Reading and Writing
https://doi.org/10.1007/s11145-020-10073-x

Initial validation of a measure of decoding difficulty as a unique predictor of miscues and passage reading fluency

Neena M. Saha[1] · Laurie E. Cutting[1] · Stephanie Del Tufo[2] · Stephen Bailey[3]

Abstract

Quantifying the decoding difficulty (i.e., 'decodability') of text is important for accurately matching young readers to appropriate text and scaffolding reading development. Since no easily accessible, quantitative, word-level metric of decodability exists, we developed a decoding measure (DM) that can be calculated via a web-based scoring application that takes into account sub-lexical components (e.g. orthographic complexity), thus measuring decodability at the grapheme-phoneme level, which can be used to judge decodability of individual words or passages. Here we report three experiments using the DM: two predicting children's word-level errors and one predicting passage reading fluency. Generalized linear mixed effect models showed that metrics from the DM explained unique variance in children's oral reading miscues after controlling for word frequency in two samples of children (experiments 1 and 2), and that more errors were made on words with higher DM scores for poor readers. Furthermore, the DM metrics predicted children's number of words read correctly per minute after accounting for estimated Lexile passage scores in a third sample (experiment 3). These results show that after controlling for word frequency (experiments 1 and 2) and estimated Lexile scores (experiment 3) the model including the DM metrics was significantly better in predicting children's word reading fluency both for individual words and passages. While further refinement of this DM measure is ongoing, it appears to be a promising new measure of decodability at both the word and passage level. The measure also provides the opportunity to enable precision teaching techniques, as grapheme-phoneme correspondence profiles unique to each child could facilitate individualized instruction, and text.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11145-020-10073-x) contains supplementary material, which is available to authorized users.

✉ Laurie E. Cutting
Laurie.Cutting@vanderbilt.edu

Extended author information available on the last page of the article

Published online: 06 August 2020

Keywords Decoding · Decoding measure · Reading errors · Text leveling · Validation evidence

Skilled reading is of critical importance to function in today's world, and, is associated with a number of positive academic and social outcomes (Kern & Friedman, 2009; McLaughlin, Speirs, & Shenassa, 2014). Given the numerous positive outcomes associated with adequate reading skill, it is unsettling that national reading assessments demonstrate that upwards of 60% of children in the United States are not reading at grade-level (U.S. Department of Education, 2017). Due to the longstanding concern in the United States regarding widespread reading failure, in 1998, the National Reading Panel was mandated by Congress to determine the research-based components of high-quality reading instruction. Findings from the National Reading Panel (2000) identified five major areas of reading instruction: phonemic awareness, phonics, vocabulary, fluency, and reading comprehension. Thus, the use of direct, systematic, explicit, code-based approaches to enhance reading outcomes (phonics) was identified as a key component of effective reading instruction, and one that is particularly important for beginning readers. Phonics-based approaches help children decipher the alphabetic code through instruction of the grapheme-phoneme correspondences (GPCs) in the English language. The literature supports phonics instruction as a necessary (but not sufficient) component of reading instruction. Numerous studies show that children who have proficient decoding skills are able to use their knowledge of GPCs to read novel words, which subsequently influences their reading outcomes, including reading fluency and reading comprehension (e.g., Eldredge, 2005; Kuhn & Stahl, 2003). Explicit and systematic phonics instruction teaches children GPCs in a direct and structured manner, and typically includes intensive practice of learned GPCs with 'decodable' text, or *highly-controlled levelled text intended for beginning readers*.

More specifically, research examining exposure and interaction with decodable text has shown that practicing newly learned GPCs in context builds decoding skill, likely because it allows children to solidify their knowledge of them (e.g., Juel & Roper-Schneider, 1985; Mesmer, 2008). Indeed, it has been noted that students who practice the GPCs that they have learned within the context of decodable text apply more letter-sound knowledge strategies when reading other non-controlled texts (Mesmer, 2008). Other findings suggest that decodable text directly supports the development of phonemic awareness, phonics, and fluency (e.g., Cheatham & Allor, 2012), three of the five key areas that the National Reading Panel (2000) identified that should be included in a reading program. However, a recent review found that as text difficulty increases, children's reading fluency decreased (Amendum, Conradi & Hiebert, 2018). This study's findings suggest that children need to practice GPCs in controlled texts that are adequately matched to their decoding level. Selecting the appropriate text level is perhaps most important for beginning or 'emerging' readers. Appropriate levels of accuracy for reading words aloud (typically ~ 95%) are optimal for building key reading skills such as fluency and comprehension. Optimal levels of reading difficulty

are also important for viewing reading as a fun and engaging activity (i.e., "Matthew Effects"; Stanovich, 2009), which increases motivation to continue reading. Increased reading means more exposure to texts and vocabulary building, which aids in comprehension, the ultimate goal of reading (e.g., Hoover & Tunmer, 2018).

In response to research citing the importance of decodable text as a needed component of reading instruction, several state education agencies, such as those in Texas and California, mandated the use of decodable text in classroom instruction in 2000 (Hoffman, Sailors & Patterson, 2002). Publishers responded to this mandate by attempting to provide decodable text in their reading programs (Hoffman et al., 2002). In addition to being incorporated in several basal reading programs, decodable text has also been employed in reading intervention programs (e.g., *Imagine it!*). However, currently, there is no standard way of determining decoding level of texts, and publishers and researchers use various methods when attempting to create decodable texts. Current ways used to determine how decodable a text is either rely on more subjective methods, such as lesson-to-text matching and qualitative leveling, or more quantitative leveling methods, such as readability formulas. Nonetheless, as we will elaborate more in the subsequent sections, none satisfactorily capture, in an objective and quantitative manner, a way to pinpoint the decoding difficulty of a word or a passage.

A Review of Current Metrics Used to Level Texts

Lesson-to-Text Matching

One method of creating decodable text is based on a student's "potential for accuracy," or, the probability that a child will read a given word correctly. Potential for accuracy is operationally defined as the match between the teacher edition and the text in students' books at a given point in the curriculum. The assumption is that if certain GPCs have been taught, then students will have greater probability of accuracy when they encounter it in text. Others have echoed this definition of decodability: Stein, Johnson, and Gutlohn (1999, p. 277) refer to decodable text as "… connected text containing a high percentage of words conforming to the letter-sound correspondences that have previously been introduced."

For many decades, researchers have used the 'lesson-to-text' definition of decodable text when assessing decodability of texts for beginning readers (Beck & McCaslin, 1978; Chall & Read, 1967; Meyer, Greer, & Crummey, 1987). For example, in the 1960s and 1970s, researchers examined how popular basal reading series aligned with teacher's introduction of GPCs, and found that the texts that were used were not aligned with the teachers' introduction of GPCs (Beck & McCaslin, 1978; Chall & Read, 1967). Furthermore, Meyer, Greer, and Crummey (1987) examined four major basal programs in the 1980s and concluded that three of the four had less than 10% decodable text (i.e., did not align with the GPCs that had been introduced). In 1999, Stein, Johnson, and Gutlohn analyzed several commercially published first-grade basal reading programs and found that six of the seven programs had little or

no relationship to the sequence presented in the teacher guides. In those programs, only 15% of the text was decodable according to the lesson-to-text matching definition of decodability. More recently, Maslin (2007) analyzed decodability, again defined using lesson-to-text matching, using texts from Harcourt, Houghton Mifflin, McMillan, McGraw-Hill, Open Court and Scott Foresman reading programs. Their results showed that 16 to 25% of the text was not decodable. Maslin concluded that this was well within the frustration level for children, and there was much room for improvement. Thus, use of lesson-to-text matching, while perhaps a useful way of assessing the decodability of a text, thus far has resulted in minimal alignment between the GPCs taught and the accompanying texts. However, it is important to note that even if publishers were more appropriately aligning their texts with the sequence of GPCs being taught, the validity of this method has not been established. For example, the exact percentage of decodable words a text must contain in order to be deemed 'decodable' is not clear (Cheatham & Allor, 2012; Menton & Hiebert, 1999). Furthermore, there is no research that clearly defines how many exposures or 'lessons' a child requires before a particular GPC becomes decodable.

Qualitative Leveling

Qualitative text leveling systems are often referred to as 'holistic text leveling' because they use a variety of subjective factors to gauge the difficulty of a text. Qualitative leveling has a history that goes back as far as the mid-1800s when the first graded school opened in Boston replete with sets of books that were specifically prepared for each grade (DuBay, 2007). From early on, the consideration of 'decoding difficulty' within qualitative text leveling has not really been aligned with the principles of phonics and explicit GPC instruction, but rather focuses more generally on being able to recognize the words, versus applying systematic knowledge of GPCs to decode them. This is an important distinction, as children may be able to learn to recognize words by sight, particularly in the early grades, but lack the ability to apply GPC knowledge to decode novel words. Decades of research has determined that teaching children word recognition (i.e., by memorizing words) does *not* generally transfer to knowing how to decode words (e.g., Adams, 1994). Qualitative leveling approaches take a variety of different factors into account to determine the difficulty level of a book, such as choosing specific vocabulary words and emphasizing repetition and repeated exposure to the same types of words; however, these approaches are capturing word recognition, versus systematically focusing on decoding difficulty/GPCs. Furthermore, specific information as to why a certain text is assigned a given level is not available (even from the publishers) for both Fountas & Pinnell and Reading Recovery, two of the most prominent qualitative leveling approaches. For example, Fountas & Pinnell's (1999) guided reading levels consider: (1) book and print features, (2) content, themes, and ideas, (3) text structure, and (4) language and literacy elements, but there is no evidence that these variables increase by the same amount from level to level. And, despite widespread use by textbook publishers, there are not any reports of interrater reliability for coders scoring text for either Fountas & Pinnell's Guided Reading Levels or Reading Recovery

(Pearson & Hiebert, 2014). Other criticisms of this approach have also been raised. For example, Hiebert (2002) stated that even experienced teachers found the Fountas & Pinnell leveling criteria too ambiguous to apply, resulting in lack of fidelity of procedures. Similarly, Brabham and Villaume (2002) noted teacher and student frustrations as children tried to apply letter-sound relationships in text that was supposed to be predictable, yet, in fact, contained several irregular words, and Fawson and Reutzel (2000) found that basal texts sequenced using the Fountas & Pinnell guidelines did not end up ordering correctly (from simple to more complex). Finally, Pearson and Hiebert (2014, p. 7) state, "we could find no studies that examined how instruction with texts ordered according to either Reading Recovery or guided reading levels influenced reading acquisition." In summary, despite the popularity of qualitative text leveling programs, the validity of predicting text difficulty for developing readers is has not been established (Hatcher, 2000; Hiebert, 2002; Hoffman, Roser, Patterson, Salas, & Pennington, 2000).

Quantitative Systems: Readability Formulas

Unlike lesson-to-text matching and qualitatively based readability formulas, quantitative readability formulas incorporate objective variables. Readability formulas generally consist of linguistic variables such as the number of syllables in a word or the number of words in a sentence that are given different weights depending on the formula. Readability formulas were first created in the 1920s, a period marked by the enthusiastic application of science to the field of education (Fry, 2002). Their popularity grew after several psychologists and journalists noted that readership increased as readability of magazines and newspapers was simplified to match the average difficulty of American adult readers. While these efforts were directed at adult readers, the first readability formulas were actually created for children. DuBay (2007) notes that the increasing population of immigrant children that found textbooks too difficult was a major impetus in the development of readability formulas. Currently, about 89% of textbook publishers use readability formulas to level their texts, and textbook publishers rate leveling as the most important factor in creating textbooks (Brabham & Villaume, 2002; DuBay, 2007). Begeny and Greene (2014) give several examples of how publishers incorporate readability formulas when planning curriculum materials: *Imagine it!* (2011) applies the Flesch-Kincaid readability estimate, *Read Well* (Sprick, Howard, & Fidanque, 1998) uses both the Spache and Dale-Chall formulas, and *Reading Mastery* (2011), an intervention program, uses Lexile rankings to compare text. Briefly, Lexile scores for texts incorporate variables such as word length, and word frequency and represent the percent of text that a given child (whose reading ability has been tested and linked using the same scale) can comprehend (e.g., MetaMetrics, 2007, p. 4–5). Readability formulas have even been used to equate curriculum-based measurement (CBM) probes (Francis et al., 2008). Readability formulas are popular because they are often free, easy to implement, and not time-intensive. However, while they may capture some elements that align with decodability (or, more likely, word recognition), their use for creating decodable text has not been thoroughly vetted.

Indeed, several criticisms of readability formulas have been raised in the literature. First, the creators of readability formulas often examine their validity by comparing them with the 'gold standard' readability formulas at the time of their creation. However, this type of criterion validity is flawed if the gold standard at the time does not itself have strong evidence of validity. For example, Olson (1984) examined the validation efforts of six popular readability formulas and found that several were criterion-validated using student scores on passages from the McCall-Crabbs Standard Test Reading Lessons. However, the creators of the McCall-Crabbs lessons state that they were only created as practice exercises in reading, and never as measures of text comprehensibility (Bruce, Rubin, & Starr, 1981).

Others have criticized readability formulas because they do not include all the psychological factors known to contribute to reading (Bruce & Rubin, 1988; Davison & Kantor, 1982). Even readability formulas designed for beginning readers (such as Spache) only use two variables: sentence length and word frequency. Furthermore, Bailin and Grafstein (2001) critiqued the linguistic assumptions that readability formulas are based on, noting that word length is a poor proxy for word difficulty, as longer words are usually the result of affixations that carry meaning, and children know what they mean.

Central to the need for a metric that captures decoding difficulty, readability formulas have also been criticized as not being sensitive to the needs of beginning readers in the primary grades, where fine-grained distinctions (subtler than grade level) are required to match readers to appropriate text (Cunningham et al., 2005; Fry, 2002; Hoffman et al., 2001). For example, Hiebert and Pearson (2010) compared Degrees of Reading Power, Fry, Spache, Lexile, and Coh-metrix formulas across seven types of kindergarten through second grade texts. They found that the formulas were all consistent with the rankings of several categories of texts (i.e., "Trade", "Trade- Instructional", "Textbook Core- Current" etc.), except in the category of decodable texts ("Text Ancillary- Decodable"), which varied widely across indices, further bolstering the argument that they are not accurate for emerging reader texts.

Summary

In summary, of the three different major approaches taken to level texts, the only one that directly considers decodability is lesson-to-text matching, in that it attempts to explicitly align decodable texts with GPCs. However, this method offers no specific metric for text difficulty, or manner to level texts, which is problematic in terms of being able to guide teachers in selecting appropriately leveled books across children's literature. However, the other methods ('holistic' qualitative and quantitative leveling) are also problematic. While qualitative methods are potentially more sensitive to the needs of emerging readers and quantitative readability formulas are generally low-cost, fast, and easy to implement, both are limited in terms of capturing decoding difficulty, because neither accounts for features at the letter-sound level (GPCs), and instead use approaches that may index more global features of words. Accounting for sub-lexical components of words (rather than whole word-level or sentence-level) is important, given research showing that text levelling systems that

incorporated sub-lexical components, such as the number of syllables in a word, appear to be more strongly correlated with children's oral reading performance, or reading fluency (Saha & Cutting, 2019) (Note that syllables can contain several GPCs, so, while they are sub-lexical- they are not as fine-grained as the GPC level.). These findings are consistent with Harris & Jacobson's (1979, p. 395) observation almost 40 years ago that "refinement of the GPC variables in readability formulas research" was needed. Yet, despite this decades-old observation coupled with the overwhelming support for needing to use decodable texts in reading instruction, to date there is no method available that quantifies text decodability in an automated manner that accounts for features at the letter-sound level (GPCs); such a metric is important for accurately matching children to appropriate text. This is especially critical for emerging readers, since accurate word recognition by applying GPCs is the primary goal of emerging readers using decodable texts. To address this limitation in the literature, we sought to create a measure of decodability to help match emerging readers to appropriate text. Prior to embarking on this task, we outlined two stages that would be required in order to provide *initial* validation for such a measure. First, it would need to take into account emerging readers' sensitivity to different types of sub-lexical information by accounting for different elements that have been theoretically (Adams, 1990; Ehri, 1995) and empirically (e.g., Balota et al., 2007; Steacy et al., 2019; Treiman, Kessler, Zevin, Bick, & Davis, 2006, Treiman, 1991; Wimmer & Goswami, 1994) linked to developing proficient GPC knowledge. Second, it would need to be 'real-world' validated, showing that it actually accounted for variance in predicting children's accuracy at pronouncing words. Stage 1 of this work required examining the theoretical and empirical literature in order to determine a way to systematically measure the sub-lexical features of words. Below we outline each component of the measure, which we colloquially named the Decoding Measure (or, DM; Saha, Bailey, & Cutting, 2017, Cutting, Saha, & Hasselbring, 2017), that resulted in a metric that allowed us to proceed to the initial validation stage (stage 2). It is important to note that we are not claiming that our measure is one that is exhaustively refined. We expect that any metric developed will need further iterative refinement and validation and that these two stages are the first of many before we reach the end goal of providing the field with an objective, automated manner of capturing decoding difficulty.

The Components of the DM

The DM is a quantitative measure of word-level decoding difficulty that incorporates metrics of how frequent a word is, along with specific metrics of sublexical elements of decoding difficulty. These metrics were determined by surveying the literature for aspects of GPCs that are challenging for beginning readers, as well as those with dyslexia (a reading disability which is specifically characterized by difficulty in learning GPCs). Notably, these are the two populations for which having a quantitative metric of decoding difficulty would be of most use. Areas discerned to be critical elements of decoding difficulty include degree of phoneme-grapheme discrepancy; how common (or rare) the vowel and consonant sounds in the word

are; and, the number of blends in a word. Each of the aforementioned variables are reported in previous research as important predictors of visual word recognition accuracy and reaction time. Importantly, all but the first component is focused on sub-lexical components of words at the GPC level. Below we review the rationale for each of these components of the DM.

Word Frequency

Word frequency has long been known to be a critical piece of recognizing words (e.g., Balota et al., 2007; Joseph, Nation, & Liversedge, 2013). Specifically, frequent words are identified and processed more easily than less frequent words. This effect is present in both adults (Balota et al., 2007) and children (Joseph et al., 2013) and is believed to be a result of accumulated knowledge from repeated exposure (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Reichle, Pollatsek, Fisher, & Rayner, 1998). However, as mentioned above, word frequency measures are not sub-lexical level metrics. Nevertheless, it is a critical component of calculating an individual word's difficulty. Words that are highly frequent are going to be recognized more quickly, regardless of whether they follow standard GPC rules. This is because it is more likely that a child has memorized them such that employing decoding strategies (GPC) is not necessary. Thus, any measure of how difficult a word is to decode must take into consideration how frequently it occurs in written text.

Letter-sound Discrepancy

A critical aspect of word decoding difficulty is the discrepancy between the number of letters and phonemes in a word. The degree of the discrepancy between graphemes and phonemes in a language is known as orthographic transparency, or orthographic depth (e.g., Frith, Wimmer, & Landerl, 1998). A language with a shallow orthography might be described as having an alphabet comprised of letters that only correspond to one sound, with no two letters making the same sound. This is not the case in English, where the orthography is said to be "deep" not "shallow." In English, letters can represent several sounds (for example, the letter "c" representing the/s/sound in "cent" but also the more common/k/sound as in "cat").

Research has shown that for beginning readers and those with dyslexia, learning to read is relatively effortless in transparent languages with shallow orthographies (Frith et al., 1998; Goswami, Gombert, & de Barrera, 1998; Wimmer & Goswami, 1994). Additionally, Seymour, Aro, and Erskine (2003) showed that children who were learning to read in languages with more transparent (regular) mappings between letters and phonemes (e.g., Greek, Finnish, German, Italian, Spanish) were near perfect when reading monosyllabic words by the middle of first grade, yet English-speaking children performed extremely poorly in comparison (34% correct). Thus, degree of orthographic transparency is a key consideration in capturing decoding difficulty. We operationalized transparenecy as the absolute value of the number of letters in a word minus the phonemes. Note that the choice to subtract the phonemes of the word from the number of letters (rather than the graphemes)

was selected because subtracting phonemes from graphemes would always result in a score of 0. This is because graphemes are commonly defined as the written representation of phonemes (Rey, Ziegler, & Jacobs, 2000). The number of phonemes in the word was delineated according to Carnegie Mellon's Pronunciation Dictionary (see http://www.speech.cs.cmu.edu/cgi-bin/cmudict for the full details about how pronunciation is determined), The CMU pronunciation dictionary is open-source and machine-readable and has over 134,000 words broken down into 39 phonemes. Scoring was done in an automated manner via a web-based application that subtracted the number of phonemes from the number of letters in the word.

Vowel and Consonant Grapheme-Phoneme Complexity

In addition to our measure of overall orthographic complexity or regularity ("Letter-sound Discrepancy" detailed above), we also included sub-lexical measures of complexity at the grapheme-phoneme level. Previous research has shown that emerging readers struggle with "complex" graphemes such as vowels. For example, children who are learning how to read, or who struggle with reading at the word level (i.e., have dyslexia), have difficulty learning how to apply appropriate vowel sounds within words (Steacy et al., 2019; Treiman, et al., 2006). Other prior work on word difficulty (e.g., Spencer, 2000) included a measure of "tricky letters or letter combinations." While Spencer's (2000) work was focused on the "relative ease which pupils can spell words," it is important to consider "tricky letter combinations" (graphemes) when developing a measure of decodability. An example of the variation in consonant grapheme-phoneme complexity in the English language can be illustrated by the following words: cat, cent, cello, and ocean. In each of these words the letter (grapheme) 'c' makes a different sound:/k/,/s/,/tch/and/sh/, respectively. This can be contrasted with the letter 'b', in which there are only two sounds:/b/ as in 'bat' and silent b as in 'debt'. Therefore, 'c' can be said to have more complexity than 'b'. The same can be said of vowel sounds, which, on average, have more complexity than consonants in the English language. For example, the single-letter vowel grapheme 'a' can assume up to 8 different sounds, while the multi-letter vowel grapheme 'ee' has only two: 'seen' and 'been'.

Thus, capturing the probability that a certain vowel or consonant grapheme will correspond to a particular phoneme was included as a component in our decoding measure. Corpora have been created to capture GPC 'complexity' (e.g., Berndt, Reggia, & Mitchum, 1987; Fry, 2004). In order to automatically calculate the probabilities, we created a web-based scoring application that mapped the GPCs from Berndt et al. (1987) onto the CMU pronunciation dictionary.

Number of Blends

Another consideration when determining decoding difficulty is how many blends (sometimes referred to as "consonant clusters") a word has. Blends are defined as "…combinations of two or three consonants which, when pronounced, blend into sounds which still retain elements of the individual consonants." (Groff, 1971, p.

59). It is important to account for the number of blends in a word when calculating decoding difficulty because blends are traditionally hard for emerging and struggling readers to master (Bruck & Treiman, 1990; Miller & Limber, 1985; Read, 1975; Treiman, 1985, 1991). Indeed, consonant blends are so hard that teachers have been directed to teach them only once less complex letter-sounds have been mastered as evidenced by reports to teachers (e.g., Shanahan, 2005) as well as policy (e.g., Common Core State Standards; CCSSs: National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010).

Additional Considerations

Other considerations in determining decoding difficulty could be the number of syllables, a metric often used in quantitative leveling systems (e.g., Flesch, 1948; Fry, 1968; Gunning, 1952). However, because the DM algorithm captured the difficulty of vowel graphemes, also including the number of syllables in a word is redundant. This is because each syllable has a vowel; therefore, as long as a decoding metric derived from the above components considered the probability of a vowel grapheme making a particular phoneme, then it would necessarily account for the number of syllables in the word.

Research Questions and Hypotheses

Any measure of decoding difficulty would need to be validated by examining if it, in fact, actually predicted the degree to which a child could accurately read words, both at the individual word level, as well as at the passage level. Furthermore, the decoding difficulty of the words themselves (measurements of text characteristics), if valid, should interact with child level measures (i.e., a child's ability to sound out novel words, as determined by standardized tests). We considered each of these issues carefully when formulating the research questions guiding our initial validation efforts, and posited the following:

(1) Do sub-lexical metrics of decoding difficulty (letter-phoneme discrepancy, grapheme-phoneme complexity, and the number of blends in a word) individually and in combination account for additional variance above and beyond word frequency in predicting oral word reading errors (classified as 1/0 incorrect/correct)? We hypothesized that letter-phoneme discrepancy, grapheme-phoneme complexity, and the number of blends in a word would individually and collectively account for variance in predicting the word reading errors.
(2) If the identified sub-lexical components do predict oral word reading errors, is there furthermore an interaction between them and the child's decoding ability (as defined by lower scores on standardized tests of decoding)? We hypothesized that this interaction would be present if we did indeed find that the DM components individually and collectively predicted word reading errors (RQ 1), with higher DM scores (harder to decode) showing that poorer readers (as defined

by lower scores on standardized tests of word reading) were more likely to read a word incorrectly.

(3) While it is valuable to know if a decoding difficulty measure predicts word reading errors and is furthermore modulated by a child's decoding skill, another way to examine the validity of such a metric would be if the measure also predicted outcomes at the passage level, even after accounting for an existing quantitative measure of passage difficulty. Thus, our third question examined whether the median DM score per passage accounted for variance in predicting overall words read correctly (WCPM) after controlling for estimated Lexile scores, the central measure of text level difficulty currently used by publishers and local and state education agencies.

Methods

Text Measures

Experiments 1 and 2 used the DM only, while Experiment 3 also incorporated a widely used metric of passage-level reading difficulty: Lexile scores. Below we describe the exact way we calculated each component of the DM, followed by a brief description of Lexile scores.

DM Scores

In the DM, the word frequency component score is derived from the Standard Frequency Index corpus (Zeno, Ivens, Millard, & Duvvuri, 1995). Within the DM, it is calculated by subtracting the word frequency percentile score because in the DM harder words are indicated by higher word scores. So, for example, if the word "you" had an SFI score of .80, then .8 would be subtracted from 1 to get .2 (a relatively small amount of points since "you" is a frequent word). This number (.2) would comprise the word frequency score component and would then be added to the remaining three components. For the second component, "letter-phoneme discrepancy", we simply subtracted the number of phonemes from the number of letters in the word. For example, in the case of the word "you", two phonemes would be subtracted from three letters, yielding a score of 1 for this component. To capture the consonant as well as vowel graphemes' probability of making a particular phoneme, the conditional probabilities of the consonant and vowel grapheme-phoneme correspondences in the word were calculated. This metric is captured by determining the frequency of a particular grapheme-phoneme match (i.e., "ou" -/oo/in "you", has a .04 frequency in the Berndt corpus (Berndt, Reggia, & Mitchum, 1987)). Then, in order to align it with the direction of the other components in the DM (i.e., harder to decode words have more points), the frequency is subtracted from 1; in this case, .04 would be subtracted from 1 yielding .96, which would form the conditional probability score component. This was calculated for each GPC match in the word, and the resulting points were added together.

The final DM component (number of blends) was calculated by tallying the number of consonant blends (or clusters) within a word. Digraphs were not included in this component because they are already accounted for in the discrepancy component. Analyses were first conducted to validate that each of the aforementioned components of the DM individually contributed to the probability of a child making an oral reading miscue. Each of these components were then added together to form an overall DM score. Of note, in order to show that the DM contributes variance above and beyond word frequency (a component found in several readability formulas, and the only component that is not measured sub-lexically), the SFI component was removed from the rest of the DM formula. Specifically, the DM word score for experiments 1 and 2 reflect the combinations of only the three sublexical components: number of blends, conditional probability of the vowel and consonant GPCs, and the absolute value of number of letters of words less phonemes in the words. SFI was entered as separate predictor. We refer to this "reduced" form of the DM (incorporating only the sub-lexical components) as the "subDM." In all, the DM's web-based application has the ability to quickly calculate the various components of the DM as well as the total score of the individual words (note that consonant probabilities are currently being integrated into the automated scoring program) in its database of over 100,000 English words.

Lexile Scores

The Lexile scores used in this manuscript are estimated Lexile scores and not certified Lexile scores. Estimated Lexile scores were obtained through the Lexile Analyzer online text scoring tool for researchers (MetaMetrics, 2018). While the formula used in the Lexile Analyzer is proprietary, previous publications (e.g., MetaMetrics, 2007, p. 4–5) on the Lexile framework state that Lexile scores include both average word frequency and average sentence length and is scaled using regression models.

Testing the Individual Contributions of the Components of the Decoding Measure

Analyses were conducted to show that each of the individual components of the DM (word frequency, letter-sound discrepancy, GPC complexity, and the number of blends) contribute to the probability that a child will make a word-reading error. Specifically, a generalized linear mixed effects model was fitted in R version 1.1.423 (R Core Team, 2019) using the lme4 package (Bates, Maechler, Bolker, & Walker, 2014). Five models were run, one for each of the individual components of the DM described in the manuscript: word frequency, letter-sound discrepancy, GPC complexity (separated by consonants and vowels), and the number of blends. In each model the dependent variable was reading miscue (1/0 for incorrect/correct). Participant ID and Item (word) were also in the models as fixed effects. The four models (one for each component) and their statistics (the estimate and the pseudo-R^2) are in Table 2. The function "r.squaredGLMM" was used to determine the pseudo-R^2. This function reports four statistics related to

Table 1 Demographic information across all samples

	Experiment 1 Word-level dependent variable: word-reading error (1/0)	Experiment 2 Word-level dependent variable: word-reading error (1/0)	Experiment 3 Passage level dependent variable: WCPM
Age (years)	8.51 (.32)	8.35 (.34)	11.65 (1.39)
Gender	46% Male	40% Male	50% Male
WASI FSIQ	107 (16)*	105 (14)*	108 (15)
WJBR scaled score	105 (11)	107 (12)	102 (15)
TOWRE	104 (14)	102 (14)	99 (17)

*Denotes that the variable was not used as a covariate in the final analyses since it as not a significant predictor

SFI standard frequency index (Zeno et al., 1995); *WASI FSIQ* Wechsler abbreviated scale of intelligence; *WJBR* Woodcock Johnson basic reading score (which is a composite of the Letter-Word Identification and Word Attack subtests). *TOWRE* test of word reading efficiency; ()=standard deviation

Table 2 Parameter estimates for the individual components of the decoding measure

DM component	Sample 1 estimate and R^2	Sample 2 estimate and R^2
SFI	3.12** (.24)	5.09** (.45)
Letter discrepancy	.20** (.24)	.26** (.42)
Consonant complexity	.33** (.23)	.09* (.41)
Vowel complexity	.18* (.23)	.68** (.42)
Number of blends	.20* (.23)	.88** (.43)

**Indicates significance of $p<.001$ * indicates significance of $p<.05$

pseudo-R^2: marginal delta, marginal theoretical, conditional delta, and conditional theoretical. Since we were interested in the effect of the fixed effects (given the random effects) and did not use model build-up procedures, the conditional theoretical R^2 are reported below. Demographic information on the two samples can be found in the manuscript under "Sample" in Experiments 1 and 2 (see Table 1).

The results of these analyses showed that each individual component of the DM significantly contributed to word-reading errors in both samples (see Table 2). Therefore, the individual components (except SFI, which was entered as a separate predictor) were added together in the subsequent analyses.

General Methods for Experiment 1 and 2: Validation at the Word Level

To address the first two research questions, we used two separate samples (experiments 1 and 2). For both experiments 1 and 2, participants completed behavioral tests over the span of a day. Testing took place in a university campus lab. All participants were consented in accordance with the university's Institutional Review Board. In both experiments, children completed standardized tests of reading ability. These included the Test of Word Reading Efficiency, Second Edition (Torgesen, Wagner, & Rashotte, 1999; Sight Word Efficiency and Phonemic Decoding Efficiency subtests that together provide an overall Test of Word Reading Efficiency score (TOWRE)), and the Woodcock Johnson Tests of Achievement, Third Edition (WJ; Woodcock, McGrew, & Mather, 2001) Letter Word Identification and Word Attack subtests that together form a Basic Reading Cluster score (WJBR). Children's standard scores on the TOWRE and WJBR were used as measures of participants' word-reading abilities. Analyses confined to pseudoword reading abilities (Word Attack and Phonemic Decoding Efficiency only) were conducted and are in the supplemental information. The overall pattern of these results do not differ from those from the main analyses.

After administration of the standardized testing the children read one of two sets of passages (each set comprising two passages) aloud. Children were randomly counterbalanced to passage set as this was part of a larger experiment on the development of children's word reading. Both sets of passages were carefully created to be of similar difficulty according to several variables that were measured in Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004): number of words, sentence length, word length, word frequency, word concreteness, Flesch Reading Ease level and Flesch-Kincaid grade level. The scores for these variables across the passages are in the supplemental information. Briefly, a Flesh-Kincaid grade level of 1 means that a child in first grade would be able to read the text easily. Errors, defined here as words that were read incorrectly, were recorded using a standardized protocol developed frequently used for progress monitoring. For example, inserted words or repeated words are not counted as errors, neither are self-corrections made within 3 s. Incorrectly pronounced words, a hesitation for more than 3 s, and omitted words are all marked as an error. Only errors (not the type of error) were analyzed in these experiments.

The words in the passages were scored in the DM web application, currently in beta version (Saha, et al., 2017, Cutting, et al., 2017). The application provides scores for individual words as well as longer text such as passages. Word scores are a summed total of the elements that increase decoding difficulty (the components described above in "Components of the DM"): therefore, higher word scores indicate that words are harder to decode.

General Data Analysis for Experiment 1 and 2

For experiments 1 and 2 data were analyzed in R version 3.6.2 (R Core Team, 2019) using the lme4 package (Bates, Maechler, Bolker, & Walker, 2014). A generalized linear mixed effects model was used to predict accuracy (coded as a 1 for an error or 0 for correct). Participants and item number were entered as random effects (both as random intercepts). Item number refers to the individual words in the order they were read. We confirmed the need for random effects by analyzing unconditional models with random intercepts for participants and item number. Model fit was examined using Chi square and deviance statistics.

The following fixed effect predictors were entered in a step-wise procedure: reading skill (TOWRE/WJBR), Standard Frequency Index (SFI), subDM score per word (which is the DM without the SFI component included), and then the interaction between basic reading skill (either the WJBR or TOWRE) and the subDM score.

Experiment 1

Sample

This validation study used a sample of participants that was originally recruited for a study investigating reading skill development. Participants were recruited from a metropolitan area in the southeastern United States through flyers, phone calls, and similar methods.

Participants were first screened over the phone to see if they met eligibility criteria. Participants were excluded from the study if he/she met any of the following criteria: (1) previous diagnosis of Intellectual Disability; (2) known uncorrectable visual impairment; (3) treatment of any psychiatric disorder (other than ADHD) with psychotropic medications; (4) history of known neurologic disorder (e.g., epilepsy, spina bifida, cerebral palsy, traumatic brain injury); (5) documented hearing impairment greater than or equal to a 25 dB loss in either ear; (6) individuals known to have IQs below 80; (7) the history of or presence of a pervasive developmental disorder; and, (8) if during testing, parental responses from the Diagnostic Interview for Children and Adolescents-IV (Reich et al., 1997) indicated the presence of any severe psychiatric diagnoses, including major depression, bipolar disorders, and conduct disorder. Children who met criteria for ADHD, Oppositional Defiant Disorder (ODD), adjustment disorder, and mild depression were not excluded, however, children with ADHD who were treated with medications other than stimulants were excluded from this study.

The sample consisted of 62 children with ages ranging from 7.75 years to 9.08 years old ($M = 8.51$, $SD = .32$), and was 46% male (see Table 1). The sample reported their race as follows: 61% of the children were White, 27% were Black/African American, 10% identified as more than one race, and less than 1% identified as Asian or Native American.

Passages

The passages used for this experiment came from the Qualitative Reading Inventory (QRI; Leslie & Caldwell, 2000). The QRI is an informal reading inventory, meaning that it is not standardized or norm-referenced for ability. For this study, children read two QRI passages out loud (See Supplemental Information for characteristics of the passages). Children's passage reading was recorded and later scored for word reading errors. Number of errors ranged from 1 to 171 ($M = 27.90$, $SD = 26.98$) across both passages. Performance was scored by trained graduate students and interrater reliability was conducted on 30% of the participant records and was above 90% each time. Ongoing reliability checks were conducted so as to prevent coder drift. In cases of discrepant results, the first author made the final decision.

Experiment 1 results

TOWRE results

Model fit statistics showed that a generalized mixed effects model with participants and items (ordered individual passage words) entered as random effects (intercepts) was the best model. In the final best-fitting model (see Table 3) SFI was a significant predictor of children's word reading miscues ($b = 2.70$, $SE = .31$, $z = 8.74$, $p < .001$), CI [2.09, 3.32], with fewer word reading errors made on words with lower SFI scores (words with low SFI scores are common words since SFI score is 1- frequency to make it commensurate with DM scoring where higher scores represent harder to decode words). The subDM was also a significant predictor after controlling for SFI and TOWRE ($b = 1.56$, $SE = .19$, $z = 8.33$, $p < .001$), CI [1.19, 1.94]. TOWRE was not significant at the .05 level ($b = -.01$, $SE = .007$, $z = -1.84$, $p = .066$), CI [−.03, −.09], but the interaction between TOWRE and subDM was significant ($b = -.01$, $SE = .002$, $z = -7.74$, $p < .001$), CI [−.02, −.01], with more errors made on words with higher subDM scores (harder to decode) for children with lower TOWRE scores (poorer readers, see Fig. 1).

WJBR results

The same pattern of results held when using WJBR instead of TOWRE. SFI was a significant predictor of children's word reading miscues ($b = 2.70$, $SE = .31$, $z = 8.72$, $p < .001$), CI [2.09, 3.31], with fewer word reading errors made on words with lower SFI scores (words with low SFI scores are common words since SFI score is 1-frequency to make it commensurate with DM scoring where higher scores represent harder to decode words). The subDM was also a significant predictor after controlling for SFI and WJBR ($b = 2.09$, $SE = .25$, $z = 8.69$, $p < .001$), CI [1.59, 2.59]. While WJBR was not significant ($b = -.02$, $SE = .009$, $z = -1.77$, $p = .07$), CI [−.04, .002], the interaction between WJBR and subDM ($b = -.02$,

Table 3 Model fit statistics, and parameter estimates when predicting word reading errors in experiment 1

	With TOWRE						With WJBR					
	Parameters		Model fit				Parameters		Model fit			
	B	*p*<	deviance	χ^2 *p*<	*M* R^2 *T*(*D*)	*C* R^2 *T*(*D*)	*B*	*p*<	deviance	χ^2 *p*<	*M* R^2 *T*(*D*)	*C* R^2 *T*(*D*)
Final Model			9869	.001	.11 (.03)	.25 (.08)			9868	.001	.11 (.04)	.25 (.08)
(intercept)	− 2.10	.01					− 1.68	n.s.				
TOWRE/WJBR	− .01	n.s.					− .02	n.s.				
SFI score	2.70	.001					2.70	.001				
subDM	1.56	.001					2.09	.001				
TOWRE/WJBR* subDM	− .01	.001					− .02	.001				

For all models, data was coded so as 0 = correct, 1 = error. SFI = Standard Frequency Index (Zeno et al., 1995)

SubDM sublexical DM and refers to the reduced form of the DM which does not included the SFI score; *M* Marginal; *C* Conditional; *T* Theoretical; *D* Delta

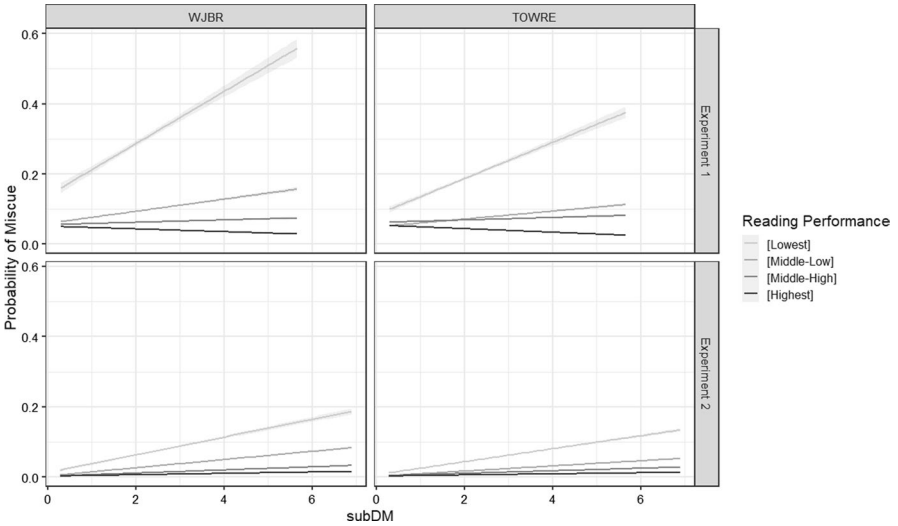

Fig. 1 A graph depicting the interaction of reading ability (WJBR or TOWRE) and the subDM when predicting children's word-reading miscues in experiment 1 and 2. Note that miscues were coded as 1 and correctly read words were coded as 0

$SE = .002$, $z = -7.85$, $p < .001$), CI [−.23, −.01], with more errors being made on words with higher subDM scores (harder to decode) for children with lower WJBR scores (poorer readers, see Fig. 1).

Experiment 2

Sample

Participants for the second sample were recruited through schools and public bulletin boards in and around a metropolitan area in the southeastern United States as part of a larger study of reading development. Participants who showed interest were screened through telephone interviews and excluded if any of the following criteria were present: (1) history of severe psychiatric problems, (2) diagnosis of autism spectrum disorder, including Asperger's, (3) history of uncontrolled seizures, (4) English was not the native (first) language, (5) receiving occupational or physical therapy for motor control issues. Participants were included at a single time point.

The final sample consisted of 88 children (35 males; age range 7.75–9.58 years old; ($M = 8.35$, $SD = .34$). Racial demographics were as follows: 81% white, 13% Black/African American, 5% Asian, 4% more than one race, 3% preferred not to answer, and none reported Alaska native, native American, Native Hawaiian, or Pacific Islander. Furthermore, 6% of the participants were Hispanic/Latino, 3% did not answer, and the rest stated they were not Hispanic/Latino.

Passages

Each child read two experimenter-created passages as part of a larger study on the development of reading ability (See Supplemental Information for characteristics of the passages). One of the passages was expository, and one was narrative. All passages had 350 words and were closely matched using Coh-metrix (Graesser, McNamara, Louwerse, & Cai, 2004) on the following variables: sentence length, word length, word frequency, word concreteness, and Flesch-Kincaid grade level. The order of presentation was counterbalanced across participants.

While these passage level variables are reported, it is important to note that the dependent variable for experiments 1 and 2 is error at the *word* level. While we believe the DM should predict word reading errors regardless of passage level characteristics, we report them for the sake of thoroughness. Children's passage reading was recorded and later scored for word reading errors. Number of errors ranged from 0 to 166 ($M = 17.45$, $SD = 26$) across both passages. Performance was scored by trained graduate students and interrater reliability was conducted on 30% of the participant records and was above 90% each time. Ongoing reliability checks were conducted so as to prevent coder drift and in cases of discrepant results the first author made the final decision.

Experiment 2 results

Towre

Model fit statistics showed that a generalized mixed effects model with participants and items (individual words in the passages) entered as random effects (intercepts) was the best model (see Table 4). In the best-fitting model, SFI was a significant predictor of children's word reading miscues ($b = 4.30$, $SE = .32$, $z = 13.6$, $p < .001$), CI [3.69, 4.93], with fewer word reading errors made on words with lower SFI scores (words with low SFI scores are common words since SFI score is 1-frequency to make it commensurate with DM scoring where higher scores represent harder to decode words). TOWRE was significant ($b = -.04$, $SE = .007$, $z = -6.20$, $p < .001$), CI [−.06, −.03]. The subDM was also a significant predictor after controlling for SFI and TOWRE ($b = .74$, $SE = .13$, $z = 5.80$, $p < .001$), CI [.49, .99] and the interaction between TOWRE and the subDM was also significant ($b = -.01$, $SE = .001$, $z = -3.76$, $p < .001$), CI [−.01, −.003], with more errors being made on words with higher subDM scores (harder to decode) for children with lower TOWRE scores (poorer readers, see Fig. 1).

WJBR results

The same pattern of results held when WJBR was used instead of TOWRE. In the best-fitting model, SFI was a significant predictor of children's word reading miscues ($b = 4.30$, $SE = .32$, $z = 13.6$, $p < .001$), CI [3.69, 4.92], with fewer word reading

Table 4 Model fit statistics, and parameter estimates when predicting word reading miscues in experiment 2

	With TOWRE						With WJBR					
	Parameters		Model fit				Parameters		Model fit			
	B	*p*<	Deviance	χ^2 *p*<	*M* R^2 *T*(*D*)	*C* R^2	*B*	*p*<	Deviance	χ^2 *p*<	*M* R^2 *T*(*D*)	*C* R^2
Final Model			9593	.001	.20 (.02)	.45 (.05)			9575	.001	.22 (.03)	.45 (.06)
(intercept)	− 2.28	.01					− .52	n.s.				
WJBR/TOWRE	− .04	.001					− .06	.001				
SFI score	4.30	.001					4.30	.001				
subDM	.74	.001					.89	.001				
WJBR/TOWRE * subDM	− .01	.001					− .01	.001				

For all models, data was coded so as 0 = correct, 1 = error. SFI = Standard Frequency Index (Zeno et al., 1995)

SubDM sublexical DM and refers to the reduced form of the DM which does not included the SFI score; *M* marginal; *C* conditional; *T* theoretical; *D* delta

errors made on words with lower SFI scores (words with low SFI scores are common words since SFI score is 1-frequency to make it commensurate with subDM scoring where higher scores represent harder to decode words). WJBR was significant ($b = -.06$, $SE = .01$, $z = -7.18$, $p < .001$), CI [−.07, −.04]. The subDM was also a significant predictor after controlling for SFI and WJBR ($b = .89$, $SE = .16$, $z = 5.57$, $p < .001$), CI [.58, 1.21], and the interaction between WJBR and the subDM was also significant ($b = -.01$, $SE = .002$, $z = -3.94$, $p < .001$), CI [−.01, −.003], with more errors being made on words with higher subDM scores (harder to decode) for children with lower WJBR scores (poorer readers, see Fig. 1).

Model convergence

It is important to note that the full model with all three fixed effects and the interaction of TOWRE with subDM yielded a convergence warning for both samples. This can happen when models become too complex (e.g., Bates, Kliegl, Vasishth, & Baayen, 2015; Matuschek, Kliegl, Vasishth, Baayen & Bates, 2017). That said, simplifying the random effects structure did not eliminate the warning nor did we feel it was theoretically justified. Instead, as statisticians have suggested, we opted to use a convergence tolerance of .001 (a less stringent convergence criteria); which all the models met (Lüdecke, 2019). Moreover, the fact that the interaction effect replicated across both samples lead the authors to believe that this is both a real and robust effect.

R^2 in multilevel models

R^2 are often used to quantify the amount of variance that a model captures, however, there are no equivalent R^2 metrics for logistic regression (generalized linear mixed effect) models. Instead, *pseudo*-R^2 measures have been developed specifically for logistic regression models. The pseudo-R^2 for all models are reported in the corresponding results tables. While reported, the authors are aware of the controversy surrounding the appropriateness of interpreting such measures for multi-level models (e.g., Rights & Sterba, 2018; Sotirchos, Fitzgerald, & Crainiceanu, 2019), and generalized multi-level models specifically (Bates, 2014) as their calculation and interpretation are not straightforward. It is also important to note that R^2 was not used to determine model fit. Like others (Sotirchos, et al., 2019), we used AIC, BIC, and deviance, and Chi square tests to determine model fit.

General discussion for experiments 1 and 2

The results show that word frequency, as measured by the SFI, was a significant predictor of children's word reading errors. This is not surprising given abundant research showing frequency effects in children's word reading (e.g., Metsala, 1997). The subDM was also significant across all four samples after controlling for word frequency (SFI), suggesting that the DM is able to explain unique variance in childrens' word-reading errors above and beyond current proxies of

decoding. While novel, this result is consistent with what should be observed with a valid decodability measure, because word frequency is a poor proxy for decodability: frequent words are not always easy to decode. Interestingly, the interaction between the subDM and word reading ability was significant in both samples, with more errors made on words with higher subDM scores (harder to decode) for children who are poor readers. Importantly, this finding replicated across both samples, and across two separate ways of measuring word reading ability. This suggests that the DM is a valid measure of word reading miscues, particularly for those who struggle to read.

Validation evidence for reading fluency: experiment 3

Sample

To address our final research question (experiment 3), we used a third sample. These data were collected as part of a larger study that took place in an urban area in the southeastern United States. Participants were recruited through advertisements in schools, clinics, and doctors' offices. The sample included 215 native English-speaking adolescents ranging from 9 to 14 years ($M = 11.65$, $SD = 1.39$) with a broad range of reading abilities. The sample was 71% Caucasian, 11% African American, 2% Asian, and 4% multi-racial; 9% did not specify. Half of the sample was male (50%).

Participants were excluded if they had: (a) previous diagnosis of intellectual disability; (b) known uncorrectable visual impairment; (c) treatment of any psychiatric disorder (other than ADHD) with psychotropic medications; (d) history of known neurologic disorder; (e) documented hearing impairment greater than or equal to a 25 dB loss in either ear; (f) known full-scale IQ below 80 prior to testing, or a score below 70 on either performance or verbal scales using the Wechsler Intelligence Scale for Children—4th edition (WISC; Wechsler, 2003) determined after enrollment; and/or (g) the history of or presence of a pervasive developmental disorder. Children with ADHD who were treated with medications other than stimulants were not included.

Passages

Children's I.Q. was assessed on the Wechsler Abbreviated Scale of Intelligence (Wechsler, 1999) .[1] Then, children read ten experimenter-created passages as part of a larger study on text complexity (Spencer, Gilmour, Miller, Emerson, Saha, & Cutting, 2019). In experiment 3, all passages were expository (rather than narrative) and covered obscure topics in order to reduce participants' reliance on background knowledge. The order of presentation was counterbalanced across participants.

The passages were submitted to the Lexile analyzer (MetaMetrics, 2018 retrieved from https://lexile.com/educators/tools-to-support-reading-at-school/tools-to-determine-a-books-complexity/the-lexile-analyzer/) and estimated Lexile scores were calculated for each passage (Note that these Lexile scores were obtained for research purposes, and thus per Lexile's guidelines, are to be reported as *not certified*). The estimated Lexile scores ranged from 550 to 850 with a mean of 650 and standard deviation of 100.

In order to calculate the median DM score per passage, the text was scored according to the DM algorithm (Saha, et al., 2017a, Cutting, et al., 2017). This application provided a DM score for every word in the passage, and we calculated the median and mean of the passage. The mean DM score per passage ranged from 2.05 to 2.39. The median DM score per passage ranged from 2.20 to 2.8. Higher DM scores indicate that text is more difficult to decode. Additional details about passage characteristics can be found in (Spencer, et al., 2019).

Data Analysis

Data were analyzed in R version 3.6.2 using the lme4 and MuMIn package (Bartoń, 2014; Bates, et al., 2014). A linear mixed effects model was used to predict words read per minute. Random effects were included after analyzing unconditional models with random intercepts for students and passage topic and comparing model fit using AIC and BIC (see supplemental information for full model build-up statistics). After establishing the need for random effects, the following fixed effects were entered: participant's age, TOWRE standard score, Lexile score per passage, and median DM score per passage. While not specifically text factors, participant age and reading ability were included as fixed effects since, theoretically, they could explain variation in WCPM. Therefore, we investigated variation in performance

[1] Note that for all experiments IQ was assessed by the Wechsler Abbreviated Scale of Intelligence Full Scale I.Q. (First Edition for experiment 1; Wechsler, 1999, and Second edition for experiment 2; Wechsler, 2011). However, analyses including IQ for experiments 1 and 2 showed that IQ was not a significant contributor. These findings were not surprising given a substantial amount of research indicating that decoding is relatively independent from IQ (e.g., Fuchs & Young, 2006; Lyon, Shaywitz, & Shaywitz, 2003). While IQ was significant in experiment 3, the results were the same with or without the inclusion of IQ. Therefore, given these findings, combined with prior empirical findings suggesting that word-level processes are relatively independent from IQ, IQ was not included as a covariate in the final/main models.However, because IQ was significant in the Experiment 3 models, results from those models are included in supplemental information..

Table 5 Model fit statistics, and parameter estimates when predicting WCPM in experiment 3 using the DM

	Parameters			Goodness-of-fit statistics					
	B	CI	*p*	AIC	BIC	χ^2	*Sig.* χ^2	*M* R^2	*C* R^2
Final Model				11,565	11,608	30.62	< .001	.658	.950
(intercept)	203.74	[189.31, 217.99]	< .001						
Age	14.51	[11.56, 17.47]	< .001						
TOWRE	28.81	[25.82, 31.78]	< .001						
Lexile score	− .05	[− .06, − .04]	< .001						
DM median	− 16.12	[− 21.77, − 10.42]	< .001						

Lexile scores are estimated and not certified

SFI standard frequency index (Zeno et al., 1995); *M* marginal; *C* conditional

Table 6 Model fit statistics and parameter estimates when predicting WCPM in experiment 3 using the subDM

	Parameters			Goodness-of-fit statistics					
	B	CI	*p*	AIC	BIC	χ^2	*Sig.* χ^2	*M* R^2	*C* R^2
Final Model				11,564	11,606	31.96	< .001	.659	.950
(intercept)	197.40	[185.11, 209.54]	< .001						
Age	14.51	[11.56, 17.47]	< .001						
TOWRE	28.82	[25.85, 31.78]	< .001						
Lexile score	− .05	[− .06, − .04]	< .001						
subDM median	− 15.75	[− 21.15, − 10.33]	< .001						

Lexile scores are estimated and not certified

SubDM sublexical DM and refers to the reduced form of the DM which does not included the SFI score; *M* marginal, *C* conditional

due to text difficulty (specifically decodability) using one popular measure (Lexile), and our new measure (DM), after controlling for participant characteristics contributing to WCPM such as age, and reading ability (see Table 5). Then, we ran the same models as above using the subDM (Table 6).

Results

Model fit statistics showed that a linear mixed effects model with participants and passages entered as random effects (intercepts) was the best model. Age improved model fit and significantly predicted the number of words read correct per minute, with older children able to read more WCPM (see Table 6).

After controlling for age, reading ability, and estimated Lexile score of the passage, adding the median DM score per passage to the model improved model fit (see Table 5). In the final, best-fitting model, Lexile scores were significant ($b = -.05$,

$SE = .005$, $t = -10.60$, $p < .001$), CI [−.06, −.04], with higher Lexile scores predicting fewer WCPM. The addition of the median DM score per passage further improved model fit ($\chi^2 = 30.62$, $df = 1$, $p < .001$) and was significant ($b = -16.12$, $SE = 2.90$, $t = -5.56$, $p < .001$), CI [−21.77, −10.42], with higher DM scores per passage predicting fewer WCPM.

The results remained even when the subDM median score was calculated per passage. After controlling for Age, TOWRE, and estimated Lexile score of the passage, adding the median subDM score per passage to the model improved model fit (see Table 6). In the final, best-fitting model, Lexile scores were significant ($b = -.05$, $SE = .004$, $t = -10.71$, $p < .001$), CI [−.06, −.04], with higher Lexile scores predicting fewer WCPM. Adding in the median subDM score per passage further improved model fit (see Table 6; $\chi^2 = 31.96$, $df = 1$, $p < .001$) and was significant ($b = -15.75$, $SE = 2.76$, $t = -5.70$, $p < .001$), CI [−21.15, −10.33], with higher subDM scores per passage predicting fewer WCPM.

Discussion for experiment 3

These results provide initial evidence that the DM can predict children's word reading fluency (defined in this study as words read correctly per minute). As the decoding difficulty of the text increased, the number of words read correct per minute decreased. Importantly, the DM predicted fluency after controlling for both participant characteristics known to influence fluency (age and reading ability) as well as a popular measure of text difficulty: Lexile scores.

General discussion

Decoding is a critical step in learning to read and decodable text is used for a wide variety of purposes. However, there is not currently a quantitative measure of text decodability. Word frequency is a poor proxy for decodability as frequent words are often hard for children to decode, particularly for young children. Readability formulas are often not sensitive at lower grade levels, and rarely take into account sub-lexical features of words. Text leveling methods are ambiguous, have issues with reliability, and also lack evidence for validity. It is problematic that decodable text is so widely used yet little empirical work has been conducted to assess whether existing measures are valid. Here, we introduced initial validation evidence for a new measure of decoding difficulty: the DM, which can be used to equate or create decodable text. These preliminary results show that the models that included the DM were a better fit (when predicting children's word-reading miscues) than the models with current (and popular) methods of controlling text at the word level (i.e., word frequency) and at the passage level (Lexile scores).

Limitations

Validation is an ongoing process. While we showed that the DM was able to predict children's word reading errors in three samples, further replication in different

samples would provide additional evidence of validity. Our samples were all from one metropolitan area in the southeastern United States, and using geographically diverse samples would strengthen evidence of validity of the DM. Additional evidence of validity will be needed as this line of research moves forward, and there may be other important aspects of sub-lexical sensitivity that should be part of the construct definition of decoding difficulty. Specifically, we did not include a measure of contextual facilitation. For example, when a given phoneme has multiple possible pronunciations, often the context can facilitate which phoneme is the correct choice. We did not include this because we hypothesized young children would not be sensitive enough to context-dependent pronunciations, but this is certainly an area that should be explored in the future.

In addition to exploring other factors or components that could be included in a decoding measure, one could argue that the way the DM operationalized the components is a potential limitation, and, as such, is an important future direction. For example, to map the graphemes in a word to the phonemes we used CMU's pronunciation dictionary. We used this dictionary because it was open-source, machine-readable, and extensive. Therefore, we were limited to the pronunciations coded by their dictionary. Given that words can be pronounced in various ways, depending on regional dialects, use of a different dictionary is certainly a consideration for additional development of this line of work. We view the findings reported in this paper as a starting point, thus laying the foundation for other dictionaries to be used.

Another limitation and potential future direction is determining the ideal weighting of the individual components of the DM. As described earlier, the components were simply added together in the interest of parsimony; however, advanced statistical modeling techniques could inform ideal values for the coefficients.

Future directions

One future direction is to continue developing the web application and make it available for a wide audience including researchers so that the DM could be used to control for decoding difficult across experimental passages. Or, perhaps teachers could input text to assess the decoding difficulty with respect to an individual student. In a similar vein, we eventually envision creating an assessment (or matching text based on the results of existing decoding assessments) or scale so that children could be matched to texts, much in the same manner as the Lexile Framework. Furthermore, the DM web application might one day be used to create texts that are decodable based upon very specific increments of GPC instruction, and possibly even tailored to individual children as an extreme form of precision teaching.

While the DM is presented here as a model of *text* difficulty, it might also be able to inform *person* characteristics since there is a push for the use of formal, mathematical models in the behavioral sciences (e.g., Muthukrishna & Henrich, 2019). Currently, the DM incorporates several sub-lexical variables and we feel that an exciting future direction would be to investigate how these sub-lexical variables change within an individual over time. For example, in a second-grade reader with poor decoding skills, which grapheme-phoneme correspondences are most

problematic and require the most instruction? Are there certain 'high-utility' GPCs that occur more frequently in children's literature, and therefore should be prioritized? Answers to questions like these could contribute to precision teaching efforts by providing data-based individualization that does not require a large time commitment on the part of the teacher.

Conclusion

Given the wide use of decodable text it is important to have a valid measure of text decodability. Here we presented initial evidence showing that the DM is able to predict children's word reading errors across three samples. Furthermore, we showed that the DM predicts children's fluency (WCPM) across several passages. Not only did it predict children's fluency, but it did so after controlling for the (arguably) best measures available at the word and passage level (word frequency, and Lexile scores, respectively). While validation efforts will continue, we believe that the DM is a promising new measure of text decoding difficulty.

Acknowledgements This work was supported through the following Grants: NICHD/NIH R01 HD044073, NIH/NICHD P50 HD052121, NIH/NICHD R01HD067254, NICHD/NIH R01 HD046130, and NICHD/NIH P30HD015052, R01 HD067253, and OSEP training grant H325D140073.

References

Adams, M. J. (1994). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Amendum, S. J., Conradi, K., & Hiebert, E. (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,* 121–151.

Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language & Communication, 21,* 285–301.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., et al. (2007). The English lexicon project. *Behavior Research Methods, 39,* 445–459.

Bartoń, K. (2014). *MuMIn: Multi-model inference*. R package version 3.5.3. http://CRAN.R-project.org/package=MuMIn.

Bates, D. (2014). *R-Sig-ME mailing list* (*web post*). https://stat.ethz.ch/pipermail/r-sig-mixed-models/2014q4/023007.html.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. Retrieved on May 14, 2020 from https://arxiv.org/abs/1506.04967.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 3.5.3. http://CRAN.Rproject.org/package=lme4.

Beck, I. L., & McCaslin, E. S. (1978). An analysis of dimensions that affect the development of code-breaking ability in eight beginning reading programs. In *Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada*.

Begeny, J. C., & Greene, D. J. (2014). Can readability formulas be used to successfully gauge difficulty of reading materials? *Psychology in the Schools, 51,* 198–215.

Berndt, R. S., Reggia, J. A., & Mitchum, C. C. (1987). Empirically derived probabilities for grapheme-to-phoneme correspondences in English. *Behavior Research Methods, Instruments, & Computers, 19,* 1–9.

Brabham, E. G., & Villaume, S. K. (2002). Leveled text: The good news and the bad news. *The Reading Teacher, 55,* 438–442.

Bruce, B., & Rubin, A. (1988). Readability formulas: Matching tool and task. In A. Davison & G. M. Green (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidered* (pp. 5–22). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Bruce, B., Rubin, A., & Starr, K. (1981). Why readability formulas fail. *IEEE Transactions on Professional Communication, 1,* 50–52.

Bruck, M., & Treiman, R. (1990). Phonological awareness and spelling in normal children and dyslexics: The case of initial consonant clusters. *Journal of Experimental Child Psychology, 50,* 156–178.

Caldwell, J., & Leslie, L. (2000). *Qualitative reading inventory* (3rd ed.). Boston, MA: Allyn & Bacon.

Chall, J. S., & Read, L. T. (1967). *The great debate*. New York: McGraw-Hill.

Cheatham, J. P., & Allor, J. H. (2012). The influence of decodability in early reading text on reading achievement: A review of the evidence. *Reading and Writing: An Interdisciplinary Journal, 25,* 2223–2246.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108,* 204–256.

Cunningham, J. W., Spadorcia, S. A., Erickson, K. A., Koppenhaver, D. A., Sturm, J. M., & Yoder, D. E. (2005). Investigating the instructional supportiveness of leveled texts. *Reading Research Quarterly, 40,* 410–427.

Cutting, L., Saha, N., & Hasselbring, T. (2017). U.S. Patent Application No. 62509856. Washington, DC: U.S. Patent and Trademark Office. System, Method And Computer Program Product For Determining A Decodability Index For One Or More Words.

Davison, A., & Kantor, R. N. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly, 17,* 187–209.

DuBay, W. H. (2007). *Smart language: Readers, readability, and the grading of text*. Costa Mesa: Impact Information.

Ehri, L. C. (1995). Phases of development in learning to read words by sight. *Journal of Research in Reading, 18,* 116–125.

Eldredge, J. L. (2005). Foundations of fluency: An exploration. *Reading Psychology, 26,* 161–181.

Fawson, P. C., & Reutzel, D. R. (2000). But I only have a basal: Implementing guided reading in the early grades. *The Reading Teacher, 54,* 84–97.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32,* 221–233.

Fountas, I. C., & Pinnell, G. S. (1999). *Matching books to readers: Using leveled books in guided reading*. Portsmouth, NH: Heinemann.

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,* 315–342.

Frith, U., Wimmer, H., & Landerl, K. (1998). Differences in phonological recoding in German- and English-speaking children. *Scientific Studies of Reading, 2,* 31–54.

Fry, E. (1968). A readability formula that saves time. *Journal of Reading, 11,* 513–516.

Fry, E. (2002). Readability versus leveling. *The Reading Teacher, 56,* 286–291.

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

Fuchs, D., & Young, C. L. (2006). On the irrelevance of intelligence in predicting responsiveness to reading instruction. *Exceptional Children, 73,* 8–30.

Goswami, U., Gombert, J. E., & de Barrera, L. F. (1998). Children's orthographic representations and linguistic transparency: Nonsense word reading in English, French, and Spanish. *Applied Psycholinguistics, 19,* 19–52.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh–Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36,* 193–202.

Groff, P. (1971). Sequences for teaching consonant clusters. *Journal of Reading Behavior, 4,* 59–65.

Gunning, R. (1952). *The technique of clear writing*. New York, NY: McGraw-Hill.

Harris, A. J., & Jacobson, M. D. (1979). A framework for readability research: Moving beyond Herbert Spencer. *Journal of Reading, 22,* 390–398.

Hatcher, P. (2000). Predictors of reading recovery book levels. *Journal of Research in Reading, 23,* 67–77.

Hiebert, E. H. (2002). Standards, assessment, and text difficulty. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (pp. 337–369). Newark, DE: International Reading Association.

Hiebert, E. H., & Pearson, P. D. (2010). *An examination of current text difficulty indices with early reading texts* (Reading Research Report No. 10-01). Santa Cruz, CA: TextProject, Inc.
Hoffman, J. V., Roser, N. L., Salas, R., Patterson, E., & Pennington, J. (2001). Text leveling and "little books" in first-grade reading. *Journal of Literacy Research, 33,* 507–528.
Hoffman, J. V., Sailors, M., & Patterson, E. U. (2002). Decodable texts for beginning reading instruction: The year 2000 basals. *Journal of Literacy Research, 34,* 269–298.
Hoover, W. A., & Tunmer, W. E. (2018). The simple view of reading: Three assessments of its adequacy. *Remedial and Special Education, 39,* 304–312.
Imagine It! (2011). *Leveled readers*. Retrieved May 14, 2020, from http://www.imagineitreading.com/documents/Docs/Leveled%20Readers%20Readability%20Summary%20Grades%201-8.pdf
Joseph, H. S., Nation, K., & Liversedge, S. P. (2013). Using eye movements to investigate word frequency effects in children's sentence reading. *School Psychological Review, 42,* 207–223.
Juel, C., & Roper-Schneider, D. (1985). The influence of basal readers on first grade reading. *Reading Research Quarterly, 20,* 134–152.
Kern, M. L., & Friedman, H. S. (2009). Early educational milestones as predictors of lifelong academic achievement, midlife adjustment, and longevity. *Journal of Applied Developmental Psychology, 30,* 419–430.
Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology, 95,* 1–47.
Lüdecke, D., (2019). *Sjstats: Statistical functions for regression models* (Version 0.17.4). https://doi.org/10.5281/zenodo.1284472; https://CRAN.R-project.org/package=sjstats
Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2003). A definition of dyslexia. *Annals of Dyslexia, 53,* 1–14.
Maslin, P. (2007). Comparison of readability and decodability levels across five first grade basal programs. *Reading Improvement, 44,* 59–75.
Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94,* 305–315.
McLaughlin, M. J., Speirs, K. E., & Shenassa, E. D. (2014). Reading disability and adult attained education and income: Evidence from a 30-year longitudinal study of a population-based sample. *Journal of Learning Disabilities, 47,* 374–386.
Menton, S., & Hiebert, E. H. (1999). *Literature anthologies: The task for first grade readers* (*Ciera Report No. 1-009*). Ann Arbor, MI: Center for the Improvement of Early Reading Achievement.
Mesmer, H. A. E. (2008). *Tools for matching readers to texts: Research-based practices*. New York, NY: Guilford Press.
MetaMetrics, Inc. (2018). *Lexile analyzer*. Retrieved from https://lexile.com/educators/tools-to-support-reading-at-school/tools-to-determine-a-books-complexity/the-lexile-analyzer/.
MetaMetrics, Inc. (2007). *The lexile framework: A metametrics white paper*. Durham, NC: MetaMetrics.
Metsala, J. L. (1997). An examination of word frequency and neighborhood density in the development of spoken-word recognition. *Memory & Cognition, 25,* 47–56.
Meyer, L. A., Greer, E. A., & Crummey, L. (1987). An analysis of decoding, comprehension, and story text comprehensibility in four first-grade reading programs. *Journal of Reading Behavior, 19,* 69–98.
Miller, P., & Limber, J. (1985). The acquisition of consonant clusters: A paradigm problem. In *Paper presented at the annual Boston University Conference on Language Development, Boston*.
Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour, 3,* 221–229.
National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Washington, DC: CCSSO & National Governors Association.
National Reading Panel (U.S.), & National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel: Teaching children to read: an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: reports of the subgroups.* Washington, D.C.: National Institute of Child Health and Human Development, National Institutes of Health.
Olson, Arthur V. (1984). *Readability formulas: Fact or fiction*. Washington, DC: U.S. Department of Education.
Pearson, P. D., & Hiebert, E. H. (2014). The state of the field: Qualitative analyses of text complexity. *The Elementary School Journal, 115,* 161–183.

R Core Team. (2019). *R: A language and environment for statistical computing. [Computer Software]*. Vienna, Austria: R Foundation for Statistical Computing.

Reading Mastery. (2011). *Lexile levels for SRA Reading Mastery Signature* Editions. Retrieved May 14, 2020, from https://www.mheducation.com/prek-12/program/reading-mastery-signature-edition-2008/MKTSP-UQM08M02.html?page=1&sortby=title&order=asc&bu=seg.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review, 105,* 125–157.

Rey, A., Ziegler, J. C., & Jacobs, A. M. (2000). Graphemes are perceptual reading units. *Cognition, 75,* B1–B12.

Rights, J. D., & Sterba, S. K. (2018). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods, 24,* 309–338.

Saha, N., Bailey, S., & Cutting, L.E. The Decoding System Measure Web Application (Beta version). December 2017.

Saha, N., & Cutting, L. (2019). Exploring the use of network meta-analysis in education: Examining the correlation between ORF and text complexity measures. *Annals of Dyslexia, 69,* 335–354.

Seymour, P. H., Aro, M., Erskine, J. M., & Collaboration with COST Action A8 Network. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94,* 143–174.

Shanahan, T. (2005). *The national reading panel report: Practical advice for teachers*. Naperville, IL: Learning Point Associates.

Sotirchos, E. S., Fitzgerald, K. C., & Crainiceanu, C. M. (2019). Reporting of R2 statistics for mixed-effects regression models. *JAMA Neurology, 76,* 507.

Spencer, K. (2000). Is English a dyslexic language? *Dyslexia, 6,* 152–162.

Spencer, M., Gilmour, A. F., Miller, A. C., Emerson, A. M., Saha, N. M., & Cutting, L. E. (2019). Understanding the influence of text complexity and question type on reading outcomes. *Reading and Writing: An Interdisciplinary Journal*, *32*, 603–637.

Sprick, M. M., Howard, L. M., & Fidanque, A. (1998). *Read well: Critical foundations in primary reading*. Dallas, TX: Sopris West.

Stanovich, K. E. (2009). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Journal of Education, 189,* 23–55.

Steacy, L. M., Compton, D. L., Petscher, Y., Elliott, J. D., Smith, K., Rueckl, J. G., et al. (2019). Development and prediction of context-dependent vowel pronunciation in elementary readers. *Scientific Studies of Reading, 23,* 49–63.

Stein, M., Johnson, B., & Gutlohn, L. (1999). Analyzing beginning reading programs: The relationship between decoding instruction and text. *Remedial and Special Education, 20,* 275–287.

Treiman, R. (1985). Onsets and rimes as units of spoken syllables: Evidence from children. *Journal of Experimental Child Psychology, 39,* 161–181.

Treiman, R. (1991). Children's spelling errors on syllable-initial consonant clusters. *Journal of Educational Psychology, 83,* 346.

Treiman, R., Kessler, B., Zevin, J. D., Bick, S., & Davis, M. (2006). Influence of consonantal context on the reading of vowels: Evidence from children. *Journal of Experimental Child Psychology, 93,* 1–24.

Wechsler, D. (1999). *Wechsler abbreviated scale of intelligence* (2nd ed.). San Antonio, TX: NCS Pearson.

Wechsler, D. (2003). *Wechsler intelligence scale for children* (4th ed.). Cleviland: Psychological Corporation.

Wechsler, D. (2011). *Wechsler abbreviated scale of intelligence* (3rd ed.). San Antonio, TX: NCS Pearson.

Wimmer, H., & Goswami, U. (1994). The infuence of orthographic consistency on reading development: Word recognition in English and German children. *Cognition, 51,* 91–103.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson test of achievement III*. Itasca, IL: Riverside Publishing.

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

Affiliations

Neena M. Saha[1] · Laurie E. Cutting[1] · Stephanie Del Tufo[2] · Stephen Bailey[3]

[1] Department of Special Education at Peabody College, Vanderbilt University, 416C One Magnolia Circle, Nashville, TN 37230, USA

[2] Education Department, University of Delaware, Newark, USA

[3] Columbus, OH, USA