reading tutors, likely oversampling motivated readers and undersampling children with specific learning disabilities, such as dyslexia. This limits the generalizability of results to struggling readers, who are poised to benefit the most from regu-

lar formative ORF assessment. Lastly, the largest Whisper model evaluated in this study was Whisper medium. We did not examine Whisper large due to resource constraints, and it is likely to provide improved results.

Acknowledgments

References

- Paul D. Allison. 2009. *Fixed Effects Regression Models*. Sage, Los Angeles, CA.
- Saeed S. Alqahtani. 2020. [Technology-based interventions for children with reading difficulties: A literature review from 2010 to 2020](#). *Educational Technology Research and Development*, 68(6):3495–3525.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. [Common Voice: A massively-multilingual speech corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.
- Vivek Bhardwaj, Mohamed Tahar Ben Othman, Vinay Kukreja, Youcef Belkhier, Mohit Bajaj, B. Srikanth Goud, Ateeq Ur Rehman, Muhammad Shafiq, and Habib Hamam. 2022. [Automatic Speech Recognition \(ASR\) Systems for Children: A Systematic Literature Review](#). *Applied Sciences*, 12(9):4419.
- Daniel Bolaños, Ronald A Cole, Wayne Ward, Eric Borts, and Ed Svirsky. 2011. FLORA: Fluent oral reading assessment of children’s speech. In *ACM Transactions on Speech and Language Processing (TSLP)*, volume 7, pages 1–19. ACM New York, NY, USA.
- Wei Chen and Jack Mostow. 2011. A tale of two tasks: Detecting children’s off-task speech in a reading tutor. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, pages 2085–2088.
- Theodore J. Christ. 2018. *Formative Assessment System for Teachers™ Technical Manual*. Author and FastBridge Learning, Minneapolis, MN.

- Michael A. Covington and Joe D. McFall. 2010. [Cutting the Gordian knot: The moving-average type-token ratio \(MATTR\)](#). *Journal of Quantitative Linguistics*, 17(2):94–100.
- Stanley L. Deno, Phyllis Mirkin, and Ben Chiang. 1982. [Identifying valid measures of reading](#). *Exceptional Children*, 49(1):36–45.
- Griffin Dietz Smith, Dianna Yee, Jennifer King Chen, and Leah Findlater. 2025. [Prompting Whisper for Improved Verbatim Transcription and End-to-End Miscue Detection](#). In *Proceedings of Interspeech 2025*, pages 1943–1947, Rotterdam, The Netherlands.
- Lingyun Gao, Cristian Tejedor-Garcia, Catia Cucchiari, and Helmer Strik. 2025. [Fine-tuning Wav2vec 2.0 for child speech recognition and reading miscue diagnosis in low-resource conditions](#). *Computer Speech & Language*, 82:101627.
- Yaroslav Getman, Ragheb Al-Ghezi, Tamas Grosz, and Mikko Kurimo. 2023. [Multi-task Wav2vec2 serving as a pronunciation training system for children](#). In *Proceedings of the 9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 36–40.
- Jan Hasbrouck and Gerald Tindal. 2017. [An update to compiled ORF norms](#). Technical Report 1702, Behavioral Research and Teaching, University of Oregon, Eugene, OR.
- Jiarui Hou, James F. Lee, and Stephen Doherty. 2024. [What is next in mobile-assisted reading? Insights from a decade of eye tracking research into cognitive processes](#). *Educational Research Review*, 45:100643.
- Rishabh Jain, Andrei Barcovschi, Mariam Yiwere, Peter Corcoran, and Horia Cucu. 2023. [Adaptation of Whisper Models to Child Speech Recognition](#). In *Proceedings of Interspeech 2023*, pages 5242–5246, Dublin, Ireland.
- Rishabh Jain, Andrei Barcovschi, Mariam Yahayah Yiwere, Peter Corcoran, and Horia Cucu. 2024. [Exploring native and non-native English child speech recognition with Whisper](#). *IEEE Access*, 12:41601–41610.
- Scott Jarvis. 2013. [Defining and measuring lexical diversity](#). In Scott Jarvis and Michael Daller, editors, *Vocabulary Knowledge: Human Ratings and Automated Measures*, volume 47 of *Studies in Bilingualism*, pages 13–44. John Benjamins Publishing Company, Amsterdam.
- Jitsi. 2024. [JiWER: Evaluate speech-to-text systems with word error rate](#). Version 3.0, Apache License 2.0.
- Yusuf Kara, Akihito Kamata, E. E. Ozkeskin, Xin Qiao, and Joseph F. T. Nese. 2023. Predicting oral reading fluency scores by between-word silence times using natural language processing and random forest algorithm. *Psychological Test and Assessment Modeling*, 65(1):36–54.
- Young-Suk Grace Kim, Jamie M. Quinn, and Yaacov Petscher. 2021. [Reading prosody unpacked: A longitudinal investigation of its dimensionality and relation with word reading and listening comprehension for children in primary grades](#). *Journal of Educational Psychology*, 113(3):423–445.
- Holly B. Lane and Valentina Contesse. 2022. *UFLI Foundations: An Explicit and Systematic Phonics Program*. Ventris Learning.
- Walter L Leite, Stephanie Hammerschmidt-Snidarich, A Corinne Huggins-Manley, Hatice K Karakis, Holly Lane, Seyedahmad Rahimi, Qian Shen, Xinyi Tai, and Matthew Schmidt. 2025. Storiza: A platform to support children’s oral reading fluency development with generative ai. In *Proceedings of the 19th International Conference of the Learning Sciences-ICLS 2025*, pp. 1574–1578. International Society of the Learning Sciences.
- Wei Liu, Ying Qin, Zhiyuan Peng, and Tan Lee. 2024. [Sparsely shared LoRA on Whisper for child speech recognition](#). In *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11751–11755. IEEE.
- Kathrin E. Maki and Stephanie M. Hammerschmidt-Snidarich. 2022. [Reading fluency intervention dosage: A novel meta-analysis and research synthesis](#). *Journal of School Psychology*, 92:148–165.
- Bo Molenaar, Cristian Tejedor-Garcia, Catia Cucchiari, and Helmer Strik. 2023. [Automatic Assessment of Oral Reading Accuracy for Reading Diagnostics](#). In *Proceedings of Interspeech 2023*, pages 5232–5236, Dublin, Ireland.
- Andrew Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In *Proceedings of Interspeech*.
- Jack Mostow. 2016. [Project LISTEN’s reading tutor](#). In Scott A. Crossley and Danielle S. McNamara, editors, *Adaptive Educational Technologies for Literacy Instruction*, pages 263–267. Routledge, New York, NY.
- Joseph F. T. Nese and Akihito Kamata. 2021. [Evidence for automated scoring and shorter passages of CBM-R in early elementary school](#). *School Psychology*, 36(1):47–59.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

- Linkai Peng, Kaiqi Fu, Binghuai Lin, Dengfeng Ke, and Jinsong Zhan. 2021. [A study on fine-tuning Wav2vec 2.0 model for the task of mispronunciation detection and diagnosis](#). In *Proceedings of Interspeech 2021*, pages 4448–4452.
- Philip M. Podsakoff, Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff. 2003. [Common method biases in behavioral research: A critical review of the literature and recommended remedies](#). *Journal of Applied Psychology*, 88(5):879–903.
- Alexandros Potamianos and Shrikanth Narayanan. 2003. [Robust recognition of children’s speech](#). *IEEE Transactions on Speech and Audio Processing*, 11(6):603–616.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28448–28481. PMLR.
- Seyedahmad Rahimi and Valerie J. Shute. 2024. [Stealth assessment: A theoretically grounded and psychometrically sound method to assess, support, and investigate learning in technology-rich environments](#). *Educational Technology Research and Development*, 72(5):2417–2441.
- Madison Rose, Michael Bennie, Valeria Pagliai, Hatice Kubra Karakis, Qian Shen, Xinyi Tai, Walter L. Leite, and Zoey Liu. 2026. [A dataset for oral reading in young English readers](#). In *Proceedings of the 30th Conference on Computational Natural Language Learning*, pages 588–600, San Diego, California, USA. Association for Computational Linguistics.
- R. Shaw and D. Shaw. 2002. DIBELS oral reading fluency-based indicators of third grade reading skills for Colorado State Assessment Program (CSAP). Technical report, University of Oregon, Eugene, OR.
- Valerie J. Shute, Xi Lu, and Seyedahmad Rahimi. 2022. [Stealth assessment](#). In J. M. Spector, editor, *The Routledge Encyclopedia of Education*. Routledge, London, UK.
- W. A. Silva, L. C. Carchedi, J. Gomes Júnior, J. Victor de Souza, E. Barrere, and J. Francisco de Souza. 2021. [A Framework for Large-Scale Automatic Fluency Assessment](#). *International Journal of Distance Education Technologies*, 19(3):70–88.
- George Spache. 1953. [A new readability formula for primary-grade reading materials](#). *The Elementary School Journal*, 53:410–413.
- Benedikt Szmrecsányi. 2004. [On operationalizing syntactic complexity](#). In *Le poids des mots: Actes des 7es Journées internationales d’analyse statistique des données textuelles (JADT 2004)*, volume 2, pages 1032–1039, Louvain-la-Neuve. Presses universitaires de Louvain.
- University of Oregon. 2023. *DIBELS 8th Edition Administration and Scoring Guide*.
- Sheila W. Valencia, Antony T. Smith, Anne M. Reece, Min Li, Karen K. Wixson, and Heather Newman. 2010. [Oral reading fluency assessment: Issues of construct, criterion, and consequential validity](#). *Reading Research Quarterly*, 45(3):270–291.
- Alissa P. Wolters, Young-Suk Grace Kim, and John William Szura. 2022. [Is reading prosody related to reading comprehension? a meta-analysis](#). *Scientific Studies of Reading*, 26(1):1–20.
- Xiaoshuo Xu, Yueteng Kang, Songjun Cao, Binghuai Lin, and Long Ma. 2021. [Explore Wav2vec 2.0 for mispronunciation detection](#). In *Proceedings of Interspeech 2021*, pages 4428–4432.
- Zhihong Xu, Kausalai Wijekumar, Gilbert Ramirez, Xueyan Hu, and Robin Irey. 2019. [The effectiveness of intelligent tutoring systems on K-12 students’ reading comprehension: A meta-analysis](#). *British Journal of Educational Technology*, 50(6):3119–3137.

7 Tables and Figures

Table 1: Wav2Vec2 evaluation on sentence-level data (n = 3013).

Model	WER	CER
Wav2Vec2 Base	0.768	0.695
Wav2Vec2 Large	0.727	0.683

Table 2: Sentence-level ASR evaluation (n = 3013, human WCPM = 76.90)

Model	WER	CER	WCPM	Rel. Bias
Whisper tiny	0.450	0.283	63.57	−0.1733
Whisper base	0.379	0.234	69.88	−0.0913
Whisper small	0.294	0.183	77.75	+0.0111
Whisper medium	0.2613	0.165	80.61	+0.0482

Table 3: Story-level ASR evaluation (n = 320, human WCPM = 54.81)

Model	WER	CER	WCPM	Rel. Bias
Whisper tiny	0.4324	0.2735	49.45	−0.0978
Whisper base	0.358	0.228	54.38	−0.0078
Whisper small	0.294	0.194	58.23	+0.0624
Whisper medium	0.2768	0.1871	60.00	+0.0947

Table 4: Regression Coefficients (Standard Errors) of estimates from fixed-effects regression model

	Whisper Tiny	Whisper Base	Whisper Small	Whisper Medium
Manual WCPM	0.7314*** (0.0528)	0.8304*** (0.0559)	0.8425*** (0.0519)	0.8704*** (0.0625)
Readability	−2.615 (1.346)	−2.481 (1.444)	−1.970 (1.196)	−1.926 (1.180)
Syntactic complexity	0.6770 (0.7926)	0.7917 (0.8610)	0.7730 (0.8540)	0.7321 (0.8164)
Lexical diversity	43.89** (15.65)	24.60 (12.75)	23.81* (9.902)	15.29 (10.32)

Note. Standard errors in parentheses. $\cdot p < .10$; $* p < .05$; $** p < .01$; $*** p < .001$

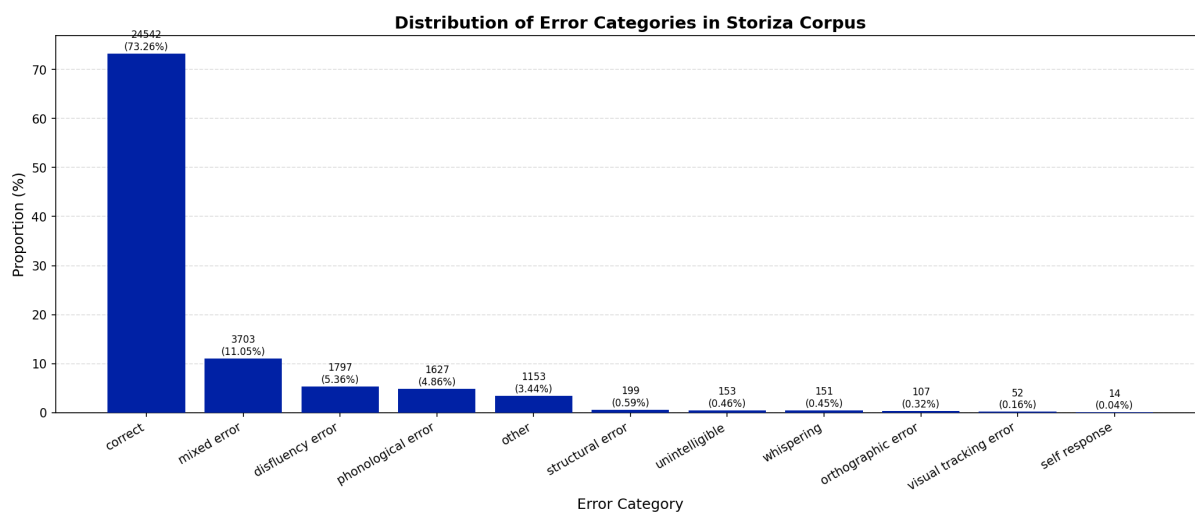


Figure 1: Distribution of error categories in the corpus.

Table 5: Test–Retest Reliability Estimates with 95% Confidence Intervals by Group and Measure

Group	Measure	r [95% CI]
Overall	Manual	0.693 [0.640, 0.740]
	Whisper Tiny	0.744 [0.699, 0.784]
	Whisper Base	0.730 [0.682, 0.771]
	Whisper Small	0.756 [0.712, 0.794]
	Whisper Medium	0.754 [0.709, 0.792]
≥ 60 sec	Manual	0.696 [0.638, 0.747]
	Whisper Tiny	0.762 [0.715, 0.803]
	Whisper Base	0.755 [0.705, 0.797]
	Whisper Small	0.772 [0.726, 0.812]
	Whisper Medium	0.782 [0.737, 0.820]
< 60 sec duration	Manual	0.332 [0.115, 0.519]
	Whisper Tiny	0.276 [0.054, 0.472]
	Whisper Base	0.303 [0.083, 0.494]
	Whisper Small	0.377 [0.165, 0.555]
	Whisper Medium	0.335 [0.119, 0.521]
Kindergarten	Manual	0.298 [−0.180, 0.663]
	Whisper Tiny	0.271 [−0.209, 0.646]
	Whisper Base	0.252 [−0.228, 0.634]
	Whisper Small	0.253 [−0.228, 0.634]
	Whisper Medium	0.276 [−0.203, 0.649]
1st Grade	Manual	0.829 [0.771, 0.873]
	Whisper Tiny	0.822 [0.762, 0.868]
	Whisper Base	0.828 [0.770, 0.873]
	Whisper Small	0.852 [0.801, 0.891]
	Whisper Medium	0.861 [0.813, 0.897]
2nd Grade	Manual	0.592 [0.493, 0.675]
	Whisper Tiny	0.660 [0.574, 0.732]
	Whisper Base	0.627 [0.535, 0.705]
	Whisper Small	0.661 [0.575, 0.733]
	Whisper Medium	0.652 [0.564, 0.725]
3rd Grade	Manual	0.718 [0.560, 0.825]
	Whisper Tiny	0.800 [0.680, 0.878]
	Whisper Base	0.835 [0.733, 0.900]
	Whisper Small	0.825 [0.717, 0.894]
	Whisper Medium	0.836 [0.735, 0.901]

Table 6: Mean WCPM by grade level at sentence level ($n = 3013$).

Method	Kinder. (146)	1st Grade (941)	2nd Grade (1391)	3rd Grade (535)	Mean (3013)
Manual	47.66	69.32	80.66	88.42	76.90
Whisper medium	49.70	72.60	84.04	94.21	80.61
Rel. Bias	+0.043	+0.047	+0.042	+0.065	+0.048

Table 7: Mean WCPM by grade level at story level ($n = 320$).

Method	Kinder. (20)	1st Grade (122)	2nd Grade (132)	3rd Grade (46)	Mean (320)
Manual	26.33	49.01	58.69	71.46	54.81
Whisper medium	27.82	53.50	64.13	79.39	60.00
Rel. Bias	+0.056	+0.091	+0.092	+0.111	+0.095