JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education
DOI: 10.1111/jedm.70040
Journal of Educational Measurement
Summer 2026, Vol. 63, No. 2, e70040

Word-Features Influencing Second Graders' Word Recognition in Connected Texts: Secondary Analysis of Oral Reading Fluency Data Using Explanatory Item-Response Models

Yukie Toyama
University of California, Berkeley
Elfrieda H. Hiebert
TextProject
Robin Irey
University of California, Berkeley and Stanford University
Cai Qing
University of California, Berkeley

This study presents a secondary analysis of word recognition patterns of 650 second-grade students on an untimed oral reading fluency assessment, using explanatory item-response models (EIRMs). The analytic sample included 50 passages containing 1,267 word-tokens. Rasch model calibration showed most words clustered around the lower third of student ability distribution, capturing variability among struggling readers. Subsequent EIRM analyses indicated that several factors had significant effects on word recognition difficulty: Lexical features including word frequency, age of acquisition, and first-syllable vowel patterns (especially r-controlled vowels, diphthongs, and long vowels) were key predictors. Content words (nouns and verbs) proved easier than function words, while more concrete words were unexpectedly more difficult. Word position within passages significantly predicted difficulty. When examined simultaneously, first-syllable vowel patterns emerged as the most robust decoding-related predictor, explaining variance previously attributed to broader decoding measures. The final model explained 40% of item variance. By identifying specific item features that challenge developing readers in oral reading, this research provides a foundation for more precise instructional approaches that target particular sources of word recognition difficulty rather than relying on conventional aggregate reading indicators.

Introduction

Oral reading fluency assessments have become deeply embedded in American educational practice. Implemented as curriculum-based measurement reading (CBM-R) such as DIBELS, these assessments have evolved considerably from their original role as diagnostic screeners. Researchers have increasingly reconceptualized CBM-R data as tools capable of informing instructional differentiation and guiding ongoing instructional decisions (King et al., 2022; Tindal, 2013). A line of inquiry termed "intervention validity" has examined how using CBM-R measures for instructional placement—particularly for tiered intervention decisions—can improve student outcomes (Hagans, 2008; Hosp et al., 2016; King et al., 2022; Tindal, 2013). Platforms

such as the widely used DIBELS now generate multiple reports designed to facilitate this expanded role, from class-level summaries to dedicated instructional grouping tools.[1]

Yet a critical limitation persists: Oral reading fluency (ORF) assessments contain far richer diagnostic information than current practice extracts. Fuchs et al. (2007) and more recently Toste et al. (2025) have argued persuasively that detailed error analysis could illuminate the specific word-recognition difficulties driving poor performance, potentially refining both our conceptualization of reading difficulties and our instructional response. Despite these calls, the diagnostic potential of these assessments remains largely untapped. Educators and researchers have relied almost exclusively on aggregate metrics—words read correctly per minute and overall accuracy percentages—to characterize student performance and guide instructional decisions. Consequently, there is limited understanding of which specific words confound struggling readers, what patterns characterize their errors, or how error profiles vary across the performance distribution.

This diagnostic void is most consequential for students failing to attain proficient reading standards. Consider the evidence at hand (University of Oregon, 2022): By the end of Grade 2, students at the 45th percentile and above achieve accuracy rates approaching 99% on oral reading fluency assessments, indicating that proficient word recognition is emerging across the upper portion of a cohort. This ceiling effect, however, obscures substantial variability in the lowest-performing third. Students at the 15th and 25th percentiles achieve 87% and 95.5% accuracy, respectively. These seemingly modest gaps carry profound consequences: At the 15th percentile, 13 errors per 100 words can substantially impede both comprehension and fluency, while students at the 25th percentile (4-5 errors per 100 words) face compounding difficulties when processing typical instructional texts of 400-500 words. Yet aggregate error counts obscure the fundamental question underlying instructional response: What specific words are students failing to recognize, and what do these failures reveal about underlying word-learning mechanisms?

This level of specificity is especially profound in Grade 2, where Connor et al. (2007) identify a moment for second chances and Stahl and Heubach (2005) mark a pivotal developmental transition. Students whose word recognition remains fragile at this juncture face increasing vulnerability to cumulative reading difficulty as comprehension demands intensify. Understanding not simply the aggregate count of errors, but rather which words drive those errors and what patterns characterize them, permits the kind of targeted instructional design that addresses the specific word-recognition vulnerabilities distinguishing the lowest performers from their more proficient peers.

Literature Review

The Study of Word-Level Reading Errors in Oral Reading

Oral reading errors provide a window into the processes students use to recognize and understand words. The way students read unfamiliar words—and the mistakes they make—can reveal both the strategies they attempt and the constraints imposed by the words themselves. This is particularly relevant for students in the bottom

third of achievement distributions, for whom word-level difficulty often represents a substantial barrier to comprehension and fluency.

Research on oral reading errors can be understood through three interrelated perspectives: (a) errors as evidence of linguistic strategies, (b) effects of assessment context (i.e., connected texts, isolated word lists) on performance, and (c) analyses of specific word-level features of errors. Taken together, these perspectives inform both the strategies students attempt and the constraints that limit successful word recognition.

Errors as Evidence of Linguistic Processing

Historically, miscue analysis emerged as a primarily qualitative approach to the study of oral reading errors, emphasizing detailed, interpretive analysis of individual miscues rather than reliance on aggregate error counts. Rather than viewing errors as simple failures, it proposed that miscues could be analyzed as approximations to correct responses, revealing the sources of linguistic information students attempted to use when encountering unfamiliar words (Goodman, 1969). Although influential historically, this perspective is no longer dominant in research on reading development.

Weber (1970) provides an illustrative example of the approach. She analyzed first graders' oral reading errors across multiple linguistic dimensions, showing that misreadings often reflected systematic attention to available cues. Errors were evaluated for graphic similarity to the target word, morphological structure, grammatical acceptability, and semantic plausibility. A student might substitute a morphologically related word (*running* for *runner*) or a semantically plausible alternative, reflecting reliance on context. Even struggling readers drew on graphic, grammatical, and semantic cues when attempting to identify words. From this perspective, errors were seen as revealing strategy rather than deficit, highlighting what students could do even if incompletely. Struggling readers were thought to compensate for incomplete word-level knowledge by leveraging semantic and contextual information.

Subsequent research, however, has challenged this view. Nicholson (1991) replicated aspects of Goodman's (1969) findings on the benefits of context but found variation by age and proficiency. While poor readers and younger average readers gained from context, skilled second- and third-grade readers sometimes performed better on isolated word lists than in passages. These results indicate that while miscue analysis highlighted potentially strategic behavior, it does not fully explain word-level reading development: decoding skill remains central, and context provides only limited compensation for struggling readers.

Assessment Context Shapes Reading Performance

The context in which words appear can influence recognition and fluency. Words in connected text are often read more accurately than in isolation because semantic, syntactic, and discourse cues provide additional support. However, the degree to which readers benefit from context depends heavily on proficiency: Students with weaker decoding skills derive less advantage from contextual information (Ardoin et al., 2013).

For beginning readers, limited word-level recognition constrains how much context can help. In a study of 143 first-grade and 147 second-grade students, Ardoin et al.(2013) found that students read words faster in connected CBM-R passages than on word lists. For second graders, passage performance explained unique variance in comprehension beyond isolated word reading, demonstrating that CBM-R captures more than simple decoding. Yet context effects were weaker for first graders and low-achieving students, indicating that automatic word recognition remains the primary driver of reading success.

Contextual effects are also sensitive to the specific letter-sound patterns present in words. Flynn et al. (2011) tested 69 second- through fourth-grade students on 599 nonsense words representing 89 distinct letter-sound patterns, both in isolation and within passages. Recognition of some patterns occurred only in connected text, particularly for struggling readers, showing that context can selectively support decoding but cannot fully compensate for underlying word-level difficulties. These studies highlight that oral reading errors reflect both word-level constraints and the ability to integrate contextual cues, with proficiency determining the relative contribution of each.

Word-Level Features as Predictors of Reading Errors

The properties of words themselves—length, frequency, orthographic regularity, and morphological complexity—create predictable constraints on reading, and their effects vary systematically by proficiency (Hiebert et al., 2020): Descriptive analyses of 411 first graders' performance on DIBELS oral reading fluency assessments illustrate these patterns. High-performing students primarily struggled with long, unfamiliar multisyllabic words. Middle-proficiency students encountered bottlenecks on multisyllabic words, which comprised 73% of their unknown words in winter assessment, even as they successfully read simpler words. Low-performing students were constrained by word frequency, length, and decodability, showing minimal recognition of multisyllabic or morphologically complex words. The study also highlighted a mismatch between instruction and the assessment's text demands: Although first-grade phonics curricula emphasize short and long vowels, only 36% of words in the ORF passages contained these patterns; 40% featured rare vowel combinations, and 24% were multisyllabic.

More recently, Bruner et al. (2025) have shown how orthographic and phonological features predict first graders' errors in CBM-R through generalized linear modeling (GLM): phonological complexity increased error likelihood by 21% per additional phoneme, whereas predictable spelling-to-sound correspondences reduced errors by 88%. Later-acquired words were misread three times more frequently than early-learned words, and irregular words such as *though*, despite only two phonemes, were highly error-prone due to low orthographic consistency (.29). These findings demonstrate that reading errors follow predictable patterns reflecting both developmental progression and word-level constraints.

Recognizing these patterns allows educators to anticipate challenges and target instruction toward orthographic, phonological, and morphological bottlenecks, rather than relying solely on general text-level difficulty ratings. For struggling readers,

aligning instruction with these predictable sources of difficulty can accelerate progression through typical developmental sequences while reducing persistent word-level barriers (Bruner et al., 2025; Hiebert et al., 2020).

Understanding Word-Features Influencing Reading Proficiency

Structural variables—orthography, length, familiarity, and morphology—and, to a lesser extent, semantic variables, such as dispersion and polysemy, are robust predictors of early reading acquisition (Compton et al., 2023). As children build lexical networks, polysemous and widely dispersed words may provide multiple, reinforced connections. While structural influences on word recognition are well documented, semantic features may offer additional explanatory power, particularly for struggling readers. This section examines structural, lexical, and contextual predictors of word recognition in connected text, with emphasis on students in the bottom third of the achievement distribution.

Word frequency and dispersion. Word frequency is the strongest single predictor of word recognition, accounting for roughly 30%-40% of variance in word recognition tasks (Brysbaert et al., 2018). High-frequency words are recognized more quickly, more accurately, and by more readers than low-frequency words (Monsell et al., 1989), with larger frequency effects for individuals with smaller vocabularies (Preston, 1935). For struggling readers, frequency is a major constraint on accurate word recognition. Beyond raw counts, dispersion and contextual diversity—how widely a word is distributed across subject areas or texts—also shape recognition efficiency (Adelman et al., 2006; Carroll et al., 1971), with words appearing in many contexts often processed more efficiently than equally frequent but niche words.

Orthographic structure. Orthographic structure—including length, syllabic structure, and grapheme–phoneme correspondences (GPCs)—strongly affects reading performance. Length effects mark the transition from letter-by-letter decoding to more holistic processing (Zoccolotti et al., 2005) and are amplified by the high prevalence of multisyllabic words in primary-grade texts: Approximately half of unique words in Grade 1 and two-thirds in Grade 3, which often remain bottlenecks for struggling readers (Kearns & Hiebert, 2022).

GPC properties further condition difficulty. Spencer (2010) identified 217 GPC combinations in the most frequent 1,000 words, with word and GPC frequency emerging as major predictors of 6-year-olds' recognition. Mastery tends to progress from simpler consonant–vowel patterns to more complex vowel patterns (Guthrie & Seifert, 1977), with nuances such as earlier mastery of short /i/, /e/, and /u/ than short /a/ (Pirani-McGurl, 2009). Several metrics quantify decoding demands, including type- and token-consistency indices (Chee et al., 2020), human transparency ratings capturing deviation from regularized pronunciation (Edwards et al., 2024; Steacy et al., 2023), and the Decoding Measure (DM; Saha et al., 2021), which combines frequency, orthographic transparency, GPC probability, and consonant blends to index sublexical demands.

Instructionally, teachers typically organize decoding around vowel families—short and long vowels, digraphs, diphthongs, and variant patterns (Fry, 2004; Moats,

2000). The Menon and Hiebert (2005) vowel pattern system, aligned with this practice, predicts word-reading performance across studies (Compton et al., 2004; Fitzgerald et al., 2015) but collapses all multisyllabic words into a single category. Given the prevalence and heterogeneity of multisyllabic words (Kearns & Hiebert, 2022), analyzing their vowel patterns is essential for characterizing decoding demands, particularly for students with persistent word-recognition difficulties.

Other lexical features. Age of acquisition (AoA) indexes when words are typically learned in oral language (Kuperman et al., 2012). Words acquired later (e.g., *yawl*, *brawl*) are recognized more slowly than earlier-acquired phonological neighbors (e.g., *draw*, *yawn*), even at comparable printed frequency (Steacy & Compton, 2019), underscoring the role of oral vocabulary—especially for struggling readers. Concreteness, the extent to which a word refers to perceptible entities, also facilitates recognition: concrete words elicit mental images more readily (Brysbaert et al., 2014; Steacy & Compton, 2019), and can even offset orthographic irregularity (Strain et al., 1995).

Contextual Features in Oral Reading

Context supports word recognition, with developing readers typically reading words more efficiently in passages than in isolation (Jenkins et al., 2003). Yet roughly half of the variance in text-based oral reading fluency remains unexplained after accounting for standard text complexity indices (Francis et al., 2008). Word position within a passage may contribute to this residual variance. First encounters with a difficult word slow reading more than later occurrences, and experimental timing data show that position influences rate and accuracy (Liceralde et al., 2022). Words appearing later in a passage may be harder due to fatigue or accumulated cognitive load. From a psychometric perspective, explanatory item-response models have been used to treat item position as an explicit predictor, documenting practice and fatigue effects (e.g., Debeer & Janssen, 2013; Hohensinn et al., 2008). Such models illustrate how position effects can bias item parameter estimates if ignored, and they provide a framework for quantifying how these effects differ across proficiency levels.

Explanatory Item-Response Models

To identify the specific drivers of word recognition difficulty, we employ Explanatory Item-Response Models (EIRMs; De Boeck & Wilson, 2004). Unlike traditional analytic approaches that aggregate performance into a single accuracy score, EIRMs simultaneously account for the shared variance among students, words, and passages. This framework allows for a decomposition of item difficulty, isolating the unique impact of linguistic features (e.g., vowel patterns) while statistically controlling for both the reader's latent ability and the contextual effects of the passage.

A critical advantage of EIRMs for oral reading research is the treatment of words and passages as random effects. While single-level regression models (e.g., standard GLMs) often ignore the nesting of words within texts, EIRMs recognize that words in a shared passage are not independent. By modeling this nested structure, we prevent the underestimation of standard errors and ensure that our results generalize

to a broader universe of possible texts rather than being limited to the specific passages administered (Briggs, 2008; Hartig & Buchholz, 2012; Mislevy, 1987). Although EIRMs have been successfully applied to discrete literacy skills such as lexical representation (Cho, Gilbert, et al., 2013; Cho, Goodwin, et al., 2024) and reading comprehension (Kulesze, et al., 2016; Toyama, 2021), their application to connected-text reading represents a novel extension of the framework to identify specific word-recognition difficulty in the ORF assessment context.

Summary

Oral reading errors reveal both the strategies students attempt and the constraints imposed by word characteristics and developmental stage. Struggling readers are especially limited by decoding skill and the specific features of words that exceed their current capacity. Word-level factors including frequency, orthographic complexity, age of acquisition, and decoding demand create predictable patterns of difficulty. Recognizing these patterns allows for more precise assessment and targeted instruction, aligning interventions with both developmental trajectory and the word-level challenges that shape reading accuracy and fluency.

The present study addresses this gap by using EIRMs to systematically analyze error patterns in the oral reading performance of Grade 2 students, primarily in the lowest-performing third of the distribution. By moving beyond aggregate metrics to characterize the specific errors students make and what these patterns illuminate about word-recognition vulnerabilities, this analysis aims to equip educators with the granular, theoretically-informed understanding necessary to enhance both the diagnostic validity and instructional utility of oral reading fluency measures now central to educational resource allocation. From a measurement perspective, this study demonstrate how EIRMs can decompose word level difficulty in connected-text reading, extending their application beyond traditional item formats.

The Present Study

This study aims to describe student learning trajectories through reading acquisition, focusing on second grade due to its critical role in proficient reading. Convergent evidence from multiple research teams shows second grade as both a developmental transition point in reading processes and a statistical inflection point in reading trajectories, making it essential for understanding literacy development during this critical period (Connor et al., 2007; Spira et al., 2005).

We used EIRMs to analyze word-level reading responses taken from untimed oral reading data for second-grade students. By examining word recognition patterns in connected text, we aim to move beyond aggregate measures to identify specific word-features that challenge different proficiency levels, providing insights for targeted interventions. Specifically, we address the following questions:

RQ1. What is the difficulty of the words in untimed ORF passages for second graders, relative to their ability?

RQ2. Do the contextual factors, i.e., word's position in the passage, as well as decoding factors (i.e., decoding demand, spelling-pronunciation transparency,

and first syllable vowel patterns) affect word difficulty in connected texts after controlling for other word-features with known impact on word reading?

Methods

Student Sample

We analyzed word-reading performance in passages read by 650 students across two cohorts (2017-2018: 40%, 2018-2019: 60%) from four districts and seven schools in the US pacific northwest. The sample was distributed across schools ranging from 4% to 23% of total enrollment per school. The sample was gender-balanced (49% male, 51% female) and predominantly White (65%) and Hispanic (25%). Three-quarters of students (75%) received free/reduced lunch. Students with disabilities comprised 11% of the sample, and English learners (ELs) made up 11% (though EL data was missing for one state in 2017-2018). When compared to state enrollment data, the sample's gender distribution (49% M, 51% F) matched typical K-12 enrollment. English learners (11%), students with disabilities (11%), and the sample's racial/ethnic composition reflected the Pacific Northwest region's averages. The sample's free/reduced-lunch eligibility (75%) notably exceeded state and national averages (approximately 50%-55%), reflecting geographic concentration of participating districts in high-poverty communities.

To evaluate the representativeness of reading abilities in our sample, we compared participants' ORF performance to Hasbrouck and Tindal (2017) Winter norms. Our sample's oral reading fluency matched the national norm closely, with median WCPM ranging from 11 at the 10th percentile to 129 at the 90th percentile—a distribution nearly identical to the national benchmark values across percentile levels. This alignment across 650 participants, spanning from struggling readers to high performers, suggests that our sample's ORF performance mirrors second-grade national norms, though the sample was not designed to be nationally representative.

CORE Assessment Administration Procedures

The current study represents a secondary analysis of existing data collected as part of a large-scale Computerized Oral Reading Evaluation (CORE) project (Nese & Kamata, 2021), which developed an automated system for estimating aggregate ORF scores using automatic speech recognition (ASR), psychometric modeling, and shorter passages for grades 2-4. Our secondary analysis examines specific word-level features influencing reading difficulty—research questions not addressed in the original CORE project. Consistent with established CBM-R/ORF protocols, the CORE passages were authored by a researcher with expertise in passage construction and ELA teaching (Nese & Kamata, 2021). They follow authentic narrative conventions—characters, settings, and coherent plot structures—alongside realistic, developmentally appropriate content reflecting children's everyday experiences (sports, pets, hobbies, school tasks). Three standardized readability indices confirmed grade-level appropriateness for the passages: Lexile 410-600, Flesch Reading Ease approximately 92, and Flesch grade level 2.98.[2] Empirical evidence (Nese, 2022) showed that the newly written CORE passages produce reading fluency

scores comparable to traditional ORF passages (i.e., easyCBM; Alonzo et al., 2006). The data made available for our study included medium and long CORE passages ($M = 59$ words, $SD = 15.6$).

The CORE assessment used ASR to collect and score each word reading, which has shown high reliability with human scores (.81-.94; Nese & Kamata, 2021). Students completed the assessment individually on laptops, using headphones equipped with noise-cancelling microphones. A distinctive feature of the CORE data collection was that, unlike typical timed ORF assessments, students read aloud each passage without a time limit. This feature is advantageous for word–level accuracy analysis because the number of students attempting each word remains relatively stable across the entire passage, rather than dropping sharply further into the text due to a time-limit cutoff (e.g., 1 minute). Students were randomly assigned a fixed set of 10-12 passages and progressed through them at their own pace.[3]

Text and Word Data

For our investigation, we analyzed second-grade data collected during the winter administration periods of two academic years (2017-2018 and 2018-2019). A total of 72 passages were administered to the second graders during these winter assessments. The majority of students ($\sim$85%) read 10 passages each. For each student-passage pair, we calculated word read correctly per minute (WCPM), following the method described in Kara et al.'s (2023). We focused on each student's five best passages based on WCPM to reduce the effects of fatigue and ASR's potential recording errors. We also dropped passages which had been read by fewer than 20 students. These exclusion criteria resulted in 50 passages in analytic sample. As for word-tokens (i.e., "items"), we excluded the following from our analysis: (a) very high-frequency words (e.g., *the, and*) and 2-letter words (e.g., *to, of, an)*; (b) proper nouns, contractions, possessives, and interjections; (c) items that were missing word-feature variables, (d) items that everyone read correctly; and (d) items that did not fit to the Rasch model in the initial calibrations.[4] Each word-token (or "item") was dichotomously scored by ASR, with 1 = correct, and 0 = incorrect. After applying the selection criteria, the final analytic sample consisted of 1,267 "items" (606 unique words) from 50 passages.[5]

Item Feature Variables

We included multiple item-feature predictors, the majority of which capture different dimensions of word complexity that may influence reading difficulty in connected text. These item feature variables are shown in Table 1.

Word frequency and dispersion. U-Function, derived from the Educator's Word Frequency Guide (Zeno et al., 1995), represents the appearances of a word per million words across more than 17 million words in educational texts. To reduce skewness, we applied a log transformation (logUfunction). We also included dispersion, which measures how evenly a word is distributed across different content areas and text types (Zeno et al., 1995). Higher dispersion values indicate that a word appears

Table 1
Descriptive Statistics for the Item Feature Variables

Variable	Min	q1	Median	q3	Max	Mean	*SD*
length	3	4	5	6	11	5.01	1.51
logUfunction	0	4.49	5.62	6.74	8.94	5.58	1.53
dispersion	.25	.78	.89	.96	1	.85	.14
AoA	2.22	3.83	4.67	5.41	9.58	4.78	1.23
concreteness	1.12	2.18	3.07	4.29	5	3.2	1.14
position	1.00	16.00	29.00	44.00	103.00	31.10	19.49
DM	.28	1.52	2.14	2.96	6.41	2.34	1.1
SPrating	1.07	1.52	1.79	2.19	4.22	1.89	.5

length = number of letters in a word; logUfunction = logarithmic transformation of the predicted frequency of word appearance per million words in educational texts (Zeno et al., 1995); AoA = age of acquisition rating (Kuperman et al., 2012); concretene = concreteness/abstractness rating of a given word (Brysbaert et al., 2014); position = position of a given word in a passage; DM = Decoding Measure (Saha et al., 2021); SPrating = spelling-to-pronunciation transparency rating (Edwards et al., 2024).

more consistently across various contexts, potentially enhancing recognition through diverse exposure.

Length. Word length was measured by number of letters. According to the number of syllables variable from the South Carolina Psycholinguistic metabase (SCOPE; Gao et al., 2022), our analytic word sample comprised 60.2% monosyllabic words, 34.5% disyllabic words, and 5.3% trisyllabic words.

Orthographic structure. Three oif the four components of the Decoding Measure (DM) were used: letter-sound discrepancy, number of blends, and conditional probabilities of grapheme-phoneme correspondences (Saha et al., 2021). We used the DM scores that excluded word frequency as the log-Ufunction was used as an independent frequency. The DM scores ranged from .28 to 6.41 ($M = 2.34$, $SD = 1.1$) in our sample. Spelling-to-pronunciation transparency ratings (SPrating) were obtained from Edwards et al. (2024), which ranged from 1.07 to 4.22 ($M = 1.89$, $SD = .5$) on a scale of 1-6, where 1 = "very easy to match" and 6 = "very difficult to match" to represent complexity of spelling-pronunciation. Words were also analyzed according to vowel patterns in their first syllables using a modified version of the Menon and Hiebert (2005) classification system (see Table 2).[6]

Word familiarity. Age of Acquisition (AoA), which indicates the estimated age at which a word is present in children's oral language (Kuperman et al., 2012), was used to measure word familiarity. AoA values in our sample ranged from 2.22 to 9.58 years with a mean of 5.01 years ($SD = 1.33$), where lower values indicate earlier-learned words.

Semantic properties. Concreteness ratings (concrete) reflect the degree to which a word refers to perceptible entities (Brysbaert et al., 2014), rated on a scale from 1 (very abstract) to 5 (very concrete). Our sample had a mean of 3.2 ($SD = 1.14$).

Table 2
Vowel Pattern in the First Syllable of Unique Words

Vowel Pattern	Phoneme/Type (*Examples*)	*n*	Percent
short	æ (*bag, fancy*); ɛ (*best, never*); ɪ (*sister, inches*); ɑ (*box, spotted*); ʌ (*puppy, grumpy*)	248	40.9%
long	ej (*bacon, raining*); i (*beat, realized*); aj (*drive, driving*); ow (*grow, opened*); u *(huge, use)*	175	28.9%
r-controlled	ʊɹ (*during, sure*); ɔɹ (*board, porch*); ɪɹ (*gear, here*) ɚ (*church, perfectly*); ɛɹ (*air, chair*); ɑɹ (*garage, sparkle*)	84	13.9%
diphthong	aw- (*down, town*); o- (*flour*); ɔj- (*coin, soil*); ʊ- (*good*)	59	9.7%
variant	prolific-short (*chance, dance*); prolific-long (*roller, most*); final-e (*cheese, leave*); limited (*head, eight*); pair/single (*some, great*)	40	6.6%
Total		606	100.0%

Grammatical classification. Of the 606 unique words in our sample, the categories of parts of speech (POS) were as follows: nouns (38.3%), verbs (34.8%), adjectives (11.8%), adverbs (9.0%), and function words (6.1%).

Position of a word in text. Position indexes each word's location within passages, treated as a continuous variable ranging from 1 (for the first word in the passage) to 103 (mean = 30.31, $SD = 19.79$).

Analytic Approach

We first fit the Rasch model using the TAM package in R (Robitzsch et al., 2025) to estimate item and person parameters and descriptively analyze word-token difficulty. To investigate sources of item difficulty, we employed explanatory item-response models (EIRMs)—also known as cross-classified generalized linear mixed-effects models—using the lme4 package (Bates et al., 2015). These models estimated log-odds of correct word reading as a function of lexical and contextual predictors while accounting for variance attributable to students, items, and passages. Including a random intercept for passage was deemed theoretically and statistically important, as it accounted for shared variance among words within the same passage and addressed violations of the Rasch model's assumption of local independence among items. To be sure, we evaluated the necessity of this random intercept empirically.

EIRMs extend traditional IRT by incorporating item- and/or person-level predictors directly into the model structure (De Boeck & Wilson, 2004). This framework facilitates the examination of how specific item features systematically influence item difficulty, thereby shedding light on the cognitive and linguistic mechanisms underlying performance—particularly relevant in oral reading contexts where characteris-

tics such as word length or decoding demands play a significant role. Specifically, the EIRMs used in this study can be expressed as:

$$logit\ (Pr(\ Y_{rik} = 1)) = \theta_r - \sum_{f=0}^{F} \beta_f X_{ik}^{(f)} + \delta_i + \gamma_k,$$

where:

- θ_r is the latent reading ability of respondent r, with $\theta_r \sim N(0,\ \sigma_\theta^2)$,
- $X_{ik}^{(f)}$ is the value of the fth feature for item i in passage k, where $X_{ik}^{(0)} = 1$ for all items, serving as a constant to represent the intercept β_0,
- β_f is the fixed effect coefficient for feature f,
- $\delta_i \sim N(0,\ \sigma_\delta^2)$ is the random deviation for item i, and
- $\gamma_k \sim N(0,\ \sigma_\gamma^2)$ is the random deviation for passage k.

All continuous predictors (see Table 1) were z-score transformed so coefficients represent effects of a 1 *SD* increase on log-odds probability when other variables are at their means. Because the model parameterizes item easiness (log-odds of correct response) in the lme4 package, negative coefficients indicate greater item difficulty. The intercept (β_0) represents log-odds probability for the average student ($\theta_r = 0$) reading a word with mean feature values ($X^{(f)} = 0$). R^2 was calculated as the proportion of item variance explained by item-features compared to the null model. The response matrix provided sufficient linkages through overlapping passages allocated across test forms, enabling concurrent calibrations.

For model building, we treated length, logUfunction, dispersion, AoA, concreteness, and POS as control variables, entering them simultaneously (Model 1). These features are known to impact word reading accuracy. We then added word position in passage as a contextual predictor (Model 2), followed by the examination of three decoding-related predictors: (a) decoding demand, (b) spelling-to-pronunciation rating, and (c) vowel patterns in the first syllable, added individually to Model 2 (Models 3a, 3b, 3c). Finally, we included all three decoding-related predictors together (Model 4) to determine whether each maintained unique effects on word reading accuracy while controlling for other predictors.

Before running the EIRMs, multicollinearity was examined using the variance inflation factor (VIF) statistics. No predictors reported here showed VIF greater than 5 (a VIF larger than 5 is considered moderate influence (Hair et al., 2006). Model fit was evaluated using global information criteria, specifically the Akaike Information Criterion (AIC, Akaike, 1974) and the Bayesian Information Criterion (BIC, Schwarz, 1978). Both indices balance model fit and parsimony by penalizing model complexity; lower values indicate a better trade–off between goodness of fit and the number of estimated parameters.

Results

RQ1. What Is the Difficulty of the Words in Untimed ORF Passages for Second Graders, Relative to Their Ability?

The WrightMap, Figure 1, provides an intuitive visualization of the measurement structure (Wilson, 2005, 2023), displaying item difficulty (top panel) and student

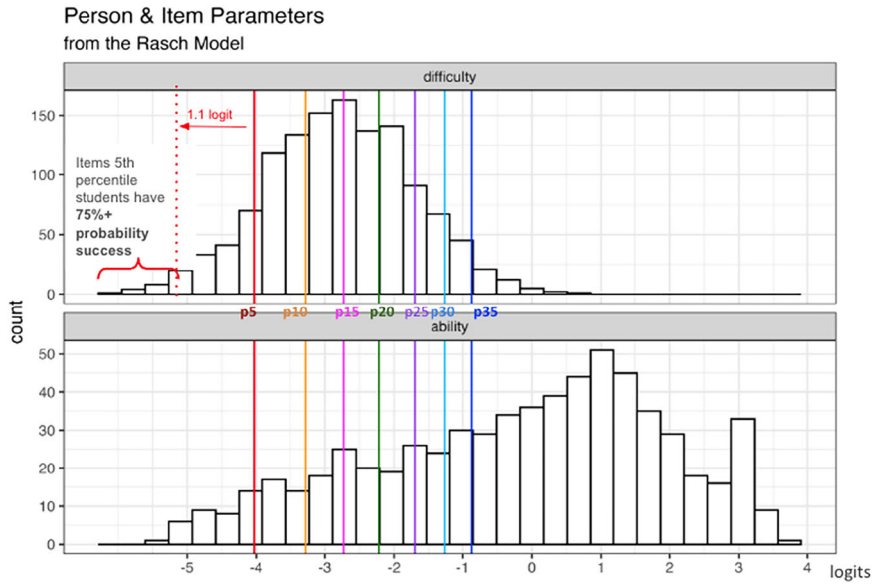

Figure 1. Wright map showing the item and student locations.

ability (bottom panel) on the same logit scale from the Rasch model. Because word responses are clustered within passages, the WrightMap is interpreted descriptively as a visualization of targeting. Student ability ranged from –5.52 to 3.58 logits ($M = -.21$, $SD = 2.1$), while item difficulty ranged from –6.28 to .68 logits ($M = -2.79$, $SD = 1.07$). Solid vertical lines indicate seven percentile locations for student ability from 5th through 35th percentiles (left to right).

Most items cluster to the left of the right-most solid line, indicating that the bulk of words are targeted to students at or below approximately the 35th percentile of the ability distribution. Because the Rasch scale is interval, distances on the logit metric have a direct probabilistic interpretation. For example, students at the 5th percentile (left-most solid line) are located 1.1 logits higher than the items left of the dotted line. This difference in logits indicates that the 5th percentile students have 75% or greater probability of correct response for those very easy items, since $logit^{-1}(1.1) \approx .75$. Examples of such very easy items include: "end," "family," "good," and "room."

Local item dependence (LID) was evaluated using residual correlations (Yen's Q3 and adjusted aQ3) (Yen, 1984). Because word-tokens are embedded within passages and many item pairs are not jointly observed by the same students due to the incomplete block design of the assessment, we examined Q3 and aQ3 both overall and after restricting to item pairs with adequate co-administration. In the full set of item pairs, Q3 exhibited an extreme upper tail, consistent with instability when overlap is sparse. After restricting to item pairs with at least 50 shared respondents, residual dependence was substantially reduced and was moderate in magnitude (e.g., Q3 95th percentile was .37-.38; approximately 8% of pairs exceeded Q3 = .30; for details, please see Table S1 in supplementary materials).[7] As expected, within-passage pairs

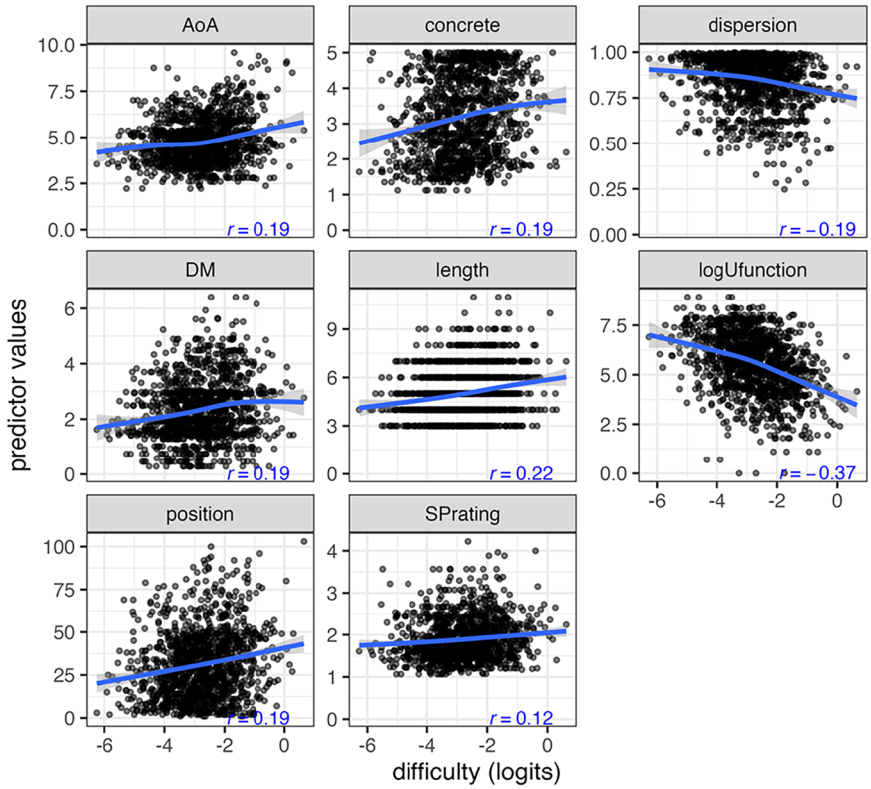

Figure 2. Scatter plots of item difficulty by continuous item-feature predictors.
Note. Pearson's correlation coefficients (*r*) are shown alongside LOESS (local polynomial regression) curves, which were included to inspect the functional relationship between each predictor and item difficulty.

showed slightly higher residual dependence than across-passage pairs. Accordingly, Rasch-based item and person locations in the WrightMap are presented as a descriptive summary of assessment targeting (RQ1). In contrast, inferential analyses (RQ2) relied on cross-classified explanatory item-response models with random passage effects to account for passage-level clustering in responses.

RQ2. Do Contextual Factors (Word Position in Passage) and Decoding Factors (Decoding Demand, Spelling-Pronunciation Transparency, and First Syllable Vowel Patterns) Affect Word Difficulty in Connected Texts after Controlling for Established Word-Features?

To understand item difficulty, we first examined bivariate relationships between the Rasch item estimates and each word-feature predictor. Figure 2 shows scatterplots for the continuous predictors, with Pearson's correlations and LOESS (local polynomial regression) curves summarizing functional form. Relationships were generally linear, with correlations ranging from –.37 (log of Ufunction) to .21 (dis-

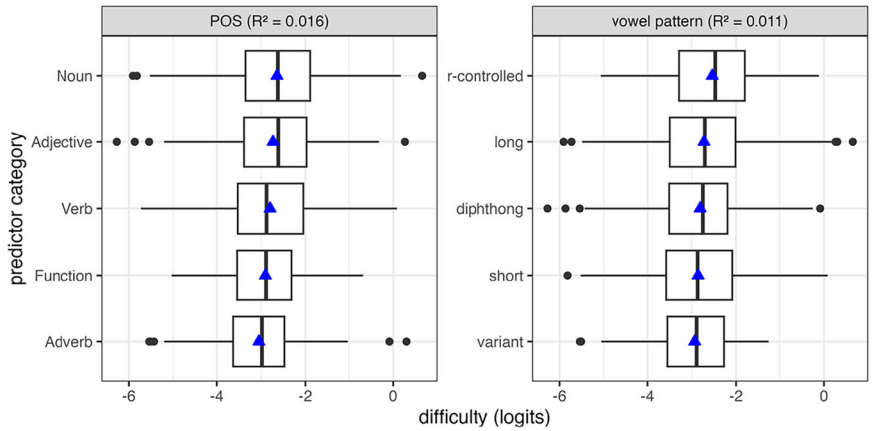

Figure 3. Boxplot of item difficulty by categorical item-feature predictors.
Note. POS = parts of speech. R^2 was derived from a one-way linear model regressing item difficulty estimates from the Rasch model to each of the categorical predictors. Triangles showing category group means.

persion and word length), and most directions were consistent with expectations. An exception was concreteness, for which more concrete words were associated with greater difficulty, contrary to theoretical predictions.

Figure 3 displays the distribution of item difficulties by (a) parts of speech (POS) and (b) vowel pattern categories, with triangles marking category means. Differences among categories were modest: One-way models regressing the difficulty estimates to each predictors indicated that these categorical variables accounted for only a small proportion of variance in difficulty (R^2 = .016 and .011 for POS and vowel patterns respectively). Overall, these bivariate associations—though largely in the expected directions—show that individual features, whether continuous or categorical, explain only limited variance in difficulty, motivating multivariable explanatory item-response models that consider their combined effects.

Explanatory Item-Response Modeling

To evaluate the need of a random effect for passages, we compared a random-item, random-person (RIRP) Rasch model to an extended model that also included a random passage effect (RIRP+RP). A likelihood ratio test favored the RIRP+RP model, $\chi^2(1) = 16.18$, $p < .001$, which also showed improved fit based on AIC (ΔAIC = –14) and BIC (ΔBIC = –5). These results supported the inclusion of the passage-level random effect in the base (null) model. Intraclass correlation coefficients (ICCs) from the RIRP+RP model indicated that 54% of variance in the log-odds of a correct response is attributable to differences between individuals, 7.7% to differences between items, and 5% to differences between passages. Notably, the inclusion of the passage-level random effect partially accounts for shared variance among items within passages, thereby mitigating local item dependence (LID). While the item

ICC was modest, it indicates sufficient variance in item difficulty that warrant explanatory modeling.

We then fit a series of EIRMs predicting word reading accuracy, beginning with the null model (RIRP+RP, Model 0). Six models of interest were run, each incrementally adding fixed effects. Model comparisons were evaluated using AIC, BIC, likelihood ratio (LR) tests, and changes in R^2 in terms of the reduction in the item variance compared against the null model.

Table 3 shows the model fit statistics and the null model (Model 0) served as the baseline for comparison. Model 1 included six control variables as fixed effects, namely *length*, *logU function, dispersion*, *AoA*, *concrete* and *POS*, yielding a significant improvement in fit, $\Delta\chi^2(9) = 301.92$, $p < .001$, with reduction in AIC and BIC relative to the null model. Model 2 further included *position*, the sole contextual variable used in the study, resulting in additional improvement (ΔAIC $= -46$; ΔBIC $= -47$; $\Delta R^2 = .05$). Subsequent models (Models 3a, 3b, and 3c) continued to show a better fit against Model 2, with Model 4—the full model including all the decoding related predictors—*decoding demand, spelling-pronunciation transparency*, and *vowel patterns in the first syllable*—along with *position* and the control variables, showing the highest explained variance ($R^2 = .40$), with reduction in AIC and the likelihood ratio showing a significant improvement in fit $\Delta\chi^2(2) = 7.87$, $p < .001$, although it showed slightly higher BIC, reflecting the trade-off between improved fit and increased model complexity.

Turning to the parameter estimates, Table 4 shows the results from the EIRMs. All but three predictors in Model 1 had significant and unique effects on the probability of correct word recognition in connected passage. The exceptions were *dispersion*, *adjective* and *adverb*. Among the ones found to be significant predictors, *concreteness* was the only one that had unexpected directionality of effect ($\beta = -.17$), indicating that 1 *SD* increase would make the logit probability to decrease by .17. Interestingly, all POS predictors had positive coefficients, of which ones for nouns and verbs were statistically significant—indicating that students have less success with function words, which was set as the reference category. Specifically, the coefficient for noun ($\beta = .43$) indicates that, holding all other predictors constant, the predicted probability of success increases by the average reader by 10.5 percentage points when the word is a noun rather than a function word.[8] This could be the characteristic of connected text reading—students may focus less on function words than content words that carry meaning, such as nouns and verbs.

Model 2 shows the negative effect of *word's position* ($\beta = -.19$), indicating that, for each one standard deviation ($\sim$20 words) increase in a word's position within the passage, the log-odds of a correct response decrease by .19. This translates into an approximate 5 percentage point reduction in the likelihood of a correct response by the average reader. Compared to Model 1, Model 2's R^2 increased by .05, indicating the word's position explained an additional 5 percentage points of the item variance, and this change was statistically significant by the likelihood ratio test.

Models 3a, 3b, and 3c examined the effect of decoding demand, spelling-to-pronunciation transparency rating, and vowel patterns of the first syllable, respectively. As can be seen in the table, decoding demand (Model 3a) and spelling-pronunciation transparency (Model 3b)—were both negatively associated with the

Table 3
Model Fit Statistics for the EIRMs

Model	Description	# para	BIC	AIC	log-Likelihood	R^2	ΔR^2 LR-Test	
0	RIRP+RP (null)	4	36826.74	36790.48	−18391.24	.000		
1	m0 + control variables	13	36624.40	36506.56	−18240.28	.319	.319	vs. Model 0***
2	m1 + position	14	36577.37	36450.46	−18211.23	.370	.051	vs. Model 1***
3a	m2 + DM	15	36580.98	36445.01	−18207.51	.376	.006	vs. Model 2*
3b	m2 + sp-rating	15	36575.17	36439.20	−18204.60	.382	.012	vs. Model 2***
3c	m2 + vowel patterns (VP)	18	36592.91	36429.74	−18196.87	.394	.024	vs. Model 2***
4	m3 + VP+ DM + sp-rating	20	36607.00	36425.90	−18192.90	.402	.008	vs. Model 3c*

RIRP+RP = random-item, random-person plus random-passage. # para = number of parameters control variables = length, dispersion, age of acquisition, concreteness, & POS (parts of speech tags, 5 categories with function word set as reference); DM = decoding demand, SPrating = spelling-to-pronunciation rating.
Note. vowel patterns (VP) is a 5-category variable with short-vowel set as reference. *$p < .05$. **$p < .01$. *** $p < .001$.

Table 4
Parameter Estimates from the EIRMs

Fixed Effects	Model 1 *Est.*	Model 2 *Est.*	Model 3a *Est.*	Model 3b *Est.*	Model 3c *Est.*	Model 4 *Est.*
(Intercept)	2.41***	2.44***	2.46***	2.47***	2.51***	2.52***
length	–.13***	–.13***	–.03	–.10***	–.15***	–.08
logUfunction	.31***	.3***	.31***	.31***	.31***	.31***
dispersion	.01	.02	.02	.02	.03	.03
AoA	–.15***	–.15***	–.15***	–.14***	–.15***	–.14***
concreteness	–.17***	–.16***	–.16***	–.17***	–.14***	–.15***
parts of speech (ref = function)						
noun	.43***	.39**	.36**	.36**	.45***	.40***
adjective	.10	.05	.04	.03	.12	.09
adverb	.13	.13	.13	.07	.23*	.18
verb	.20*	.15	.1	.13	.24*	.20*
position		–.19***	–.19***	–.19***	–.19***	–.19***
DM			–.12**			–.04
SPrating				–.1***		–.07
vowel pattern (ref = short)						
long					–.21**	–.17**
diphthong					–.30***	–.26**
variant					–.09	–.02
r-controlled					–.3***	–.28***
Random Effects						
var (item)	.456	.422	.418	.414	.406	.401
var (student)	4.690	4.700	4.700	4.700	4.700	4.700
var (passage)	.044	.041	.042	.040	.037	.038
R^2	.319	.370	.376	.382	.394	.402

var = variance.
Note. For other variables, see note for Tables 1 and 3. *$p < .05$. **$p < .01$. ***$p < .001$.

probability of a correct response: 1 *SD* increase in decoding demand was associated with a .12-logit decrease in the likelihood of a correct response, and 1 *SD* increase in spelling-pronunciation opacity was associated with a .10-logit decrease in log-odds of accuracy. When translated into probabilities from a baseline of 50% accuracy, these effects correspond to approximate decreases of 3.0 and 2.5 percentage points in predicted probability, respectively. These results indicate that words requiring more decoding effort and those with less transparent spelling-to-sound correspondences were slightly but consistently more difficult, independent of other variables in the model. Notably, once the decoding demand measure is included in Model 3a, the effect of word length is attenuated and no longer significant, suggesting that decoding demand subsumes much of the variance previously attributed to word length. In con-

trast, logUfunction, AOA, concreteness, noun, and position remain to show unique and statistically significant effects in all the models examined.

Model 3c results show that words with *r-controlled*, *diphthong*, and *long* vowels in the first syllable are more difficult than those with *short* vowels ($\beta = -.33, -.30, -.21$, respectively). Assuming baseline 50% probability for short-vowel words, these coefficients correspond to accuracy reductions of 8.2, 7.4, and 5.2 percentage points, respectively.

Notably, including vowel patterns altered effects for adverbs and verbs. In models 2, 3a, and 3b, these syntactic categories showed smaller, non-significant effects. However, in Model 3c, both showed unique, statistically significant positive associations with reading accuracy after controlling for other predictors. This suggests vowel patterns and syntactic categories share overlapping variance; accounting for the former reveals independent facilitative effects of adverbs and verbs. Regarding explanatory power, R^2 increases from Model 2 were .6 percentage points for DM (Model 3a), .12 percentage points for letter-sound transparency (Model 3b), and 1.2 percentage points for first syllable vowel patterns (Model 3c).

In the final model (Model 4), DM and spelling-to-pronunciation transparency became non-significant, while vowel patterns—*long-vowel*, *diphthong*, and *r-controlled*—remained statistically significant. This suggests vowel patterns captured variance previously explained by other two decoding-related predictors. Similarly, effects of word length and adverb diminished and became non-significant compared to Model 3c, indicating overlapping variance among predictors. The final model explained about 40% of item variance.

Lastly, to assess whether the effects of vowel patterns varied across passages, we extended the random-intercept model by allowing the vowel pattern variable to vary as random slopes across passages. However, the model comparison revealed no significant improvement in fit with this random slope model; $\chi^2(14) = 19.52$, $p = .15$, and both AIC and BIC increased relative to Model 4. These suggested no evidence for heterogeneity in vowel-pattern effects across passages.

Discussion

This study examined how specific lexical and contextual features contribute to word recognition difficulty when second-grade students read connected texts, moving beyond aggregate fluency measures to identify precise sources of word reading challenges. The WrightMap from the Rasch model revealed that most words clustered around the lower third of the student ability distribution. This suggests that the item set is well-targeted to students performing at the lower end of the reading continuum—those often designated as "below basic" or "at risk" on standardized assessments. By grounding interpretation in item-level performance, the Rasch framework affords fine-grained insights into which specific words pose greater or lesser difficulty for these learners. In fact, even students at the 5th percentile on the ability distribution on the WrightMap could correctly read the words at the very tail end of the item distribution with 75% or greater probability and it turns out "end," "good," and "room", are such words. Compared to typical test scores (e.g., scale score of 1,270) or ability-level descriptors (e.g., "advanced," "proficient," "basic"),

the logit information is powerful as students and teachers can build upon what students already know to extend their capacity of word recognition. Such item-focused diagnostic information offers a more constructive alternative to broad proficiency categories, which often obscure the diverse sources of reading difficulty thus providing limited instructional guidance (Buly & Valencia, 2002).

The explanatory IRT models revealed that several word feature predictors, included as control variables, had statistically significant and unique effects on oral word reading accuracy—even after accounting for other variables in the model. Specifically, accuracy decreased for words that were longer, more concrete, and acquired later in development (i.e., higher age-of-acquisition). However, the length effect disappeared when decoding demand was included in the models (Models 3a and 4), suggesting that it may not be length alone that influences difficulty. Rather, longer words appear more difficult because they tend to involve greater phonological and orthographic complexity. Indeed, word length was strongly correlated with the number of phonemes ($r = .86$) and decodable demand ($r = .84$) in the sample of words analyzed.

As for the facilitative factors, high-frequency words were read more accurately, and content words such as nouns and verbs were easier than function words (e.g., pronouns such as "she," conjunctions "but" or prepositions "with"), again when other lexical and contextual variables were accounted for. Word position within the passage also exerted a robust influence: words appearing later were more likely to be read inaccurately, potentially reflecting fatigue effects and the need for sustained reading stamina across the passage.

The negative association between concreteness and reading accuracy was unexpected, given that research suggests concreteness and imageability facilitate lexical processing (e.g., Steacy & Compton, 2019). This unexpected finding may reflect the lexical composition of early-grade narratives, where abstract, high-frequency words (e.g., *go, do, make, thing*) dominate (Graesser et al., 2011). Many of these high-frequency, abstract words exhibit irregular orthographic patterns (e.g., *said, one, come*) that may impede recognition despite their ubiquity in students' oral language (Edwards et al., 2024; Steacy & Compton, 2019). When orthographic irregularity combines with high frequency and low concreteness, the typical facilitative effect of concreteness may be offset by decoding demands. Additionally, studies of passage comprehension have reported an inverse relationship between concreteness and comprehension, suggesting that other text features such as narrativity and level of cohesion may also contribute to processing difficulty (Pickren et al., 2022).

When examined separately, each decoding-related predictor—decoding demand, spelling-to-pronunciation transparency, and first-syllable vowel pattern—showed significant effects on word reading accuracy. However, in the final model, only vowel patterns (*long-vowel*, *diphthong*, and *r-controlled*) remained significant, suggesting they serve as more proximal and discriminative indicators of word recognition than broader composite indices. This result suggests suppression effects, where shared variance leads one predictor to become a more effective representation of the construct (Ludlow & Klein, 2014). This highlights the importance of selecting predictors most tightly coupled with the mechanism under investigation—here, the phonological processing demands of vowel graphemes. Notably, complex vowel patterns

comprised substantial proportions of our word sample (14% r-controlled, 10% diphthongs), challenging assumptions that early reading materials predominantly feature simple phonological structures.

Parts of speech also revealed important patterns that address the research question about difficulty of particular word types in typical reading contexts. Nouns (38.3%) and verbs (34.8%) dominated the texts, while function words comprise only 6.1%. Despite their lower frequency in the sample, function words often presented greater difficulty, suggesting that these words may require different instructional approaches than content words.

Limitations

This investigation is a secondary analysis of data originally collected with different research objectives. Consequently, the specific word-level analyses were not prespecified in the original research protocol, and the selection of passages and words for analysis was constrained by the original data collection design. Future research using prospectively collected data might enable more systematic manipulation of word features to test causal relationships.

Several methodological limitations should be considered when interpreting these findings. First, our sample was drawn from specific geographical areas in the United States, which may limit generalizability. Our cross-sectional design also limits our ability to make causal inferences about the development of word recognition skills. The observed patterns of difficulty across ability groups represent snapshots in time rather than developmental trajectories, and individual students' paths through word reading acquisition may vary considerably.

Technical and measurement constraints also affect our findings. While automatic speech recognition (ASR) technology demonstrated strong agreement with human raters in the CORE project (Nese, 2022), we acknowledge potential limitations related to speech pattern variability. Students with non-standard dialects, articulation differences, or early English learner profiles may have experienced differential ASR recognition accuracy. Our analytic approach of selecting students' five best-performing passages may have partially mitigated systematic ASR errors by reducing the influence of any single passage-specific recognition failures. However, given our sample included 11% English learners and students from varied linguistic backgrounds, some differential measurement error related to ASR limitations cannot be entirely ruled out. The large sample of 1,267 word-tokens across 606 unique words and 650 students may have attenuated the effects of sporadic ASR errors on overall findings, though effects on specific word-level patterns warrant future investigation with additional validation methodologies.

While our word feature variables capture multiple dimensions of lexical complexity, they may not fully address certain aspects of word difficulty, including morphological transparency, semantic ambiguity beyond polysemy, and individual differences in instructional exposure. Additionally, sentence-, paragraph- or passage-level influences, such as levels of coherence, syntactic complexity, and background demand, have not been fully investigated. Despite these limitations, our results identify

specific lexical features that challenge developing readers and provide an empirical foundation for targeted instructional interventions.

The R^2 value of .40 in the last model suggests more than half of the item variance remains unexplained. Future research should identify more nuanced predictors with enhanced diagnostic value, such as morphological complexity and syntactic complexity at both the sentence and passage levels. Further, differential facet functioning (Xie & Wilson, 2008) could be investigated to uncover heterogeneity of effects of word-, sentence- and passage-features, by important student characteristics such as fluent vs. not-fluent readers in terms of WCPM, English learner status, and dyslexic and non-dyslexic students.

Future Directions

These findings suggest several promising research directions. First, future work should move beyond static accuracy outcomes to employ joint modeling of accuracy and word-level response times within an EIRM framework (e.g., Potgieter et al., 2026). Because modern online ORF assessments—including the one used to collect the CORE data—automatically capture word-level latencies, we are well-positioned to leverage this granular process data to better understand the reading process. By specifying a multivariate model where latent reading ability and latent processing speed are allowed to covary, we may be able to disentangle the unique impact of word features on processing speed versus accuracy. Such an investigation could capture the fluency-based slowing that often precedes an error, providing a fine-grained decomposition of how position affects the reading process. Furthermore, applying longitudinal EIRMs across multiple time points (e.g., Cho, Athay, et al., 2013; Stevenson et al., 2013; Wilson et al., 2011) would allow us to track the latent growth of word recognition in connected text, determining when sensitivity to features like vowel complexity or spelling transparency begins to diminish as students reach automaticity.

Second, integrating neurocognitive measures with behavioral data could illuminate underlying processes. Eye-tracking could reveal attention allocation differences, while neuroimaging could identify neural networks activated by different word characteristics, bridging cognitive models and classroom instruction. Finally, developing adaptive assessment and instructional technologies based on word-feature profiles represents promising application. Digital platforms could dynamically adjust text difficulty based on individual mastery of specific features, providing personalized experiences and detailed diagnostic profiles enabling precise interventions.

Conclusion

This study demonstrates that the Rasch model and its explanatory extension have the potential to transform routine oral reading assessments into powerful diagnostic tools. By moving beyond aggregate accuracy and fluency measures, we identified specific word features that systematically challenge second-grade readers. The finding that first-syllable vowel patterns—particularly r-controlled vowels, diphthongs, and long vowels—emerged as the strongest decoding-related predictors has direct implications for reading instruction. Rather than relying on broad decoding assess-

ments, educators can focus on these specific orthographic patterns when designing interventions for struggling readers.

The concentration of word difficulty in the lower third of the ability distribution makes these findings particularly relevant for students who struggle most with reading. For these learners, understanding which specific words and word types present challenges in connected text enables targeted instruction that builds on what they already know and can do rather than defaulting to generalized achievement labels grounded in deficit-based models. As schools are increasingly data-driven, the approach we presented here offers a promising pathway toward more nuanced, responsive instruction that meets students' specific needs.

Conflict of Interest Statement

None.

Notes

[1]See https://dibels.amplify.com/report.

[2]In other words, the passages were written to a common specification as narrative stories targeting the same construct of grade–appropriate reading difficulty, therefore we used a unidimentional model rather than a multidimensional model that account for passage topics or genres. Distinct topical or genre clusters were not used in the passage design.

[3]If they paused for more than 10 seconds, they were placed into the next passage.

[4]Items with infit mean square (MNSQ) values > 1.33 and corresponding t-statistics > 1.96 were flagged as underfitting and removed from the analysis. These thresholds are commonly used as indicators of misfit in Rasch modeling (Adams & Khoo, 1996).

[5]Word-tokens (each occurrence of a word) are considered as items. There are more items than unique words as some words appeared more than once within and across passages. Sixty-one percent of the unique words appeared only once among the fifty passages examined; 17% appeared twice, 7% appeared three times, and the rest appeared more than four times.

[6]The modification was in the form of attending to the first syllable of multisyllabic words to better represent the authentic decoding challenges students encounter.

[7]Results were similar for thresholds of 10 and 50.

[8]This is because a .43-logit increase from 0 logit for the reference category corresponds to an increase in the probability of success from 50% to approximately 61%, representing a 10.5 percentage point gain.

References

Adams, R. J., & Khoo, S.-T. (1996). *Quest: The interactive test analysis system*. Version 2.1. Australian Council for Educational Research. https://research.acer.edu.au/measurement/3

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Alonzo, J., Tindal, G., Ulmer, K., & Glasgow, A. (2006). *easyCBM©online progress monitoring assessment system*. Eugene, OR: Behavioral Research and Teaching.

Ardoin, S. P., Eckert, T. L., Christ, T. J., White, M. J., Morena, L. S., January, S. A. A., & Hine, J. F. (2013). Examining variance in reading comprehension among developing readers: Words in context (curriculum-based measurement in reading) versus words out of context (word lists). *School Psychology Review*, *42*(3), 243–261.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education*, *21*(2), 89–118.

Bruner, L., Hiebert, E.H., & Tortorelli, L. (2025). *The role of text vocabulary in word recognition, reading rate, and comprehension of first-grade students*. Paper presented at the annual conference of the Society for the Scientific Study of Reading, Calgary, AB.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45–50.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Buly, M. R., & Valencia, S. W. (2002). Below the bar: Profiles of students who fail state reading assessments. *Educational Evaluation and Policy Analysis*, *24*(3), 219–239.

Carroll, J.B., Davies, P., & Richman, B. (1971). *The American heritage word frequency book*. Houghton Mifflin.

Chee, Q. W., Chow, K. J., Yap, M. J., & Goh, W. D. (2020). Consistency norms for 37,677 English words. *Behavior Research Methods*, *52*(6), 2535–2555.

Cho, S. J., Athay, M., & Preacher, K. J. (2013). Measuring change for a multidimensional test using a generalized explanatory longitudinal item response model. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 353–381.

Cho, S. J., Gilbert, J., & Goodwin, A. (2013). Explanatory multidimensional multilevel random item response model: An application to simultaneous investigation of word and person contributions to multidimensional lexical representations. *Psychometrika*, *78*(4), 830–855.

Cho, S. J., Goodwin, A., Naveiras, M., & De Boeck, P. (2024). Modeling nonlinear effects of person-by-item covariates in explanatory item response models: Exploratory plots and modeling using smooth functions. *Journal of Educational Measurement*, *61*(4), 595–623.

Compton, D. L., Steacy, L. M., Gutiérrez, N., Rigobon, V. M., Edwards, A. A., & Marencin, N. C. (2023). The development of early orthographic representations in children. In S. Q. Cabell, S. B. Neuman, & N. P. Terry (Eds.), *Handbook on the science of early literacy* (312 pp.). The Guilford Press.

Compton, D. L., Appleton, A. C., & Hosp, M. K. (2004). Exploring the relationship between text-leveling systems and reading accuracy and fluency in second-grade students who are average and poor decoders. *Learning Disabilities Research & Practice*, *19*(3), 176–184.

Connor, C. M., Morrison, F. J., & Underwood, P. S. (2007). A second chance in second grade: The independent and cumulative impact of first-and second-grade reading instruction and students' letter-word reading skill growth. *Scientific Studies of Reading*, *11*(3), 199–233.

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164–185.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer.

Edwards, A. A., Rigobon, V. M., Steacy, L. M., & Compton, D. L. (2024). Spelling-to-pronunciation transparency ratings for the 20,000 most frequently written English words. *Behavior Research Methods*, *56*(4), 2828–2841.

Fitzgerald, J., Elmore, J., Koons, H., Hiebert, E. H., Bowen, K., Sanford-Moore, E. E., & Stenner, A. J. (2015). Important text characteristics for early-grades text complexity. *Journal of Educational Psychology*, *107*(1), 4.

Flynn, L. J., Hosp, J. L., Hosp, M. K., & Robbins, K. P. (2011). Word recognition error analysis: Comparing isolated word list and oral passage reading. *Assessment for Effective Intervention*, *36*(3), 167–178.

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.

Fry, E. (2004). Phonics: A large phoneme-grapheme frequency count revised. *Journal of Literacy Research*, *36*(1), 85–98.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (2007). Using curriculum-based measurement to inform reading instruction. *Reading and Writing*, *20*(6), 553–567.

Gao, Y., Stocco, A., & Prat, C. S. (2022). SCOPE: The South Carolina Psycholinguistic Metabase. *Behavior Research Methods*, *54*(6), 2740–2763.

Goodman, K. S. (1969). Analysis of oral reading miscues: Applied psycholinguistics. *Reading Research Quarterly*, *5*, 9–30.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2011). Coh–Metrix: Providing multilevel analyses of text characteristics. *Behavior Research Methods*, *43*(3), 581–601.

Guthrie, J. T., & Seifert, M. (1977). Letter-sound complexity in learning to identify words. *Journal of Educational Psychology*, *69*(6), 686–696.

Hair J. F., Black W. C., Babin B. J., Anderson R. E., Tatham R. L. (2006). *Multivariate data analysis* (6th Ed.). Pearson/Prentice-Hall.

Hagans, K. S. (2008). A response-to-intervention approach to decreasing early literacy differences in first graders from different socioeconomic backgrounds: Evidence for the intervention validity of the DIBELS. *Assessment for Effective Intervention*, *34*(1), 35–42.

Hartig, J., & Buchholz, J. (2012). *A multilevel item response model for item position effects and individual persistence*. Psychological Test and Assessment Modeling, *54*(4), 418–431.

Hasbrouck, J. & Tindal, G. (2017). *An update to compiled ORF norms* (Technical Report No. 1702). Behavioral Research and Teaching, University of Oregon.

Hiebert, E. H., Toyama, Y., & Irey, R. (2020). Features of known and unknown words for first graders of different proficiency levels in winter and spring. *Education Sciences*, *10*(12), 389.

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. *Psychology Science Quarterly*, *50*(4), 391–402.

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). *The ABCs of CBM: A practical guide to curriculum-based measurement*. Guilford Publications.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Accuracy and fluency in list and context reading of skilled and RD groups: Absolute and relative performance levels. *Learning Disabilities Research & Practice*, *18*(4), 237–245.

Kara, Y., Kamata, A., Ozkeskin, E. E., Qiao, X., & Nese, J. F. (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.

Kearns, D. M., & Hiebert, E. H. (2022). The word complexity of primary-level texts: Differences between first and third grade in widely used curricula. *Reading Research Quarterly*, *57*(1), 255–285.

King, S., Rodgers, D., & Lemons, C. J. (2022). The effect of supplemental Reading instruction on fluency outcomes for children with down syndrome: a closer look at curriculum-based measures. *Exceptional Children*, *88*(4), 421–441.

Kulesz, P. A., Francis, D. J., Barnes, M. A., & Fletcher, J. M. (2016). The influence of properties of the test and their interactions with reader characteristics on reading comprehension: An explanatory item response study. *Journal of Educational Psychology*, *108*(8), 1078–1097. https://doi.org/10.1037/edu0000126

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*, 978–990.

Liceralde, V. R. T., Loukina, A., Beigman Klebanov, B., & Lockwood, J. R. (2022). Beyond text complexity: Production-related sources of text-based variability in oral reading fluency. *Journal of Educational Psychology*, *114*(1), 16.

Ludlow, L., & Klein, K. (2014). Suppressor variables: The difference between 'is' versus 'acting as'. *Journal of Statistics Education*, *22*(2), 1–22.

Menon, S., & Hiebert, E. H. (2005). A comparison of first graders' reading with little books or literature-based basal anthologies. *Reading research quarterly*, *40*(1), 12–38.

Mislevy, R. J. (1987). *Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters* (RR-87-26-ONR). Princeton, NJ: Educational Testing Service.

Moats Louisa, C. (2000). *Speech to print: Language essentials for teachers of reading & spelling*. Baltimore: Paul H. Brookes Publishing Co.

Monsell, S., Doyle, M. C., & Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, *118*(1), 43.

Nese, J. F. (2022). *Comparing the growth and predictive performance of a traditional oral reading fluency measure with an experimental novel measure*. AERA Open, *8*.

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.

Nicholson, T. (1991). Do children read words better in context or in lists? A classic study revisited. *Journal of Educational Psychology*, *83*(4), 444.

Pickren, S. E., Stacy, M., Del Tufo, S. N., Spencer, M., & Cutting, L. E. (2022). The contribution of text characteristics to reading comprehension: Investigating the influence of text emotionality. *Reading Research Quarterly*, *57*(2), 649–667.

Pirani-McGurl, C. A. (2009). *The use of item response theory in developing a Phonics Diagnostic Inventory*. University of Massachusetts Amherst.

Potgieter, C. J., Kamata, A., Kara, Y., & Qiao, X. (2026). Joint analysis of dispersed count-time data using a bivariate latent factor model. *British Journal of Mathematical and Statistical Psychology*, *79*(1), 207–228.

Preston, K. A. (1935). The speed of word perception and its relation to reading ability. *The Journal of General Psychology*, *13*(1), 199–203.

Robitzsch, A., Kiefer, T., & Wu, M. (2025). Package 'TAM'. *Test Analysis Modules–Version*, *4*, 2–21.

Saha, N. M., Cutting, L. E., Del Tufo, S., & Bailey, S. (2021). Initial validation of a measure of decoding difficulty as a unique predictor of miscues and passage reading fluency. *Reading and Writing*, *34*, 497–527.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Spencer, K. (2010). Predicting children's word-reading accuracy for common English words: The effect of word transparency and complexity. *British Journal of Psychology*, *101*(3), 519–543.

Spira, E. G., Bracken, S. S., & Fischel, J. E. (2005). Predicting improvement after first-grade reading difficulties: the effects of oral language, emergent literacy, and behavior skills. *Developmental psychology*, *41*(1), 225.

Stahl, S. A., & Heubach, K. M. (2005). Fluency-oriented reading instruction. *Journal of literacy Research*, *37*(1), 25–60.

Steacy, L. M., & Compton, D. L. (2019). Examining the role of imageability and regularity in word reading accuracy and learning efficiency among first and second graders at risk for reading disabilities. *Journal of experimental child psychology*, *178*, 226–250.

Steacy, L. M., Edwards, A. A., Rigobon, V. M., Gutierrez, N., Marencin, N. C., Siegelman, N., …Compton, D. L. (2023). Set for variability as a critical predictor of word reading: Potential implications for early identification and treatment of dyslexia. *Reading Research Quarterly*, *58*(2), 254–267.

Stevenson, C. E., Hickendorff, M., Resing, W., Heiser, W. J., & de Boeck, P. A. (2013). 106 Explanatory item response modeling of children's change on a dynamic test of analogical reasoning. *Intelligence*, *41*(3), 157–168.

Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1140–1154.

Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *International Scholarly Research Notices*, *2013*(1), 958530.

Toste, J. R., Fluhler, S. K., Farris, E. A., & Chandler, B. W. (2025). What's in a word? Analyzing students' oral reading fluency to inform instructional decision-making. *Intervention in School and Clinic*,*60*(5), 251–261.

Toyama, Y. (2021). What makes reading difficult? An investigation of the contributions of passage, task, and reader characteristics on comprehension performance. *Reading Research Quarterly*, *56*(4), 633–642.

University of Oregon (2022). DIBELS 8th Edition 2021–2022 Percentiles (Technical Report 2201). Eugene, OR: Author. Available: https://dibels.uoregon.edu

Weber, R.-M. (1970). A linguistic analysis of first-grade reading errors. *Reading Research Quarterly*, *5*(3), 427–451.

Wilson, M. (2005, 2023). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.

Wilson, M., Zheng, X., & McGuire, L. (2011). Formulating latent growth using an explanatory item response model approach. *Journal of applied measurement*, *13*(1), 1–22.

Xie, Y., & Wilson, M. (2008). Evaluating the stability of the measurement of person fit. *Educational and Psychological Measurement*, *68*(3), 415–436.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). The educator's word frequency guide. Touchstone Applied Science Associates.

Zoccolotti, P., De Luca, M., Di Pace, E., Gasperini, F., Judica, A., & Spinelli, D. (2005). Word length effect in early reading and in developmental dyslexia. *Brain and language*, *93*(3), 369–373.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.