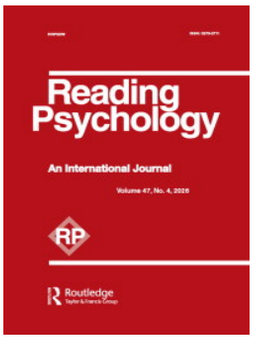

Reading Psychology

ISSN: 0270-2711 (Print) 1521-0685 (Online) Journal homepage: www.tandfonline.com/journals/urpy20

The *Word Recognition Record* (WR^2): An Oral Reading Error Analysis Framework for Use with One-Minute Fluency Probes

D. Ray Reutzel, Janice A. Dole & Parker C. Fawson

To cite this article: D. Ray Reutzel, Janice A. Dole & Parker C. Fawson (2026) The *Word Recognition Record* (WR^2): An Oral Reading Error Analysis Framework for Use with One-Minute Fluency Probes, Reading Psychology, 47:4, 661-692, DOI: 10.1080/02702711.2025.2579298

To link to this article: https://doi.org/10.1080/02702711.2025.2579298

Published online: 31 Oct 2025.

Submit your article to this journal

Article views: 93

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=urpy20

READING PSYCHOLOGY
2026, VOL. 47, NO. 4, 661–692
https://doi.org/10.1080/02702711.2025.2579298

Routledge
Taylor & Francis Group

Check for updates

The *Word Recognition Record* (WR^2): An Oral Reading Error Analysis Framework for Use with One-Minute Fluency Probes

D. Ray Reutzel, Janice A. Dole and Parker C. Fawson

Center for the School of the Future, Utah State University, Logan, Utah, USA

ABSTRACT

There were two purposes for this study. First, we describe the development of the *Word Recognition Record* (WR^2) error analysis framework for examining student oral reading errors as an augmentation to administrations of one-minute fluency probes. Second, we examined how reliably teachers can use the *Word Recognition Record* (WR^2) framework to analyze oral reading errors of students' one-minute fluency probes using two DIBELS 8 ORF benchmarking/progress monitoring passages. The results include the developed *Word Recognition Record Administration* and *Scoring Protocol and* the *Word Recognition Record Form* for analyzing oral reading error patterns from administrations of one-minute fluency probes. The Generalizability (G&D) study showed that primary grade teachers can reliably use the WR^2 to analyze oral reading errors from administration of one-minute fluency probes.

ARTICLE HISTORY
Received 13 October 2025
Accepted 20 October 2025

Individual observations of students' oral reading errors have been a cornerstone of reading assessment since the early days of reading research (Cattell, 1886; Huey, 1908; Thorndike, 1917; Wundt, 1910–1920). Oral reading remains a highly valued and valid method for assessing both the act and the processes of reading. Silent reading, as compared with oral reading, poses significant challenges for reading assessment because of its internal nature requiring advanced, expensive, difficult to access, and often impractical physiological tools (e.g., eye movement tracking, FMRI, MSI, PET, NIRS). When silent reading is assessed in schools, it is usually

CONTACT D. Ray Reutzel ray.reutzel@usu.edu Center for the School of the Future, Utah State University, Logan, Utah, USA.

© 2025 Taylor & Francis Group, LLC

done after the reading is complete, providing little information about the reading process. Because of these challenges, oral reading has become the preferred direct observation method for both laboratory and classroom settings.

In today's schools, one-minute fluency probes are widely used to benchmark and progress monitor young children's reading acquisition. While one-minute fluency probes represent an improvement over past oral reading assessment methods (e.g., informal reading inventories and running records) by including fluency measures in addition to accuracy measures, many primary-grade teachers have noted that a summary score (or "dashboard" indicator) does not provide sufficient information to guide instructional decision making. Some primary grade teachers are intuitively aware that listening to a student read provides rich insights into the reading process and have developed *ad hoc* methods to analyze oral reading errors after administering one-minute fluency probes.

Without deeper analysis of oral reading errors, the counting of students' oral reading errors cannot lead to better informed reading instruction for students. Robinson (1968, p. 401) suggested more than half a century ago that "a wealth of information about [reading] processes could be secured from carefully planned … examination of children's reading behavior [errors]." It is precisely for this reason that students' oral reading errors have been historically studied for clues or insights into how readers are processing text and how teachers might use these insights to inform instructional decision making.

This study describes the development and reliability of the *Word Recognition Record* (WR^2). This instrument and process is an augmentation for widely used one-minute fluency probes to provide an evidence-based framework for analyzing oral reading error patterns across time to potentially inform teacher instructional decision making.

Theoretical Framework and Rationale

The theoretical framework used for this study is based on Hoover and Gough (1990) expansion of Gough and Tunmer (1986) *simple view of reading*. Gough and Tunmer (1986) argued that decoding is "of central importance in reading, for without it, linguistic comprehension is of no use" (p. 128). Skilled reading, according to this view, is a complex process consisting of two critical components – decoding and language comprehension (Gough & Tunmer, 1986). While each component is separate and can be dissociated, the two components are interconnected and interdependent. Anderson and Pearson (1984) argued that the act of skilled reading is much like a symphony orchestra with each individual

instrument playing a part. Each instrument is necessary but insufficient, like the way decoding and comprehension are necessary but insufficient in themselves. Skilled reading occurs when the components are woven together.

Hoover and Gough (1990) expanded on the simple view of skilled reading by identifying specific elements of decoding and language comprehension. They defined decoding as "efficient word recognition," and language comprehension as the ability to make meaning from oral language (p. 130). Hoover and Gough (1990) verified this simple view by empirically demonstrating that decoding and comprehension accounted for a large proportion of the variance in reading comprehension from .71 to .83 at the third-grade level.

Hoover and Gough (1990) model has been further adapted into Scarborough's (2001) *Reading Rope* which represents the reemergence of a scientific view of skilled reading. Since this study addresses the assessment of oral reading errors that can be observed directly, the word recognition strand of Scarborough's rope is the focus.

Word recognition is a process used in reading, according to Willingham (2017), to recover speech. It involves recognizing written symbols, associating symbols with corresponding speech sounds, and blending these speech sounds in temporal order to say words. Many written words are efficiently recognized when specific one-syllable spelling patterns are used singly or in combination with others as is often the case in recognizing multiple syllable words. Some written words, typically those irregularly spelled, cannot be straight forwardly translated into spoken sounds and blended to say words. These are often learned through memorization of these words' orthographic patterns or spellings and in some cases through frequent encounters in print. These words are often referred to as high frequency words (Mesmer, 2024). For the word recognition process to facilitate early reading acquisition, readers must not only accurately read words but also do so fluently.

Fluent oral reading is defined as accurate, automatic or effortless, and expressive (Rasinski et al., 2011). Accurate reading means orally speaking words that are in the text. Automatic or effortless reading means orally speaking words in the text quickly, without hesitation, and with proper phrasing. Expressive oral reading means reading the text with appropriate volume to be heard while articulating the prosodic features of stress, juncture and pitch. Stress is the emphasis, usually marked by volume, that words are read with orally. Juncture is making appropriate breaks between syllables and words to allow listeners to comprehend the words, i.e., I scream vs. ice cream. Pitch is the raising or lowering of tonal aspects of reading words orally.

When readers can both read words accurately and fluently, the decoding process requires less of the limited amount of working memory capacity thus freeing up this limited resource for linking spoken words to vocabulary meanings and eventual comprehension of text (Pressley, 2000). Taken together, these elements of the decoding process are typically categorized as "foundational" reading skills (CCSS, 2010).

This study, in addition to describing the development of the *Word Recognition Record* (WR^2), of necessity, examines how reliably teachers can analyze student oral reading errors using this tool. Eventually, such analyses of oral reading errors may serve as a diagnostic guide for teachers to recognize error patterns across texts to help them plan more effective reading instruction, but not before they can reliably analyze oral reading errors. To better understand the context and need for this study, a brief review of approaches for observing and analyzing student oral reading errors is presented.

Assessing Oral Reading: Review of Literature

For more than a century of reading research and assessment, there have been three dominant examples of oral reading assessments used in classroom settings: (1) informal reading inventories (IRIs) (Betts, 1936, 1946; Ekwall, 1976; Johnson & Kress, 1965), (2) running records (RRs) (Clay, 1972, 1979, 1993), and (3) CBM one-minute oral reading fluency probes (Deno, 1985; Kaminski, 1992). Each of these three oral reading assessments are described as background for this study.

Informal reading inventories

Informal Reading Inventories (IRIs) emerged in the mid-20th century as teachers sought alternatives to standardized norm-referenced reading tests. The primary purpose of IRIs as originally proposed by E. A. Betts (1946) was to determine each student's three reading levels: *independent level* (95–100% word accuracy, strong comprehension), *instructional level* (90–94% word accuracy, adequate comprehension with support), and *frustration level* (below 90%-word accuracy or poor comprehension).

To accomplish this aim, an IRI is an individually administered assessment that evaluates each student's *word recognition, oral reading fluency,* and *comprehension*. As each student reads graded passages aloud, examiners follow along on another copy marking oral reading errors or *miscues* (Goodman, 1965, 1969) such as *substitutions, omissions, insertions,* and *self-corrections* noting the number of errors and oral reading error patterns. To assess passage reading fluency, examiners sometimes timed the oral reading of a passage to determine the number of words read correctly per minute (WCPM).

IRIs have been used to group students for instruction, select appropriate levels of texts, and identify word recognition and comprehension weaknesses to plan for targeted interventions or lessons. More recent versions of IRIs have added comprehension retellings, more sophisticated error analysis systems, the use of varied text genres, and are available in digitized formats.

IRIs have recently come under increased scrutiny where questions have been raised about the validity of the three levels of reading—*independent, instructional,* and *frustration* – originally proposed by E. A. Betts (1946). Shanahan (2011) asserts that these three levels of reading ability are unsupported by research evidence. He has also questioned the related practice of matching books to students' reading levels as potentially harmful because recent evidence suggests some students placed in more difficult text levels make greater growth in reading proficiency.

Running records

Running records (RRs) are an informal, formative reading assessment tool used to evaluate a student's oral reading performance originally developed by M. Clay (1979, 1993). Examiners capture what a child says while reading a text aloud, allowing the teacher to analyze reading accuracy, error patterns (miscues), self-corrections, and fluency. Clay's (1993) oral reading analysis framework abbreviated as—**MSV** – **M**eaning (semantic), **S**tructure (syntactic), and **V**isual (grapho-phonic) cues, attempted to understand how readers integrate these three cueing systems to comprehend text.

Like an IRI, running records (RR) involve listening to a child read a leveled text aloud, recording each word read correctly, marking any miscues, and noting self-corrections. After reading, examiners calculate accuracy and error rates as well as a self-correction ratio. Errors or miscues are then subjected to further qualitative analysis using the MSV framework. Like IRIs, RRs are also used to determine a student's *independent, instructional,* and *frustration* reading levels.

Running Records have been used to diagnose which reading cues students are using and how, grouping and regrouping students for instruction, selecting appropriate text difficulty levels, and designing targeted instruction. Running Records are central to widely adopted instructional programs like *Reading Recovery* (Clay, 1993) and Fountas and Pinnell (1996) *Guided Reading.* As in the case of IRIs, RRs have also been adapted for digital use in classrooms.

RRs have also recently come under increased scrutiny where questions have been raised about the reliability of text leveling schemes (Amendum et al., 2018; D'Agostino & Briggs, 2025; Shanahan, 2009, 2025; Strong et al., 2018), the ability of teachers' to reliably use RRs (Fawson et al., 2006; D'Agostino et al., 2021), and a reliance on the "three cueing

systems," which is not well aligned with cognitive models of how reading works (Greene, 2024; NCTQ, 2023a; Seidenberg, 2017; Shanahan, 2021, 2024; Willingham, 2017).

One-minute fluency probes

A *One-minute fluency probe (OMFPs)* is a standardized, timed assessment in which a student reads a passage aloud for 60 seconds. The OMFP, originally developed by S. Deno (1985), is the cornerstone of *Curriculum-Based Measurement* (CBM) and plays a major role in progress monitoring within *Response to Intervention/Multi-Tiered Systems of Support* (RTI/MTSS) frameworks.

Administering a one-minute fluency probe involves listening to a child read a grade level text aloud for 60 seconds, tallying the total number of words read correctly and marking any errors or miscues. The score obtained from an administration of a one-minute fluency probe is the number of words read correctly in one minute resulting in a *words correct per minute* (WCPM) score. *WCPM* scores are then compared to established benchmarks to assess overall reading proficiency (Hasbrouck & Tindal, 2006, 2017).

OMFPs were designed to be short, simple, and sensitive to small changes in reading proficiency over time and have become one of the most widely used progress monitoring assessments in schools today. OMFPs are an effective and cost-efficient tool for screening students who may be at risk for future reading difficulties and are strongly correlated with comprehension and reading proficiency (Fuchs et al., 1988; Barrett, Johnson, Truckenmiller et al., 2024). The one-minute fluency probe is a central feature of reading assessment as found in the *DIBELS (Dynamic Indicators of Basic Early Literacy Skills – ORF)* assessment suite**,** *AIMSweb***,** and *Acadience Reading*. One-minute fluency probes are closely aligned with the Science of Reading (SOR) (National Reading Panel, 2000; Reading League, 2022).

Although one-minute fluency probes are efficient and predictive of overall reading achievement, several scholars argue that OMFPs privilege speed over comprehension, potentially misrepresenting a student's ability to construct meaning from text (Samuels, 2007; Valencia et al., 2010). Others note that these brief measures reduce reading to word recognition and automaticity only, overlooking other scientifically supported elements of reading instruction (Afflerbach et al., 2008; Paris, 2005). Additional concerns include variability introduced by short timing windows, inconsistent scoring rules, and cultural or linguistic bias, particularly for English learners (Fuchs et al., 2001; Hosp & Hosp, 2003). Overall, many researchers emphasize that OMFP should not be used as a stand-alone indicator

of reading proficiency or used to make high stakes instructional decisions, but rather as only one formative component within a broader assessment framework that includes comprehension and other strategic reading measures (Kuhn & Stahl, 2003; Valencia et al., 2010).

The ORFPs and its commercialized derivatives – *DIBELS ORF, AimsWeb, and Acadience Reading*, are used today in all 50 states and with millions of students nationwide to assess oral reading. Even though oral reading errors are part of the OMFP assessment, these errors are principally recorded for the purpose of calculating a *words correct per minute* (*wcpm)* score to compare against established reading fluency standards.

Parker et al. (1992) somewhat presciently suggested many years ago that OMFP scores might be made more powerful as a tool for instructional decision making if an analysis of student's oral reading errors was added. But as it now stands, teachers get only a summary score (WCPM) or color benchmark rating from proprietary data processing companies, e.g., Acadience®, Amplify® (red, yellow, green, & blue) of their students' individual oral readings of DIBELS ORF passages. Summary data such as wcpm or color-coded benchmark levels, while useful as a "dashboard" indicator that a reading problem is or is not occurring, neither encourage that teachers gain insights into the types of errors their students are making nor information about recurring error patterns across several assessment occasions that might be used to determine which elements of word recognition skills may require added instructional support in early reading instruction. As such, studies are needed to bridge this important *practice to research gap*.

There were two purposes for this initial study. First, we describe the development of the *Word Recognition Record* (WR^2) error analysis framework for examining student oral reading errors as an augmentation to the administrations of one-minute fluency probes. Second, we examine how reliably teachers can be trained to use the *Word Recognition Record* (WR^2) framework to analyze oral reading errors of first-grade students' readings of DIBELS 8 ORF (2022) benchmark or progress monitoring passages. With these twin purposes in mind, we pursued answers to the following research questions.

Research Questions

1. What is the *Word Recognition Record* (WR^2), and how was it developed?
2. How reliably can teachers use the WR^2 framework to analyze student oral reading errors across multiple word recognition skill components?

Method

Research Design

After obtaining IRB approval from the university and the school district, this study was conducted in two phases. In Phase 1, we employed an instrument development design (Henk et al., 2011; Walpole et al., 2021) approach to create the *Word Recognition Record* (WR^2). According to Henk et al. (2011), the utility of all research findings and the effectiveness of all instructional practice decisions are necessarily predicated upon the integrity of the assessment tools and processes used to evaluate the phenomenon under study.

In phase 2, we employed a quantitative analysis, *Generalizability (G)* and *Dependability (D)* study, of experienced first-grade teachers' ability to reliably use the *Word Recognition Record* (WR^2) to analyze student oral reading errors across multiple word recognition skill components. In classical test theory, reliability addresses variability and error by factoring an observed score into a true score and an estimate of error. Generalizability theory, based on modern test theory, allows for a more granular examination of systematic sources of variability (Strube, 2000).

Participants

As previously noted, in phase 1 of this study, we worked with a team of four university SOR reading faculty and with a team of district level reading coaches to develop the *Word Recognition Record* (WR^2). All the university faculty who participated on the development team had published multiple research studies cited by the National Reading Panel (2000) report and in internationally peer-reviewed top-tier reading research journals.

Reading coaches who participated on the *Word Recognition Record* (WR^2) development team all had master's degrees in literacy or related fields and were LETRS trained. LETRS professional development is a well-established national training program in the instructional elements of the Science of Reading (SOR) (Moats & Tolman, 2019).

In phase 2 of this study, ten experienced, first-grade teachers in one school district in a state in the western U.S volunteered for participation in this project. First-grade teachers were chosen for this project because students at this level are moving from recognizing single syllable word patterns into words that employ these single syllable patterns in multiple syllable words offering the broadest possible analysis of early reader word recognition errors. Additionally, students

in subsequent grade levels are exposed to texts with the same single syllable and multiple syllable word patterns for oral reading error analysis.

The school district that was selected for the study is considered a suburban-rural district using the state education agency's description of state school districts. The schools from which these teachers were selected ranged from high poverty, low achieving, Title 1 schools with 40% diversity to schools that were affluent, high achieving schools with less than 10% diversity. The four schools in which the teachers taught in this study ranged from a low of 21% to a high of 51% of students receiving free and reduced lunch. Students in the less diverse schools in this district spoke English as their primary language. Those ML students in the more diverse schools in this district spoke primarily Spanish as their heritage language. Demographic information about the ten, first-grade classroom teacher raters in phase 2 of the study is shown in Table 1.

Table 1. Demographic characteristics of the 10 first-grade teacher raters.

First-grade teachers	Number of years teaching	Number of years teaching at first-grade level	Highest degree	Endorsement, licenses, certificates	Self-rating of teacher knowledge about reading foundation skills
Teacher 1	10	7	2 Masters Degrees	• Utah State ELL Endorsement • LETRS Trained	Moderate
Teacher 2	5	4	Bachelors	• Utah State Reading Endorsement • LETRS Trained	Moderate
Teacher 3	16	4	Masters	• LETRS Trained	High
Teacher 4	16	16	Bachelors	• Utah State Math Endorsement • LETRS Trained	High
Teacher 5	4	2	Bachelors	• LETRS Trained	High
Teacher 6	16	6	Masters	• Early Childhood Endorsement • LETRS Trained	Moderate
Teacher 7	5	5	Bachelors	• Utah State ELL Endorsement • World Languages Endorsement • LETRS Trained	High
Teacher 8	8	3	Masters	• Utah State • ELL Endorsement • LETRS Trained	Moderate
Teacher 9	7	4	Bachelors	• Utah State ELL Endorsement • LETRS Trained	High Moderate
Teacher 10	9	3	Bachelors	• LETRS Trained	Moderate

Procedures

Phase 1: Developing the Word Recognition Record (WR^2)

The WR^2 was developed iteratively using traditionally accepted instrument development procedures and guidelines (Crocker & Algina, 1986; Ebel & Fiesbie, 1991; Henk et al., 2011; Osterlind, 1989; Thorndike, 1982, Walpole et al., 2021). First, we conducted a research synthesis of national and international reports on the *Science of Reading* which were published by the federal government agencies and private foundations that described the elements of "word recognition" and "foundations of early reading" instruction (Burns et al., 1998; Foorman et al., 2016; National Early Literacy Panel. Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, DHHS, 2010; National Reading Panel, 2000, Scarborough, 2002). This information was then used to determine domain specifications for elements to be included in the pilot WR^2 instrument. The domain specifications were modified as the development team of university researchers and the cadre of K-6 school literacy coaches worked through the initial versions of the pilot WR^2 instrument to develop the training and use protocols.

Based on the initial domain specifications, a WR^2 prototype was developed by (a) writing a purpose statement, (b) clarifying domain specifications, (c) writing instrument specifications, and (d) developing an error recording and analysis coding form.

Purpose statement. A clearly defined purpose statement was developed which described the object of measurement for which the WR^2 was intended, the purposes for which the resulting scores were intended to be used, and the constraints under which the WR^2 should be used (i.e., time required and preferred conditions for use) as outlined in widely recognized classical and modern testing theory and development texts (Crocker & Algina, 1986; Ebel & Fiesbie, 1991; Osterlind, 1989; Thorndike, 1982). We determined the following purpose statement for the WR^2:

> *The* WR^2 *is intended for use with the one-minute fluency probe to mark, count, and analyze students' oral reading word recognition errors, including fluency indicators, to potentially inform teachers' future word recognition instructional decisions in the primary grades.*

Domain specifications. Based on informed professional judgment and information gleaned from readings of federal government and private foundation reports, the researchers developed domain specifications for word recognition and foundations of early reading instruction error analysis using a content analysis process called "conceptualizing decisions"

(Neuendorf, 2002). The word recognition and foundations of early reading instruction domains specified included: phonics (single syllable word patterns, multi-syllable words), high frequency words, and fluency including accuracy, rate, and expression. We omitted specification of some foundational early reading instructional elements to include phonological and phonemic awareness, concepts about print, oral language development, and alphabetic principle because these are typically considered emergent or pre-reading skills. In addition, we did not think we could directly observe these emergent skills in an error analysis of oral reading.

For phonics we focused error analysis on single syllable patterns in words using CV, VC, CVC, VCe, CVVC, and r-controlled patterns. Multiple syllable words, composed of two or more of these single syllable patterns, e.g., suntan—CVC and CVC; or comfort—CVC and C r-controlled C, etc., were also included in our phonics domain specification.

High frequency words, words that are highly frequent in texts, are a part of early word recognition processes and are of two types (Shapiro & Solity, 2016). There are those that can be approached through typical decoding processes of blending and segmenting, i.e., me, had, more, make, etc., because these are spelled regularly—meaning the sounds of letters are as expected. Then there are those high frequency words that appear often in text that are irregularly spelled—i.e., said, some, was, the, etc. These words are typically learned by memorization or partial decoding (Colenbrander et al., 2022). As for fluency, expression was viewed through the lenses of volume, inflection (pitch, juncture and stress as previously discussed), and phrasing (Zutell & Rasinski, 1991). These specified domains were then used to frame the establishment of the categories contained in the WR^2 use protocol and scoring form.

Four nationally recognized researchers in the field of early reading, employed as teacher educators at national research-extensive universities, were invited to review the pilot domain specifications to determine that no relevant dimension or category related to word recognition foundational skill areas had been overlooked in the domain specifications. These researchers indicated that the error categories in the WR^2 met expectations and included generally accepted definitions and descriptions (See Figure 1).

Writing instrument specifications. Instrument specifications define the format, characteristics of oral reading error descriptors, directions for recording each error, and a procedure or protocol for coding errors (Crocker & Algina, 1986; Ebel & Fiesbie, 1991; Osterlind, 1989; Thorndike, 1982). The item format included: (a) an error recording and analysis form containing a heading and definition for each relevant reading foundation

Word Recognition Areas	Description	Example Research Support
Phonics	Systematic phonics instruction stresses the acquisition of letter-sound correspondences and their use to read and spell words.	National Reading Panel (2000) WWC Practitioner's Guide: Foundational Skills in K-3 (2016)
High Frequency Words	High-frequency words are of two types: regularly and irregularly spelled words. Students are taught to read regularly spelled words using phonic blending skills. Irregularly spelled words are learned through memorization of spelling patterns.	National Reading Panel (2000) WWC Practitioner's Guide: Foundational Skills in K-3 (2016)
Reading Fluency	Fluent readers can read text with speed, accuracy, and proper expression.	National Reading Panel (2000) WWC Practitioner's Guide: Foundational Skills in K-3 (2016)

Figure 1. Word recognition categories obtained from a systematic literature search.

domain, (b) reading foundation categories for recording each oral reading error, and (c) a summary error tally by reading foundation category.

The general guideline established for writing item definitions specified that the definition had to be a succinct, clear description of each reading foundation domain. In addition, a pilot WR^2 use protocol was also developed which included step-by-step directions for using the WR^2 to record and analyze student's oral reading errors after they had been administered a one-minute fluency probe using grade-level benchmark or progress monitoring passage.

Error recording and analysis coding form. The number of categories written for the WR^2 corresponded with the number of word recognition and reading foundational skill areas identified in the table of specifications (Thorndike, 1982). Initial tryouts of the pilot WR^2 instrument were conducted with five trained reading coaches who rated 10 randomly selected first-grade student oral reading samples of grade level ORF passages. Follow up tryouts and refinements discussed by the raters resulted in a (95%) agreement among these raters in using the WR^2 scoring and analysis form for number of errors coded, words read per minute, and number of error categories coded. There was less agreement on fluency ratings (88%) among these initial raters. When there were disagreements in fluency ratings, we noted these were often adjacent

scores and not wide variations, e.g., a 2 or 3 or a 1 and 2, not 1 and 4 (See Appendixes A and B for the final version of the WR^2 instrument and its use protocol).

Phase 2: WR^2 reliability study

The ten first-grade teachers who volunteered for the study had participated in a *Word Recognition Record* (WR^2) professional development training module. We developed a WR^2 professional development training module to prepare 10 first-grade teachers on how to use the WR^2 in their classrooms to reliably record and categorize first-grade students' oral reading errors once the students had completed reading three test passages. A power point presentation was developed that contained step-by-step directions and examples of how to use the WR^2 instrument. The training of these 10 first-grade teachers required two and one half to three hours and was held at the district office.

We used multiple (N = 3) student audio recordings of first-grade students reading DIBELS 8 ORF benchmark/progress monitoring passages orally followed by agreement discussions until teachers felt that they could use the WR^2instrument reliably. DIBELS 8 ORF (2022) benchmark/progress monitoring passages were selected for this study because of their validated grade level reading difficulty. After the conclusion of the training session, first-grade teacher ratings of audio student reading samples were assessed for initial reliability of average total errors coded across the three training passages. These 10 first-grade teachers were able to adequately code total errors with 98% average accuracy, signaling a reasonable baseline for a further reliability study of the deployment of the WR^2 instrument and process in independent ratings.

After completion of the WR^2 professional development training module, teachers were given instructions about how to participate in the phase 2 WR^2 reliability study. Each of the ten first-grade teachers was to independently score and analyze audio recordings of 10 first-grade students' one-minute fluency probe reading a first-grade benchmark/progress monitoring passage: *Lucky Day* (N = 10 student recordings). The 10 first grade student oral reading samples used in the reliability study were chosen from three strata of first-grade student reading proficiency levels: *well below, below,* and *at or above*. Next, teachers were asked to record student oral reading errors and categorize these errors into the WR^2 categories.

As the teachers identified or categorized each student's oral reading errors, they were also asked to write comments about their decision-making processes in "comment boxes" in each section of the WR^2. Once the initial audio files for passage 1, *Lucky Day,* were scored using the WR^2,

teachers were asked to send the completed WR² forms to the researchers via a secure online BOX file.

Next, teachers were instructed to rate a second set of 10 first-grade students' one-minute fluency probe audio recordings using the passage, *Our Pond* (N = 10 student recordings). The addition of a second testing occasion allowed researchers to examine "intra-rater" or "occasion" variance among the teacher raters using the WR²instrument. Once the second task was completed, the teachers once again sent their second ratings to the researchers via the same secure online BOX file. All 10 teachers completed all 10-student oral reading audio recording for two grade level passages resulting in a complete data set with no missing data.

Data Analysis

Using the R statistical program, we conducted a two-facet (rater x occasion), fully crossed Generalizability study (G & D study – GENOVA) to examine inter and intra-rater reliability as well as to determine cost/benefit advantages for reducing rater error and increasing rater reliability for absolute and relative decisions related to the use of the WR² instrument (Brennan, 2001; Brennan & Johnson, 1995; Briesch et al., 2014; Cronbach et al., 1972; Shavelson & Webb, 1991; Strube, 2000). We selected a Generalizability (G&D) analysis approach to avoid the known problems associated with employing classical testing theory related to variability or error. This is accomplished by factoring an observed score into a true score and an estimate of error. Unlike classical test theory, generalizability studies do not ignore systematic sources of variability by sweeping them under the virtual rug as *error variance*. Instead, generalizability studies assume that the effects of these factors are systematic rather than random. Most users of tests and other psycho-social instrumentation are interested in their suitability as a function of when the measurement is taken, who takes the measurement, and how many times a measurement is administered (Shavelson & Webb, 1991). One advantage of generalizability studies is that error estimates can be factored into multiple sources of errors using a unique type of analysis of variance (i.e., GENOVA) (Brennan & Johnson, 1995; Crick & Brennen, 1983; Strube, 2000).

Thus, the purpose of this generalizability study (G study) was to evaluate 10 first-grade teachers' use of the WR² to reliably record and categorize student oral reading errors from 10 first grade readers reading two first-grade DIBELS ORF benchmark/progress monitoring passages (*Lucky Day* and *Our Pond*). If the number of raters needed to

achieve acceptable reliability (.70) was minimal (2 raters or less on one rating occasion), then the WR^2 reliability training would be regarded a success.

By using a generalizability study, the relative contribution of the potential sources of error to the total error variance (Strube, 2000) could be determined. In other words, a G study attempts to measure how various factors may affect a measurement instrument or rater score and the reliability of that instrument or rater. Generalizability studies have been conducted for very few psycho-socio measurement instruments or for rater reliability studies because of the "demanding data collection and formidable mathematics" involved in such studies (Pressley & McCormick, 1995, p. 528).

In this G study, 10 first-grade students' audio reading samples were the object (person) of measurement, and two factors were facets of measurement (i.e., teacher raters – 10 first grade teachers, and occasions—two passage readings) (Shavelson & Webb, 1991). The object of measurement (person) and each facet represented a potential source of error in the ratings. Teacher raters and rating occasions were considered random factors because each was "much smaller than the size of the universe" and "exchangeable with any other samples of the same size drawn from the universe" (Shavelson & Webb, 1991, p. 11).

By using analysis of variance (i.e., GENOVA), variance components were estimated for the object of measurement (i.e., persons), each facet (i.e., teacher raters, rating occasions, and texts), and all possible two and three-way interactions (Crick & Brennen, 1983). The resultant variance components provided an estimate of the relative contribution of each of the error sources and possible interactions to the total amount of error in scores obtained from the use of the WR^2.

Results

To examine first-grade teachers' abilities to use the WR^2 instrument to individually and independently collect and code oral reading errors during and after the administration of the one-minute fluency probe using two passages, a two-facet (rater x occasion) fully crossed Generalizability and Dependability study was used for making absolute and relative decisions. The results of the GENOVA (ANOVA) analyses for 10 first-grade teachers rating 10 first grade readers' one-minute fluency probes by reading two passages, *Lucky Day* and *Our Pond*, on two occasions are shown in Tables 2 and 3.

Table 2 contains the means and standard deviations of the ten first-grade teacher raters for the two one-minute fluency probes using two

Table 2. Means and standard deviations by DIBELS 8 passage.

	WCPM		Error total		Categories coded		Fluency	
	Mean	SDV	Mean	SDV	Mean	SDV	Mean	SDV
Passage: *Lucky Day*	33.5	25.7	5.19	2.68	2.26	.84	2.5	.75
Passage: *Our Pond*	40.5	29.4	4.48	2.60	1.83	.74	2.6	.68

Table 3. Two facet fully crossed G theory analysis of 10 first-grade teachers' abilities to reliably rate student oral reading errors using the WR[2].

Design Type 3: Two-facet fully crossed design, as in P * F1 * F2
Number of Persons ('P'): 10
Number of levels for Facet 1 ('F1' number of teacher raters): 10
Number of levels for Facet 2 ('F2' number of rating occasions): 2
One or more negative variance estimates have been set to zero

ANOVA

WCPM

	df	SS	MS	Variance	Proport.
P	9.000	1478.38	164.26	8.12	94.6
F1	9.000	9.40	10.44	6.60	.01
F2	1.000	24.36	24.36	2.25	2.6
P*F1	81.000	48.40	6.00	4.94	.01
P*F2	9.000	157.40	17.50	1.69	2.0
F1*F2	9.000	13.20	1.50	6.72	.10
P*F1*F2	81.000	46.80	6.00	5.82	.70

Error variances:		G/D-coefficients:	
Relative	Absolute	G	D
3.5	3.5	.99	.99

Error count

	df	SS	MS	Variance	Proport.
P	9.000	1083.9	120.4	4.81	62.9
F1	9.000	8.1	.9	2.39	.3
F2	1.000	25.2	25.2	2.80	.000
P*F1	81.000	34.0	.42	2.32	.000
P*F2	9.000	217.9	24.2	2.39	31.2
F1*F2	9.000	.7	.08	.00	.00
P*F1*F2	81.000	37.6	.46	4.23	5.5

Error variances:		G/D-coefficients:	
Relative	Absolute	G	D
.47	.47	.94	.94

Fluency expression rating

	df	SS	MS	Variance	Proport.
P	9.000	34.00	3.78	.14	25.9
F1	9.000	26.50	2.94	.13	24.8
F2	1.000	.50	.50	.00	.000
P*F1	81.000	17.00	.21	.03	4.8
P*F2	9.000	8.80	.98	.08	14.3
F1*F2	9.000	1.90	.21	.005	.90
P*F1*F2	81.000	12.80	.16	.16	29.3

Error variances:		G/D-coefficients:	
Relative	Absolute	G	D
.16	.31	.65	.48

(*Continued*)

Table 3. Continued.

Error category coding					
	df	SS	MS	Variance	Proport.
P	9.000	77.04	8.56	3.36	45.2
F1	9.000	1.55	.72	1.46	.000
F2	1.000	9.24	9.24	7.45	10.0
P*F1	81.000	15.50	.19	2.22	3.0
P*F2	9.000	16.21	1.80	1.66	22.3
F1*F2	9.000	.51	.06	.000	.000
P*F1*F2	81.000	12.54	.16	1.45	19.5
	Error variances:			G/D-coefficients:	
	Relative	Absolute		G	D
	.18	.18		.794	.794

benchmark/progress monitoring passages. An examination of Table 3 showed that a .99 reliability (G&D) was achieved by one teacher rater on one rating occasion for wcpm using the WR^2. For the total number of errors, Table 3 showed that a .94 reliability (G&D) was achieved by one teacher rater on one rating occasion using the WR^2. For fluency expression ratings, a .65 reliability (G only) was achieved by one teacher rater on one rating occasion using the WR$^{2.}$ Finally, for percentage of agreement for error categorization, a .79 reliability (G&D) was achieved by one teacher rater on one rating occasion using the WR^2. The results of the D study showed similar results for both the G & D studies except for the fluency expression ratings which was different, .65(G) vs. .48(D) (See Figures 2–5).

Figures 2–5 show visually the effect of adding raters and rating occasions. These figures demonstrate clearly that adding a second rating occasion typically did not increase rater reliability whereas adding a single additional rater markedly increased reliability.

Teacher comments provided in the comment boxes of the scoring protocols consistently indicated that the WR^2 instrument professional development increased teachers' knowledge of and confidence in their abilities to record and categorize student oral reading errors using the WR^2 scoring form. An example drawn from these comments illustrated one first-grade teacher's increased knowledge and confidence as follows: "Using this process [the WR^2] sure helped me to be clearer about the accuracy, speed, expression, and types of decoding errors my students were making in their one-minute fluency probes. I could see how patterns could emerge after administering several one-minute fluency probes that would help me plan my Tier 2 small group instructional targets." Another teacher commented, "This really helped me to see kids' errors which I was not doing when I had previously administered one minute fluency probes." Another teacher commented, "I really liked the WR^2 instrument

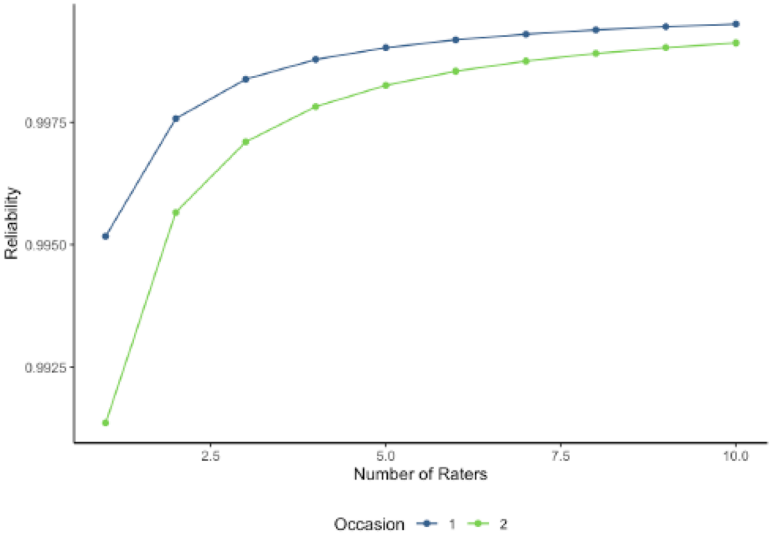

Figure 2. First-grade teacher rater reliability for WR^2 fluency accuracy and rate – Words correct per minute.

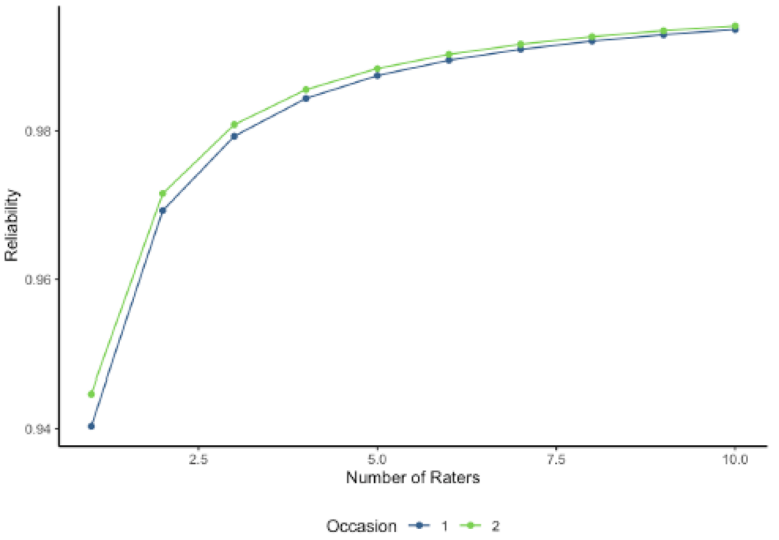

Figure 3. First-grade teacher rater reliability for WR^2 errors – Total errors.

because it made it easier for me to follow a system for recording and categorizing kid's reading errors to make decisions about my teaching." One teacher did note in her comments, "I found it difficult to make fine judgments about the fluency ratings. Maybe with more practice, I could get better."

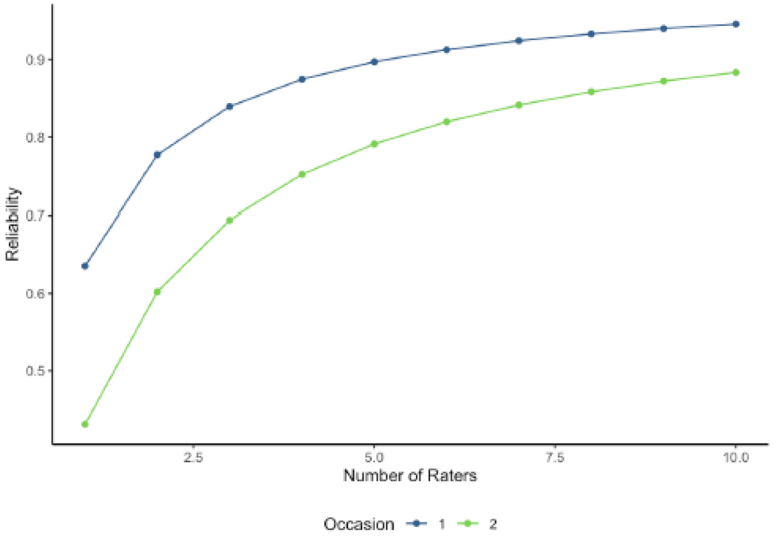

Figure 4. First-grade teacher rater reliability for WR^2 fluency – Expression ratings.

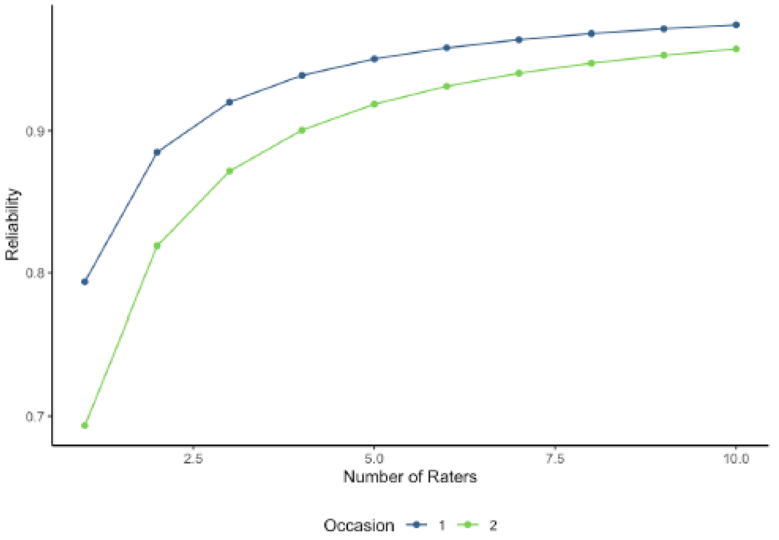

Figure 5. First-grade teacher rater reliability for WR^2 errors – Categories.

These comments seemed to reflect a new level of teachers' confidence in their ability to learn more from one-minute fluency probes than whether a student was red, yellow, green, blue, or reading a certain number of words correct per minute. Two of the 1st grade teachers who had used running records in their past noted in emails to the principal investigator that the WR^2 was easier to place word recognition errors into

word recognition categories as compared with the MSV coding system they had used in the past when taking and analyzing running records. The trained teachers in this study are now asking for their districts and schools to receive the WR^2 training and use the WR^2 oral reading error analysis form as well.

Discussion

This study had two primary purposes. First, we described the development of the Word Recognition Record (WR^2) error analysis framework, designed to augment one-minute fluency probes by providing a detailed method for examining student oral reading errors. Second, we investigated the reliability with which teachers could use the WR^2 framework to analyze multiple components of word recognition during students' one-minute fluency probes, using two DIBELS 8 ORF benchmark passages.

The development of the WR^2 was guided by research on the Science of Reading, particularly studies of word recognition processes, as well as the theoretical frameworks provided by Hoover and Gough (1990) expansion of Gough and Tunmer (1986) *Simple View of Reading* and Scarborough's *Reading Rope* (2001). The development team also drew on clinical observations of student oral reading errors obtained from routinely administering standard one-minute fluency probes. Although we recognized the potential value of capturing broader early reading skills—such as oral language structures, print concepts, and phonemic awareness—our review indicated that these could not be reliably inferred from the errors recorded during these brief oral reading probes.

Accordingly, the WR^2 was designed to focus specifically on word recognition data directly observable in a typical one-minute fluency probe. Once the framework's purpose, domains, specifications, directions, and forms were clearly defined, our next objective was to determine whether the WR2 was sufficiently structured to allow teachers to be trained to use it reliably.

Establishing teacher rater reliability on the WR^2 was a critical first step in evaluating the instrument's utility. Without demonstrated reliability, it would be impossible to determine the number of assessments required for results to generalize to other primary grade reading passages or to meaningfully analyze how error patterns might inform intervention selection and effectiveness.

The results of the Generalizability (G) and Dependability (D) study demonstrated reliability coefficients (G&D) ranging from .79 to .99, with one exception – fluency expression ratings, .65. The Generalizability and

Dependability study showed a need for only one rater on one rating occasion for making relative and absolute decisions as evidenced by obtaining a minimally acceptable reliability coefficient of .70 for three of the four measurements of oral reading errors using the WR^2.

The measurement of fluency expression required the addition of a single rater on a single rating occasion to achieve a .70 reliability coefficient or higher. When designing the WR^2 training module, the research team was aware of Zutell and Rasinski (1991) research findings showing that most teachers could be trained to rate expressive reading with a modicum of training. On the other hand, more recent research suggests that raters found expression and prosody more difficult to assess than rate and accuracy. Raters differed in severity and leniency, and phrasing was the prosodic subdimension most difficult to reliably assess (Arkan et al., 2022; Smith & Paige, 2019; Valencia et al., 2010). These studies showed that raters needed training using multiple texts and repeated practice to reduce disagreements.

These teachers had experienced a very brief training on the use of the *Multidimensional Fluency Scale* (MFS) in a district sponsored literacy professional development and during the training for using the WR^2. From the module trainings with fluency expression ratings showed the greatest variability and lowest percentages of agreement. This study's results seem to confirm that these 10 first-grade teacher raters may have benefited substantially from additional training emphasis on how to rate the expression dimensions of fluency.

These results further showed that first-grade teachers can reliably use the WR^2 administration, scoring, and analysis protocol as developed in this study for identifying rate, accuracy, and categorizing word recognition errors with adequate reliability. Additionally, these results indicated the cost of achieving a minimal .70 reliability coefficient using the WR^2 instrument to be quite reasonable, requiring only 1–2 teacher raters on one testing occasion for use within a classroom and school. It is also worth noting that a .70 coefficient obtained with only one rater occurs when a perfect agreement of oral reading error ratings is achieved by three of four raters, with a fourth person disagreeing by only one adjacent rating score. Consequently, an inter-rater agreement of .70 with only 1–2 raters would be viewed in the assessment scoring industry as a strong result (Brennan, 2001; Burdick et al., 2013). Although this study focused exclusively on first-grade students, the types of oral reading errors captured by the WR^2 are consistent with those observed in second- and third-grade students, suggesting that the findings may generalize to these later grade levels.

WR^2 oral reading error analysis allowed teachers to systematically identify instructional targets by analyzing recurring patterns in students' oral word recognition errors (See Appendix B for an example of a completed WR^2 oral reading error analysis). For example, two to three one-minute fluency probes may reveal consistent difficulties with r-controlled vowels or automaticity with grade-level high-frequency words, particularly those with irregular spellings. These findings can directly guide placement in Tier 2 intervention groups targeting specific word recognition deficits. Likewise, students with adequate accuracy and rate but poor prosody can receive targeted small-group or computer-assisted practice to improve expressive fluency, ensuring interventions are better aligned with observed performance.

Although some teachers had prior experience administering running records, their comments indicated that the WR^2 data collection, analysis, and categorization process was sufficiently clear and straightforward that even those without prior experience quickly learned to score word recognition errors reliably. The WR2 framework, viewed through a Science of Reading word recognition skill lens, has the potential to inform teachers' future decoding, word recognition, and fluency instruction (Foorman et al., 2016; Shanahan, 2021).

Primary grade teachers, as shown in this study, benefited from specific support and training to use an instrument such as the WR^2 to identify oral reading error patterns using a reading word recognition foundation skill lens (Gillett & Ellingson, 2017). DIBELS and other proprietary users of DIBELS like Acadience® and Amplify® may want to use the WR^2 or something like it as a formative assessment augmentation to help primary grade teachers gain more from the repeated administrations of one-minute fluency probes using DIBELS 8 passages than simply providing them a summary wcpm score, accuracy score, or colored risk level or benchmark report indicator. Similarly, core reading programs may also want to incorporate use of the WR^2 as an informal and formative assessment for helping teachers listen with greater purpose, intentionality, and insight to the texts they provide in their anthologies, and which are assigned to students to read. Teacher education programs may also want to once again train preservice teachers to record and analyze student oral reading errors by introducing them to the WR^2 (Gillett & Ellingson, 2017).

There is no doubt that the long-standing practice of listening intently and systematically to young children read text orally and noting their oral reading errors provides teachers a window into determining which word recognition skills students have clearly brought under their control and those that may need additional support and attention in instruction (Robinson, 1968). With careful analysis of oral reading errors using the WR^2 over time, teachers can look for consistent word recognition error patterns within known word recognition categories across multiple

administrations of one-minute fluency probes to make formative decisions about next instructional moves to assist students in the primary grades to become proficient in the vital and foundational skills of early word recognition. Our experiences as teacher educators also suggest that this is an area where considerable renewed attention may well be focused in anticipation of preparing teachers to teach word recognition skills and to assist preservice and inservice teachers in developing the insights and skills necessary to determine how they can support all young readers' with targeted word recognition instruction in their classrooms (Foorman et al., 2016; Parker et al., 1992).

Limitations

This study was limited in several ways. First, the school district in the study was not randomly selected leading to potential sampling selection bias. Although this district was purposively sampled to attract well-trained first grade teachers who understood the Science of Reading, the district may not be representative of teacher knowledge in other school districts. Second, with data from only 10 first-grade teacher raters, these results should be considered only preliminary. Finally, while the study demonstrated that teachers can reliability identify oral reading errors using the WR^2, it cannot be assumed that they can then translate these errors into effective instruction for helping students (Foorman et al., 2016).

Implications

There are several implications to be drawn from this study. The findings demonstrate that training in the use of the WR^2 resulted in increased accuracy, reliability, and teacher confidence in their abilities to record, analyze, and categorize oral reading errors using the WR^2 reading word recognition categories of phonics, high frequency words, and fluency (expression). The findings also suggest that primary grade teachers, when provided with appropriate training and tools like the $WR^{2,}$ can move beyond the summary data obtained from one-minute fluency probes to obtain deeper insights into students' word recognition challenges. The WR2 framework offers several potential benefits:

- Enhanced Diagnostic Information: It can be used to reveal specific error patterns over time and across passages that can inform targeted word recognition instruction.
- Professional Development: The WR^2 training module increased teacher confidence and interest in systematic error analysis, but it appears teachers will also need more training in rating students' oral reading expression.

- Potential for Broader Adoption: Beyond school and district use, the WR^2 could be incorporated into core reading programs and teacher preparation curricula as an informal and formative augmentation to one-minute fluency probes.

In conclusion, the WR^2 framework reestablishes the long-standing practice of listening closely to student oral reading and offers a systematic method to identify and address specific word recognition difficulties. As Robinson suggested many years ago (1968, p. 401), "a wealth of information about [reading] processes could be secured from carefully planned … examination of children's reading behavior." As we collaborated with these teachers, coaches, and other educators, we observed that some were not only listening to students read DIBELS ORF passages during one-minute fluency probes and calculating accuracy and words correct per minute (WCPM), but they were also informally analyzing these errors. Their *ad hoc* analyses appeared to draw on their newly acquired knowledge of the Science of Reading, word recognition skills, and foundational early reading skills.

Future research on the WR^2 should examine the reliability with which teachers can score commonly used classroom passages and texts. Additional studies of the WR^2 could investigate patterns of students' oral reading errors across multiple texts, as well as the number of texts required to establish these patterns with reliability. Finally, research is needed to determine whether teachers who systematically analyze students' word recognition errors using the WR^2 deliver more effective word recognition instruction or interventions than teachers who do not engage in such systematic error analysis.

Finally, the WR^2, used in conjunction with one-minute fluency probes, returns the long-standing practice of analyzing oral reading errors to its well-deserved historic place in the classroom, clinic, school, and laboratory assessment arsenal. Further, this research allowed teachers and coaches to access added value in their progress monitoring and benchmarking of students' reading using one-minute fluency probes. This added value was accomplished by augmenting one-minute fluency probes with the WR^2's systematic framework for analyzing oral reading errors that was informed by the Science of Reading using word recognition skill categories that hold the potential to inform primary-grade teachers' future word recognition instructional decisions.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

ORCID

D. Ray Reutzel http://orcid.org/0000-0002-9829-0944
Parker C. Fawson http://orcid.org/0000-0002-5417-6058

References

Afflerbach, P., Pearson, P. D., & Paris, S. G. (2008). Clarifying the differences between reading skills and reading strategies. *The Reading Teacher*, *61*(5), 364–373. https://doi.org/10.1598/RT.61.5.1

Amendum, S. J., Conradi, K., & Hiebert, E. H. (2018). Does text complexity matter in the elementary grades? A research synthesis of text difficulty and elementary students' reading fluency and comprehension. *Educational Psychology Review*, *30*(1), 121–151. https://doi.org/10.1007/s10648-017-9398-2

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P.D. Pearson, R. Barr, M.L. Kamil, & P Mosenthal (Eds.), *Handbook of reading research* (pp. 255–292). Longman.

Arkan, C. A., Uysal, P. K., Bilge, H., & Yildirim, K. (2022). Reliability of rating of Multidimensional Fluency Scale (MFS) with many-facets Rasch model. *International Journal of Assessment Toole in Education*, *9*(2), 470–491. https://doi.org/10.21449/ijate.974214

Barrett, C. A., Johnson, L. J., Truckenmiller, A. J., & VanDerHeyden, A. M. (2024). Comparing the cost-accuracy rations of multiple screening in elementary schools. *Remedial and Special Education*, *45*(2), 71–84. https://doi.org/10.1177/07419325231190809

Betts, E. A. (1936). *The prevention and correction of reading difficulties*. Row, Peterson & Company.

Betts, E. A. (1946). *Foundations of reading instruction: With emphasis on differentiated guidance*. American Book Company.

Brennan, R. L. (2001). *Generalizability theory*. Springer.

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice*, *14*(4), 9–12. https://doi.org/10.1111/j.1745-3992.1995.tb00882.x

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology*, *52*(1), 13–35. https://doi.org/10.1016/j.jsp.2013.11.008

Burdick, H., Swartz, C. W., Stenner, A. J., Fitzgerald, J., Burdick, D., & Hanlon, S. T. (2013). Measuring students' writing ability on a computer-analytic developmental scale: An exploratory study. *Literacy Research and Instruction*, *52*(4), 255–280. https://doi.org/10.1080/19388071.2013.812165

Burns, M. S., Griffin, P., & Snow, C. E. (1998). *Starting out right: A guide to promoting children's reading success*. National Academy of Sciences. Courtesy of the National Academy Press. Reprinted with permission.

Cattell, J. M. (1886). The time taken up by cerebral operations. *Mind*, *15*, 373–381.

Clay, M. M. (1972). *Reading: The patterning of complex behaviour*. Heinemann.

Clay, M. M. (1979). *The early detection of reading difficulties* (2nd ed.). Heinemann.

Clay, M. M. (1993). *An observation survey of early literacy achievement*. Heinemann.

Colenbrander, D., Kohnen, S., Beyersmann, E., Robidoux, S., Wegener, S., Arrow, T., Nation, K., & Castles, A. (2022). Teaching children to read irregular words:

A comparison of three instructional methods. *Scientific Studies of Reading*, *26*(6), 545–564. https://doi.org/10.1080/10888438.2022.2077653

Common Core State Standards. (2010). *National Governors Association Center for Best Practices and Council of Chief State School Officers*. https://www.thecorestandards.org

Crick, J. E., & Brennen, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system*. American College Testing Program.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt Brace Jovanovich.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. John Wiley.

D'Agostino, J. V., & Briggs, C. (2025). Validation analysis during the design stage of text leveling. *Education Sciences*, *15*(5), 607. https://doi.org/10.3390/educsci15050607

D'Agostino, J. V., Rodgers, E., Winkler, C., Johnson, T., & Berenbon, R. (2021). The generalizability of running record accuracy and self-correction scores. *Reading Psychology*, *42*(2), 111–130. https://doi.org/10.1080/02702711.2021.1880177

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*(3), 219–232. https://doi.org/10.1177/001440298505200303

DIBELS 8 ORF. (2022). *University of Oregon*. https://dibels.uoregon.edu

Ebel, R. L., & Fiesbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice Hall.

Ekwall, E. E. (1976). Informal reading inventories: The instructional level. *The Reading Teacher*, *29*(7), 662–665.

Fawson, P. C., Reutzel, D. R., Smith, J. A., Ludlow, B. C., & Sudweeks, R. (2006). Examining the reliability of running records: Attaining generalizable results. *Journal of Educational Research*, *100*(2), 113–126. https://doi.org/10.3200/JOER.100.2.113-126

Foorman, B., Beyler, N., Borradaile, K., Coyne, M., Denton, C. A., Dimino, J., Furgeson, J., Hayes, L., Henke, J., Justice, L., Keating, B., Lewis, W., Sattar, S., Streke, A., Wagner, R., & Wissel, S. (2016). *Foundational skills to support reading for understanding in kindergarten through 3rd grade (NCEE 2016-4008)*. National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. http://whatworks.ed.gov

Fountas, I. C., & Pinnell, G. S. (1996). *Guided reading: Good first teaching for all children*. Heineman.

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

Gillett, E., & Ellingson, S. P. (2017). How will I know what my students need? Preparing preservice teachers to use running records to make instructional decisions. *The Reading Teacher*, *71*(2), 135–143. https://doi.org/10.1002/trtr.1609

Goodman, K. S. (1965). A linguistic study of cues and miscues in reading. *Elementary English*, *43*(6), 639–643.

Goodman, K. S. (1969). An analysis of oral reading miscues: Applied psycholinguistics. *Reading Research Quarterly*, *5*(1), 9–30. https://doi.org/10.2307/747158

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*(1), 6–10. https://doi.org/10.1177/074193258600700104

Greene, T. (2024). From policy to action: Why 8 states banned three-cueing from K–3 reading instruction. *ExcelinEd in Action*. https://excelinedinaction.org/2024/01/10/from-policy-to-action-why-8-states-banned-three-cueing-from-k-3-reading-instruction/

Halliday, J. L. (2012). Revisiting key assumptions of the reading level framework. *The Reading Teacher*, *66*(1), 53–62.

Hasbrouck, J., & Tindal, G. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher*, *59*(7), 636–644. https://doi.org/10.1598/RT.59.7.3

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

Henk, W. A., McKenna, M. C., & Conradi, K. (2011). Developing affective instrumentation. In N.K. Duke & M. H. Mallette (Eds.), *Literacy research methodologies* (2nd ed., pp. 242–269). Guilford Press.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127–160. https://doi.org/10.1007/BF00401799

Hosp, M. K., & Hosp, J. L. (2003). Curriculum-based measurement for reading, spelling, and math: How to do it and why. *Preventing School Failure: Alternative Education for Children and Youth*, *48*(1), 10–17. https://doi.org/10.1080/1045988X.2003.10871065

Huey, E. B. (1908). *The psychology and pedagogy of reading*. MIT Press.

Johnson, M. S., & Kress, R. A. (1965). *Informal reading inventories*. International Reading Association.

Kaminski, R. A. (1992). *Assessment for the primary prevention of early academic problems: Utility of curriculum-based measurement prereading tasks* [Doctoral Dissertation]. University of Oregon.

Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, *95*(1), 3–21. https://doi.org/10.1037/0022-0663.95.1.3

Mesmer, H. A. (2024). *Big words for young readers: Teaching kids in grades K to 5 to decode and understand – Words with multiple syllables and morphemes*. Scholastic.

Moats, L. C., & Tolman, C. A. (2019). *Language essentials for teachers of reading and spelling (LETRS)* (Vol. 1 and 2). Voyager Sopris Learning.

Mooney, M. (1990). *Reading to, with and by children*. Richard C. Owen Publishers.

National Council on Teacher Quality (NCTQ). (2023a). *Reviews of reading instructional materials used by teacher preparation programs*. https://www.nctq.org/review/readingTextbooks

National Council on Teacher Quality (NCTQ). (2023b). *The four pillars to reading success*. https://teacherquality.nctq.org/dmsView/the_four_pillars_to_reading_success

National Early Literacy Panel. Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, DHHS. (2010). *Developing early literacy: Report of the National Early Literacy Panel (Reference Only) (NA)*. U.S. Government Printing Office.

National Reading Panel. (2000). Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, DHHS. *Report of the National Reading Panel: Teaching Children to Read: Reports of the Subgroups (Reference Only) (00-4754)*. U.S. Government Printing Office.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Sage.
Osterlind, S. J. (1989). *Constructing test items*. Kluwer.
Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly*, *40*(2), 184–202. https://doi.org/10.1598/RRQ.40.2.3
Parker, R., Hasbrouck, J. E., & Tindal, G. (1992). Greater validity for oral reading fluency: Can miscues help? *Journal of Special Education*, *25*(4), 492–503. https://doi.org/10.1177/002246699202500406
Pressley, M. (2000). What should comprehension instruction be the instruction of? In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. III, pp. 545–561). Lawrence Erlbaum Associates.
Pressley, M., & McCormick, C. B. (1995). *Advanced educational psychology for educators, researchers, and policymakers* (p. 528). Harper Collins.
Rasinski, T. V., Reutzel, D. R., Chard, D., & Linan-Thompson, S. (2011). Reading fluency. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research* (Vol. 4, pp. 286–319). Routledge.
Raygor, A. L. (1977). The Raygor readability estimate: A quick and easy way to determine difficulty. In P. D. Pearson (Eds.), *Reading: Theory, research and practice* (pp. 259–263). Twenty-Sixth Yearbook of the National Reading Conference: NRC.
Reading League. (2022). *Science of reading: Defining guide [eBook]*. https://www.thereadingleague.org/wp-content/uploads/2022/01/Science_of_Reading_Defining_Guide_eBook.pdf
Robinson, H. M. (1968). The next decade. In H. M. Robinson (Ed.), *Innovation and change in reading instruction* (pp. 401). Yearbook of the National Society for the Study of Education.
Ross, J. A. (2004). Effects of running records assessment on early literacy achievement. *Journal of Educational Research*, *97*(4), 186–195. https://doi.org/10.3200/JOER.97.4.186-195
Samuels, S. J. (2007). The DIBELS tests: Is speed of barking at print what we mean by reading fluency? *Reading Research Quarterly*, *42*(4), 563–566. https://doi.org/10.1598/RRQ.42.4.7
Scarborough, H. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory and practice. In Neuman S., & Dickinson D. K. (Eds.), *Handbook of early childhood literacy research* (Vol. 1, pp. 97–110). Guilford Press.
Seidenberg, M. (2017). *Language at the speed of sight: How we read, why so many can't and what can be done about it*. Basic Books.
Shanahan, T. (2009). *The problem with guided reading*. Shanahan on Literacy. https://www.shanahanonliteracy.com/blog/the-problem-with-guided-reading?utm_source=chatgpt.com
Shanahan, T. (2011). *Rejecting instructional level theory*. Shanahan on Literacy. https://www.shanahanonliteracy.com/blog/rejecting-instructional-level-theory
Shanahan, T. (2021). *3P versus 3-cueing: Why recommend the one and shun the other*. Shanahan on Literacy. https://www.shanahanonliteracy.com/blog/3p-versus-3-cueing-why-recommend-one-and-shun-the-other
Shanahan, T. (2024). *Three-cueing and the law*. Shanahan on Literacy. https://www.shanahanonliteracy.com/blog/three-cueing-and-the-law?utm_source=chatgpt.com
Shanahan, T. (2025). *Science of reading levels*. Shanahan it on Literacy. https://www.shanahanonliteracy.com/upload/publications/272/pdf/reading-levels-1.pdf

Shapiro, L. R., & Solity, J. (2016). Differing effects of two synthetic phonics programmes on early reading development. *British Journal of Educational Psychology*, *86*(2), 182–203. https://doi.org/10.1111/bjep.12097
Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.
Smith, G., & Paige, D. D. (2019). A study of interrater reliability: A study of reliability across multiple raters when using the NAEP and MDFS rubrics to measure oral reading fluency. *Reading Psychology*, *40*(1), 34–69. https://doi.org/10.1080/02702711.2018.1555361
Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. Guilford Press.
Stouffer, J. (2021). Seeking middle ground: Analyzing running records from top to bottom. *Reading Teacher*, *74*(6), 769–784. https://doi.org/10.1002/trtr.2012
Strong, J. Z., Amendum, S. J., & Conradi Smith, K. (2018). Supporting elementary students' reading of difficult texts. *Reading Teacher*, *72*(2), 201–212. https://doi.org/10.1002/trtr.1702
Strube, M. J. (2000). Reliability and generalizability theory. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 23–66). American Psychological Association. https://doi.org/10.1037/10317-001
Thorndike, E. L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational Psychology*, *8*(6), 323–332. https://doi.org/10.1037/h0075325
Thorndike, R. L. (1982). *Applied psychometrics*. Houghton Mifflin.
Valencia, S. W., Smith, A. T., Reece, A. M., Li, M., Wixson, K. K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly*, *45*(3), 270–291. https://doi.org/10.1598/RRQ.45.3.1
Walpole, S., Strong, J. Z., & Vitale, M. N. (2021). Instrument development. In M. H. Mallette & N. K. Duke (Eds.), *Literacy research methodologies* (3rd ed., pp. 211–233). Guilford Press.
Willingham, D. T. (2017). *The reading mind: A cognitive approach to understanding how the mind reads*. Jossey-Bass.
Wundt, W. (1908–1911). *Principles of physiological psychology.* (6th ed., Vols. 1-3).
Zutell, J., & Rasinski, T. V. (1991). Training teachers to attend to their students' oral reading fluency. *Theory into Practice*, *30*(3), 211–217. https://doi.org/10.1080/00405849109543502

Appendix A

Word Recognition Record (WR²) Administration & Scoring Process

Preparing to Administer the WR²

Step 1: Obtain copies of grade level passages. DIBELS 8th Edition ORAL READING FLUENCY Benchmarking and Progress Monitoring Passages are available at the following website address: https://dibels.uoregon.edu/materials/dibels
Step 2: Make two paper copies of the selected passage: one for you to mark up, one for the student to read.
Step 3: Obtain a stopwatch or timer for passage administration.
Step 4: Arrange a quiet space with a small desk or table for a student and you, the

examiner, to place the passage copy to be read in front of the student.
Step 5: Bring the student to this quiet area and seat them comfortably facing the one paper copy of the passage.
Step 6: Put your paper copy on a clipboard for marking the passage. Place your stopwatch or timer where you can see it but not the student.
Step 7: Give the following directions to the student prior to beginning the one-minute timing and marking of the DIBELS 8 ORF benchmarking or progress monitoring passage selected:

Please read this passage out loud (point to the passage).
If you get stuck, try to blend it or say it. When I say "stop," you can stop reading. So do your best reading. Start here (point to the first word of the first paragraph of the passage.)

Step 8: Start your stopwatch when the student says the first word and begin making a record of the student's reading of the passage on the examiner copy. If a student hesitates for 3 seconds, give the correct word, and mark the errors as shown below. Use the following marking system to record student deviations (errors) from the text:

- **Teacher Prompt: mark a TP on the word given.**
- **Omission: circle the omitted word**
- **Substitution: Write the word above the word substituted.**
- **Insertion: Carrot mark ^ and write the word inserted.**
- **Self-Correction: SC over the word they corrected.**
- **Inflectional endings: If the student does not say the ending, cross off the ending they omitted.**

Step 9: Stop the timer/stopwatch and say, "Stop!" Mark a diagonal or vertical line at the point where the student stopped reading on the paper copy of the passage read. Count the number of total words read, total errors marked, and total words correct (subtract errors from total words read) and put these scores in the WCPM column and tally totals at the bottom of the *WR²* Analysis Form.

Step 10: Pull out one copy of the *WR²* Analysis Form for error analysis. Analyze the errors made by first placing them into the categories represented on the *WR²* Analysis Form. For example, if a student omits, substitutes, inserts, or is given a word, note the type of error and the actual error on the *WR²* Analysis Form.

REPEAT THIS PROCESS WITH 2 OR MORE DIBELS 8 ORF BENCHMARK/ PROGRESS MONITORING PASSAGES to OBTAIN DATA FOR ESTABLISHING POTENTIAL ORAL READING ERROR PATTERNS.

Appendix B

*Word Recognition Record (WR²) **

Title of Passage: _____ **Student's Name:** _____
Examiner Name: _____ **Date:** _____

Phonics: Single syllable patterns	Phonics: Multi-syllable words	High frequency words		Fluency	
Single syllable Tally Total:	Multiple syllable Tally Total:	High frequency words Tally Total:		Fluency Average of Ratings:	
		D	**IS**	**Rate WCPM**	**Expression** **V** **I** **P**

Phonics
Single Syllable Patterns
CV
CVC
CVVC
CVCe
CVr

Phonics
Multiple Syllables
Number of Syllables: 2, 3, 4, etc.
Type of M.S. Words
Compound
Base/Affix
Roots
True M.S. Word
Single Syllable Pattern Combinations
Compound words
CVC/CVC
CVC/CVVC
CVVC/CVC
CVC/CVCe
CVC/CVr
CVC/Cle

High Frequency Words
D – Decodable
IS – Irregularly Spelled (Sight)

Expression
V – Volume
I – Inflection
P – Phrasing

*High tally scores (more than 3) represent SOR word recognition elements potentially needing additional or scaffolded instruction.

Example of a Completed WR²

Title of Passage: Our Pond Student's Name: HVE4.2.OP.m49
Examiner Name: _____ Date: _____

Phonics: Single Syllable	Phonics: Multi Syllable	High Frequency Words		Fluency	
		D	IS	Rate	Expression
lots/lost CVCC			that/this	WCPM	V 4
TP build CVVCC			are/is	39	
					I 2
					P 2

Phonics
Single Syllable Patterns
CV
CVC
CVVC
CVCe
CVr

Phonics
Multiple Syllables
Number of Syllables:
2, 3, 4, etc.
Type of M.S. Words
Compound
Base/Affix
Roots
True M.S. Word
Single Syllable Pattern Combinations
Compound words
CVC/CVC
CVC/CVVC
CVVC/CVC
CVC/CVCe
CVC/CVr
CVC/Cle

High Frequency Words
D - Decodable
IS – Irregularly Spelled (Sight)

Fluency
WCPM – Use norm tables

Expression* Scale 1 Poor – 4 Excellent
V – Volume
I – Inflection
P – Phrasing

Single Syllable	Multiple Syllable	High Frequency Words	Fluency
Tally Total: 11	Tally Total:	Tally Total: 11	2.7 Average of Ratings:

*High tally scores (more than 3) represent SOR foundational elements potentially needing additional or scaffolded instruction.