CAN READABILITY FORMULAS BE USED TO SUCCESSFULLY GAUGE DIFFICULTY OF READING MATERIALS?

JOHN C. BEGENY AND DIANA J. GREENE

North Carolina State University

A grade level of reading material is commonly estimated using one or more readability formulas, which purport to measure text difficulty based on specified text characteristics. However, there is limited direction for teachers and publishers regarding which readability formulas (if any) are appropriate indicators of actual text difficulty. Because oral reading fluency (ORF) is considered one primary indicator of an elementary aged student’s overall reading ability, the purpose of this study was to assess the link between leveled reading passages and students’ actual ORF rates. ORF rates of 360 elementary-aged students were used to determine whether reading passages at varying grade levels are, as would be predicted by readability levels, more or less difficult for students to read. Results showed that a small number of readability formulas were fairly good indicators of text, but this was only true at particular grade levels. Additionally, most of the readability formulas were more accurate for higher ability readers. One implication of the findings suggests that teachers should be cautious when making instructional decisions based on purported “grade-leveled” text, and educational researchers and practitioners should strive to assess difficulty of text materials beyond simply using a readability formula. © 2013 Wiley Periodicals, Inc.

CAN READABILITY FORMULAS BE USED TO SUCCESSFULLY GAUGE DIFFICULTY OF READING MATERIALS?

Readability is an attribute of written text, commonly defined by factors that theoretically make text more or less difficult to read (e.g., vocabulary, sentence complexity). To quantify readability, mathematical formulas using such factors have been derived over the last century (Harrison, 1980). From the formulas, readability is calculated and typically expressed as an estimated grade level (e.g., fourth grade reading level). The purpose of the grade level estimate is to define the difficulty of the text, meaning an average reader in that grade should be able to read or “cope” with the book or passage without undue frustration (Bailin & Grafstein, 2001; Compton, Appleton, & Hosp, 2004; Harrison, 1980). The many different formulas used to calculate readability take into account a combination of text factors including one or more of the following: percentage of high frequency easy words (i.e., words on a predetermined list defined as familiar to most students in a particular grade), percentage of hard words (i.e., words not on a list of familiar words), average number of words per sentence, average number of syllables per word, number of single syllable words, or number of words with multiple syllables. The combination of factors and the mathematical constants used in the different formulas can vary significantly from one another, even when theoretically consistent (Connatser & Peac, 1999; Harrison, 1980).

USE OF READABILITY FORMULAS: PAST AND PRESENT

Most conventional readability formulas were developed using general assumptions about reading difficulty. For example, it has been assumed that shorter words, shorter sentences, words with fewer syllables, and words that are used more frequently are easier to read (Connatser & Peac, 1999). Throughout the 20th century, several new formulas have been developed and used for many purposes, ranging from determining the readability of government documents, newspaper articles, schoolbooks, and medical documents (Bailin & Grafstein, 2001; Harrison, 1980). For example, the

Correspondence to: John Begeny, College of Humanities and Social Sciences, Department of Psychology, 640 Poe Hall, Campus Box 7650, Raleigh, NC 27695-7650. E-mail: john_begeny@ncsu.edu
Forcast formula was developed to evaluate exams and entrance forms for the United States Army, and is calculated based on the number of one-syllable words per 100 words (Sticht, 1973). The Fry formula, on the other hand, was developed in Uganda as a method of getting a rough estimate of grade level by plotting the average number of sentences and number of syllables per 100 words on the Fry Graph (Fry, 1968). Table 1 shows a list of other readability formulas and the intended purpose of each respective formula. Each formula listed in Table 1 is also used in the present study.

The growth of readability formulas prompted consumers not only to use them to determine text difficulty, but also to use them as a means of modifying reading materials to a predetermined level. For instance, according to a survey of technical communicators, over one-third of individuals who use readability formulas make modifications to their writing (e.g., by ensuring a manual is at a readability of seventh grade or below), with shortening sentences and substituting words being two primary ways to make grade level estimates lower (Connatser & Peac, 1999). This practice in technical communication and research is questionable, and although several researchers criticize the use of readability formulas and text modification in order to obtain desired readability scores, these practices are common in many fields (Connatser & Peac, 1999), including business (Shelby, 1992) and education (e.g., Bailin & Grafstein, 2001; Harrison, 1980; Kotula, 2003; Lange, 1982; Maslin, 2007).

In education, based on readability levels, librarians make book recommendations (Bailin & Grafstein, 2001), publishers determine difficulty of curriculum materials (Brabham & Villalune, 2002), researchers develop reading assessments and instructional materials (Persampieri, Gortmaker, Daly, Sheridan, & McCurdy, 2006), and teachers differentiate instruction (Kotula, 2003). As highlighted by Dzaldov and Peterson (2005), this type of leveling (i.e., assigning a difficulty ranking to specific reading material) is commonplace and there appears to be a “leveling mania” in education. Leveling may take many forms such as color-coded systems or ranked letter systems (Glasswell & Ford, 2011), but the majority of leveling is defined as grade levels according to readability formulas.

To illustrate, publishers of instructional and intervention materials use readability estimates as a metric of difficulty level and use this metric to sequence materials in reading curricula. Among popular reading curriculum materials, Imagine It! (2011) applies the Flesch-Kincaid readability estimate and Read Well (Peyton & Macpherson, 2009) uses both the Spache and Dale-Chall formulas to modify and sequence reading materials. Similarly, reading intervention programs such as Reading Mastery (2011) compare curricular materials with Lexile rankings (another type of readability measure). The widespread use of readability estimates in education highlights the need to further investigate whether meaningful differences exist between the grade level of the text (defined by readability formulas) and a measure of the actual difficulty level of the text.

**CONTEMPORARY ISSUES WITH READABILITY FORMULAS**

Logically, educators make an assumption that first grade materials are easier than second grade materials; second grade materials are easier than third grade materials, and so forth. How confident they can be in these assumptions is unknown, and how accurately they can choose the appropriate level of text for a student’s reading ability becomes complicated. For example, if an educator is selecting instructional materials for a student who is struggling to read, he or she may choose materials that are at a lower grade level according to the reading curriculum publisher and that publisher’s use of one or more readability formulas to level the reading materials. However, given the possible limitations of readability formulas for determining text difficulty, the materials selected for the student may not, in fact, be easier for the student to read. As such, the student may be assigned inappropriate material for his or her ability level and this may minimize the student’s reading progress. Similarly, teachers may assume that books with higher assigned grade levels are significantly more difficult than those assigned with lower grade levels, and may try to
## Table 1
### Summary of Readability Formula

<table>
<thead>
<tr>
<th>Name</th>
<th>Development</th>
<th>Formula</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dale-Chall</td>
<td>Assess the difficulty of comprehending reading materials for students above fourth grade level.</td>
<td>Grade (= (0.1579 \times \text{percent unfamiliar words}) + (0.0496 \times \text{word/sentence}) + 3.6365)</td>
<td>Dale &amp; Chall, 1948</td>
</tr>
<tr>
<td>Flesch-Kincaid</td>
<td>Assess comprehension difficulty of reading materials for upper elementary through secondary grades.</td>
<td>Grade (= 0.39 \times \text{average words/sentence} + 11.8 \times \text{average syllables/word} - 15.59)</td>
<td>Flesch, 1948</td>
</tr>
<tr>
<td>FOG</td>
<td>Used to assess difficulty based on the grade level necessary for comprehending at least 90% of the reading material.</td>
<td>Grade (= 0.4 \times \left(\text{average words/sentence} + \text{percent of hard words}\right))</td>
<td>Gunning, 1952</td>
</tr>
<tr>
<td>Forcast</td>
<td>Developed to test basic reading competency to enter the U.S. Army and used best for multiple-choice tests and entrance forms.</td>
<td>Grade (= 20 - \left(\text{#single-syllable words/10}\right))</td>
<td>Sticht, 1973</td>
</tr>
<tr>
<td>Fry</td>
<td>Assess readability based on a graph and validated with primary and secondary reading materials.</td>
<td>Grade (= \text{graph average (#sentence, #syllables) from three 100-word passages})</td>
<td>Fry, 1968</td>
</tr>
<tr>
<td>Lexile</td>
<td>Uses both a measure of text difficulty and reader ability to match readers with appropriate text materials.</td>
<td>Based on word frequency and sentence length (rounded to the 10L)</td>
<td>Lexile® <a href="http://www.lexile.com">www.lexile.com</a></td>
</tr>
<tr>
<td>PSK</td>
<td>Developed an improved formula based on four previously established formulas (Dale-Chall, Flesch, Gunning, and Farr-Jenkins-Paterson).</td>
<td>Grade (= 0.0778 \times \text{average sentences length} + 0.455 \times \text{number syllables} - 2.2029)</td>
<td>Powers, Sumners, &amp; Kearl, 1958</td>
</tr>
<tr>
<td>SMOG</td>
<td>Assess reading difficulty based with the goal of having a simple formula based only on the number of polysyllabic words.</td>
<td>Grade (= 3 + (\text{square root of the number of polysyllable word count}))</td>
<td>McLaughlin, 1969</td>
</tr>
<tr>
<td>Spache</td>
<td>Used to assess difficulty of reading materials for students in lower elementary grades (i.e., below third grade).</td>
<td>Grade (= (0.121 \times \text{word/sentence}) + 0.082 \times \text{percent unfamiliar words} + 0.659)</td>
<td>Spache, 1953</td>
</tr>
</tbody>
</table>
challenge advanced students with text that is presumably above the student’s grade level. But again, the supposedly more challenging text may or may not be more challenging. In either situation, teachers are making instructional decisions based on purported levels of readability, yet it cannot be guaranteed that these levels represent meaningful differences (e.g., that a first grade text is actually easier to read than a second grade text).

Additionally, educators may focus on grade levels when selecting appropriate reading assessment materials. A common method of measuring reading skills in elementary school is using oral reading fluency (ORF), which in elementary school is considered a good proxy of reading comprehension and overall reading ability (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Shinn, Good, Knutson, Tilly, 1992), and is a critical skill in early reading development. Materials from the Dynamic Indicators of Basic Early Literacy (DIBELS) system (Good & Kaminiski, 2002) and AIMSweb® system (Howe & Shinn, 2002), for example, include ORF assessments that are now used in thousands of schools throughout the United States, and both systems used readability formulas to develop the assessment passages and to help assign grade levels to those passages.

In addition, some teachers create their own reading assessment or instructional materials, and after entering the text from those materials into web-based programs such as Intervention Central (2011), teachers can determine a readability estimate of the text. Again, with each of the above examples, the assumption is that different grade level passages represent meaningful differences in text difficulty, and this assumption has direct impact on teachers’ appropriate use and selection of reading assessment and/or instruction materials.

Although it is reasonable to argue that text components used in the calculation of readability estimates (e.g., high frequency words, sentence length, number of syllables) are relevant to identify the difficulty of text, the validity of readability formulas is inconclusive in the scientific literature. That is, we do not yet know with much certainty whether each respective readability formula measures what it is intended to measure (i.e., the difficulty of a given selection of text). As such, grade leveled texts, as defined by readability estimates, may or may not discriminate difficulty based on student reading outcomes, and further study is needed to understand the strengths and limitations of readability estimates in education.

**Existing Research Evaluating the Validity of Readability Formulas**

Researchers have questioned the use of readability to assess text difficulty for over 50 years (e.g., Swanson & Fox, 1953), yet grade leveling with readability formulas is still widely used, particularly with reading materials in schools (Dzaldov & Peterson, 2005; Glasswell & Ford, 2011). In response to these concerns, researchers have tested readability estimates using elementary student reading outcomes (e.g., Compton et al., 2004). Because ORF assessment is commonly used in schools to test a student’s overall reading skills, most of the recent studies related to this topic have used ORF as a metric for determining how well readability formulas predict reading outcomes. A small number of studies have attempted to answer the question “Do readability formulas actually predict text difficulty?”

Initial studies tested basic correlations and addressed the question: Are readability estimates good correlates with ORF? Powell-Smith and Bradley-Klug (2001) studied curriculum based measures of reading (CBM-R)—a common measure of ORF—with 36 second graders and assessed whether the Dale-Chall and Fry formulas correlated with actual reading fluency. Students read 20 passages over five weeks; half of the passages came from a basal reading series and half came from “Tests of Oral Fluency” (TORF) passages which are considered generic (i.e., not obtained from the curriculum). Correlations between ORF rates and readability were not significant, as defined by the Fry formula \( r = -.06 \) and the new Dale-Chall formula \( r = -.001 \) for the sample of 60 available passages. The researchers concluded that readability estimates did not improve predictions.
of CBM-R. But despite this finding, many CBM-R passages sets (e.g., AIMSweb, DIBELS) were since developed, at least in part, using readability formulas.

Similarly, Compton et al. (2004) investigated the relationship between a variety of text-leveling systems (including readability formulas) and measures of ORF and accuracy with 248 average or below average second grade readers. Text-leveling systems included readability (i.e., Flesch-Kincaid and Spache), decodability, sentence length, percentage of multisyllabic words, and percentage of high frequency words as defined by Zeno, Ivens, Millard, and Duvvuri (as cited in Compton et al., 2004). Each student in the study read 15 passages considered to be relatively equivalent and all at a second grade level. Percentage of high frequency words was the only variable that was significantly correlated with reading accuracy, while both high frequency words and decodability (i.e., percentage of decodable words) both correlated moderately with reading fluency ($r = .56$ and $r = .58$, respectively). Conversely, although many readability indices include high frequency words as part of the formula, Compton et al. found that readability estimates themselves did not correlate significantly with fluency (Flesch-Kincaid $r = -.08$; Spache $r = .20$). Thus, in both of the aforementioned studies, there was little evidence that readability formulas produce estimates of text difficulty which correlate well with ORF rates.

Other researchers addressed the question by testing whether readability formulas are useful for developing equivalent ORF assessment passages, such as those used by educators to monitor student reading progress and make educational decisions. Ardoin, Suldo, Witt, Aldrich, and McDonald (2005) explored the accuracy of eight different readability formulas in predicting reading fluency rates for 99 third graders. The goal of that study was to determine which formula was the best predictor of reading fluency and whether isolated text components could predict fluency rates. Readability estimates from eight formulas (i.e., Dale-Chall, Flesch-Kincaid, FOG, Forcast, Fry, Powers-Sumner-Kearl [PSK], SMOG, Spache) were compared with fluency rates obtained from six CBM-R passages. Each student read a total of six passages, three from a third grade reading curriculum and three from the same curriculum at fourth grade. Their exploration revealed three primary findings. First, the grade levels estimated by different readability formulas were vastly different. For example, one of the publisher-defined fourth grade passages had a Spache readability estimate of 3.53, but a Forcast readability estimate of 9.60. Second, the average correlations between ORF and readability estimates varied from one another based on a series of Wilcoxin tests that revealed the Forcast formula, which is not even intended for narrative texts nor is it used widely in reading research, had the largest mean tau correlation ($\tau = -.48$). Additionally, the Spache and Dale-Chall formulas, which are commonly used in research and reading material development, had among the smallest magnitudes of correlations ($\tau = -.07$ and $-.03$, respectively) between reading fluency and readability estimates across the different passage length conditions. There were important limitations of this study (e.g., averaging nonsignificant correlation coefficients and drawing conclusions from subsequent analyses of those averages), but limitations notwithstanding, Ardoin et al. offered additional evidence for questioning the predictive validity of readability formulas with actual reading performance.

Ardoin, Williams, Christ, Klubnik, and Wellborn (2010) further tested readability estimates—using Spache, Forcast, and the Lexile Framework—with a 50-passage CBM-R set and compared students’ ORF rates to readability estimates. By comparing the results from 88 participants in second and third grade, Ardoin et al. found that the average tau correlations of readability estimate and ORF were not significant (e.g., Lexile $\tau = .10$, Spache $\tau = .01$). Although tau correlations were significant for 42% of the participants using the Forcast formula, the mean correlation was not significantly different from zero, demonstrating that none of the three estimates were significant predictors in this study. However, results were significant when different students’ ORF rates ($r = .51$) were correlated with one another. The researchers concluded that readability estimates are problematic when equating CBM-R passages, and student data are needed to achieve a robust measure of passage equivalency.
Based on the available research, it appears that readability formulas do not provide estimates that are precise enough to equate passages. However, developing exactly equivalent reading passages may never be possible because the relationship between the readers and reading materials is complex and depends on factors such as each reader’s background (e.g., vocabulary, experiences related to text content), which cannot be captured by readability formulas or equating methods (Bailin & Grafstein, 2001).

Although the aforementioned studies have suggested that readability formula estimates are sometimes poor correlates to ORF and are not adequate for developing equivalent text materials, there is some support that readability formulas provide at least a coarse estimate of relative difficulty. For instance, Hintze and Christ (2004) found that readability estimates could be used to control for difficulty of CBM-R. In their study, 99 students from second to fifth grade served as participants. The researchers created two passage sets: one set consisted of 20 passages selected at random from grade level texts and the other 20-passage set was “controlled” using readability. They chose to use the Spache readability for second and third grade materials, given that the formula was designed for younger grade levels (Spache, 1953). The Dale-Chall formula was used to evaluate passages at fourth and fifth grade levels, as this formula is reportedly intended for older grades (Dale & Chall, 1948). The authors demonstrated that controlling for passage readability resulted in a significantly smaller measurement error (i.e., lower standard error and standard error of estimates), suggesting improved reliability of CBM-R progress monitoring when readability indices are used. This study therefore showed at least some usefulness of readability indices. Francis and colleagues (2008) found similar results when testing passage presentation with DIBELS oral reading fluency passages (DORF). They found that passages from the same reported grade level were highly correlated with one another even though they were not of equal difficulty and could not be considered equivalent.

In summary, depending on the particular research question, design, and analyses—and notwithstanding the limitations of earlier studies—readability formulas have shown mixed evidence for predicting text difficulty with ORF measures. For example, readability formulas appear to be poor correlates with ORF (Compton et al., 2004; Powell-Smith & Bradley-Klug, 2001) and seem to lack the precision necessary to define equivalent assessment measures (Ardoin et al., 2005; 2010). However, other researchers have successfully used formulas to control for general passage difficulty (Hintze & Christ, 2004) and found evidence that grade leveled passages of the DIBELS leveling system (which relied heavily on the Spache formula) were highly correlated with one another (Francis et al., 2008). The studies by Hintze and Christ (2004) and Francis et al. (2008) therefore suggest that readability estimates may have at least some utility in assisting educators with instructional decision-making (e.g., helping educators determine which materials are appropriate for instruction or assessment based on readability formulas). However, before educators can be confident in using readability formulas as a general gauge of text difficulty between closely leveled text (e.g., first versus second grade text, second versus third grade text), research is needed to more specifically address the question that appears to be most practical for educators: Is, for example, third grade text easier than fourth grade text when grade levels are determined by readability levels, and which readability formulas most reliably identify the expected difficulty levels between grade-leveled text?

**Purpose of Current Study**

The goal of the current study is to identify which readability formulas (if any) show an actual correspondence between grade level and difficulty level, when difficulty level is determined by students’ actual reading performance. Important to highlight, unlike most previous research examining readability formulas, it is not a goal of this study to examine whether readability estimates can be used to create “equally difficult” passages or define the precise difficulty of passages, as previous research suggests that such a task is unlikely. Rather, in the current study, differing grade
levels, as determined by eight commonly used readability formulas, were examined to see whether grade levels predicted more or less difficult text, as determined by students’ ORF scores.

The interest in examining grade levels and their general correspondence with difficulty level is because most education practitioners use grade levels to determine the appropriateness of instruction or assessment materials, and are less interested in specifically equating materials or deriving correlations between specific text and a precise readability formula. To examine the general correspondence of grade-leveled text (determined by a readability formula) and the text’s difficulty, this study aimed to address the following two research questions: (a) which readability formulas, if any, are the most valid in discriminating grade level differences as measured by ORF; and (b) does the validity of a readability formula depend on a student’s reading ability level? These types of comparisons have not been previously evaluated, therefore our research questions are exploratory in nature. Although the absolute grade level comparison is unique to this study, it is anticipated that the Spache and Dale-Chall may be the most valid given that there is some research to support these formulas (Francis et al., 2008; Hintze & Christ, 2004).

**Method**

**Participants**

Participant data from 360 students in second (n = 87), third (n = 83), fourth (n = 96), and fifth (n = 94) grades in 21 different classrooms in an elementary school located in the Southeastern United States was used in this study. Prior to participation, we obtained parental consent for each student. The ethnicity of the participants consisted of 41% Black, 40% White, 13% Latino, and 1% Biracial; ethnic background was not available for 5% of the participants. Further, 58% were male, 12% had been retained, 10% received special education services, and 79% received free or reduced-price lunch. Students identified as nonreaders or those with severe reading disabilities were not included.

**Measures**

DIBELS winter benchmark passages (Good & Kaminiski, 2002) were used to measure ORF. DIBELS are used in most, if not all states, according to an article about the increased use of DIBELS in schools (Dessoff, 2007). Benchmark passages are leveled according to grade and are generally used to assess reading levels of students in a school at three points during the school year (i.e., fall, winter, and spring). For the purposes of this study, winter benchmark passages were used because assessments occurred during the winter. Twelve passages, two from each grade (1–6) were used in this study, and were divided into passage sets by grade level.

**Procedure**

**Passage Sets.** Each participant read a set of six DIBELS benchmark passages, determined by the student’s grade level. A set of passages consisted of two passages below the student’s grade level, two passages at grade level, and two passages above grade level (e.g., a second grade student read two first grade, two second grade, and two third grade DIBELS benchmark passages). Overall, there were 12 total DIBELS passages used in the study, two per grade from first through sixth grade.

**Passage Administration.** Standard procedures for assessing ORF using CBM-R, based on Shinn (1989) recommendations, were used. Performance is measured as the number of words read correctly in one minute (WCPM). Passages were administered to students in a quiet hallway outside of the student’s classrooms. During the data collection for this study, all teachers in the school were instructed to keep hallways completely quiet because assessments were taking place throughout the
school. Each student read all six passages in the specific grade level set during one testing period. All passages were counterbalanced to control for potential practice effects.

**Readability Formulas.** Eight readability estimates for each of the 12 passages were calculated using the computer software program Readability Studio. The entire passage was included in the calculation of each, given that previous studies found little to no effect when passage length was varied (Ardoin et al., 2005). The average passage length was 250 words, with a range of 198 to 292 words. The formulas used to calculate estimates included Dale-Chall, Flesch-Kincaid, FOG, Forcast, Fry, PSK, SMOG, and Spache. Additionally, a Lexile score for each of the 12 reading passages was calculated using the web-based Lexile Analyzer (http://www.lexile.com/analyzer/). These readability formulas were chosen because the indices (a) are commonly used in reading research, (b) were tested in the development of DIBELS, and/or (c) were evaluated previously in the literature (e.g., Ardoin et al., 2005; 2010).

Each of the traditional formulas in Readability Studio provides an estimate of readability as a grade level; however, some estimates are more specific to difficulty within grade level, as indicated by a decimal number. For the purposes of this study, “absolute” grade levels were used, meaning that passages were considered at the specific grade level regardless of the decimal number given (e.g., an estimate of 3.8 or 3.1 were both considered to be a third grade passage). Lexile estimates are not given as grade levels; however, the publishers provide a Reading Growth Chart as a means of matching ratings with grade levels (MetaMetrics, 2000). This chart was used to code grade level texts within the Lexile Framework. A summary of the readability scores and corresponding grade levels for the 12 passages used is listed in Table 2.

**Data Analysis Overview**

A nominal coding strategy was developed to compare whether passages at different grade levels correspond with ORF rates that are meaningfully different from one another (e.g., is a third grade passage easier for the student to read than a fourth grade passage?). The ORF rate is defined as WCPM and will henceforth be referred to as such. Each grade level comparison for each student was coded as “expected” versus “unexpected” (as will be explained in detail below). The code took into account the following: WCPM for each participant, the standard error of measurement for CBM-R

---

**Table 2**

**Readability Estimates and Corresponding Grade Levels**

<table>
<thead>
<tr>
<th>Passage</th>
<th>DORF</th>
<th>Dale-Chall</th>
<th>F-K</th>
<th>FOG</th>
<th>Forcast</th>
<th>Fry</th>
<th>Lexile</th>
<th>PSK</th>
<th>SMOG</th>
<th>Spache</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1st</td>
<td>1.0 (1st)</td>
<td>0.9 (1st)</td>
<td>3.3 (3rd)</td>
<td>7.3 (7th)</td>
<td>1 (1st)</td>
<td>350 (1st)</td>
<td>3.7 (3rd)</td>
<td>4 (4th)</td>
<td>2.0 (2nd)</td>
</tr>
<tr>
<td>B</td>
<td>1st</td>
<td>1.0 (1st)</td>
<td>2.2 (2nd)</td>
<td>4.5 (4th)</td>
<td>8 (8th)</td>
<td>2 (2nd)</td>
<td>470 (2nd)</td>
<td>4.1 (4th)</td>
<td>5 (5th)</td>
<td>2.4 (2nd)</td>
</tr>
<tr>
<td>C</td>
<td>2nd</td>
<td>3.0 (3rd)</td>
<td>3.2 (3rd)</td>
<td>5.0 (5th)</td>
<td>7.8 (7th)</td>
<td>3 (3rd)</td>
<td>670 (3rd)</td>
<td>4.2 (4th)</td>
<td>4 (4th)</td>
<td>2.8 (2nd)</td>
</tr>
<tr>
<td>D</td>
<td>2nd</td>
<td>4.0 (4th)</td>
<td>3.7 (3rd)</td>
<td>5.3 (5th)</td>
<td>8.6 (8th)</td>
<td>4 (4th)</td>
<td>750 (4th)</td>
<td>4.4 (4th)</td>
<td>5 (5th)</td>
<td>2.7 (2nd)</td>
</tr>
<tr>
<td>E</td>
<td>3rd</td>
<td>4.0 (4th)</td>
<td>3.8 (3rd)</td>
<td>6.3 (6th)</td>
<td>7.9 (7th)</td>
<td>4 (4th)</td>
<td>750 (4th)</td>
<td>4.4 (4th)</td>
<td>6 (6th)</td>
<td>3.1 (3rd)</td>
</tr>
<tr>
<td>F</td>
<td>3rd</td>
<td>4.0 (4th)</td>
<td>5.7 (5th)</td>
<td>7.7 (7th)</td>
<td>9.2 (9th)</td>
<td>7 (7th)</td>
<td>880 (4th)</td>
<td>5.1 (5th)</td>
<td>8 (8th)</td>
<td>3.2 (3rd)</td>
</tr>
<tr>
<td>G</td>
<td>4th</td>
<td>5.5 (5th)</td>
<td>5.7 (5th)</td>
<td>8.4 (8th)</td>
<td>9.5 (9th)</td>
<td>7 (7th)</td>
<td>880 (4th)</td>
<td>5.2 (5th)</td>
<td>8 (8th)</td>
<td>3.4 (3rd)</td>
</tr>
<tr>
<td>H</td>
<td>4th</td>
<td>4.0 (4th)</td>
<td>6.8 (6th)</td>
<td>8.0 (8th)</td>
<td>9.3 (9th)</td>
<td>7 (7th)</td>
<td>1080 (5th)</td>
<td>5.2 (5th)</td>
<td>7 (7th)</td>
<td>3.8 (3rd)</td>
</tr>
<tr>
<td>I</td>
<td>5th</td>
<td>5.5 (5th)</td>
<td>6.3 (6th)</td>
<td>8.8 (8th)</td>
<td>9 (9th)</td>
<td>7 (7th)</td>
<td>900 (4th)</td>
<td>5.1 (5th)</td>
<td>9 (9th)</td>
<td>3.5 (3rd)</td>
</tr>
<tr>
<td>J</td>
<td>5th</td>
<td>5.5 (5th)</td>
<td>8.1 (8th)</td>
<td>10.4 (10th)</td>
<td>9.8 (9th)</td>
<td>8 (8th)</td>
<td>1140 (6th)</td>
<td>5.6 (5th)</td>
<td>9 (9th)</td>
<td>3.9 (3rd)</td>
</tr>
<tr>
<td>K</td>
<td>6th</td>
<td>5.5 (5th)</td>
<td>8.7 (8th)</td>
<td>10.2 (10th)</td>
<td>9.3 (9th)</td>
<td>8 (8th)</td>
<td>1180 (6th)</td>
<td>5.6 (5th)</td>
<td>9 (9th)</td>
<td>4.4 (4th)</td>
</tr>
<tr>
<td>L</td>
<td>6th</td>
<td>5.5 (5th)</td>
<td>8.9 (8th)</td>
<td>9.6 (9th)</td>
<td>11.6 (11th)</td>
<td>10 (10th)</td>
<td>1050 (5th)</td>
<td>6.2 (6th)</td>
<td>10 (10th)</td>
<td>3.9 (3rd)</td>
</tr>
</tbody>
</table>
Table 3
Same Grade Level Groups for Passages A-L

<table>
<thead>
<tr>
<th>Formula</th>
<th>1st Grade</th>
<th>2nd Grade</th>
<th>3rd Grade</th>
<th>4th Grade</th>
<th>5th Grade</th>
<th>6th Grade</th>
<th>7th Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dale-Chall</td>
<td>A B</td>
<td>C</td>
<td>D E F H</td>
<td>G I J K L</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Flesch-Kincaid</td>
<td>A B C D E</td>
<td>–</td>
<td>F G H I</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FOG</td>
<td>–</td>
<td>–</td>
<td>A B C D E</td>
<td>E F G H</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forecast</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>A C E</td>
</tr>
<tr>
<td>Fry</td>
<td>A B C D E</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexile</td>
<td>A B C D E</td>
<td>–</td>
<td>D E F G I</td>
<td>H L J K</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSK</td>
<td>–</td>
<td>–</td>
<td>A B C D E</td>
<td>F G H I J K</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMOG</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>A C B D E</td>
<td>E H</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spache</td>
<td>–</td>
<td>A B C D E</td>
<td>F G H I J L</td>
<td>–</td>
<td>–</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

passages, and grade level of passages determined by readability estimates. Below is a description of each factor in the coding system.

**Participant WCPM.** Each student read a total of six passages based on his or her grade level (according to the passage sets described earlier), and from each passage a WCPM score was obtained. The objective of this study is to determine whether WCPM differences between passages were associated with how each readability formula estimated the passage’s difficulty level (e.g., based on the student’s WCPM scores, passages estimated at second grade according to the Spache formula should result in more WCPM than passages estimated at third grade level with the Spache formula). WCPM scores were considered meaningfully different if the difference exceeded the standard error of measurement for CBM-R (Christ & Silberglitt, 2007).

**Standard Error of Measurement.** The standard error of measurement (SEM) is used in CBM-R to identify how much error around the observed score is attributable to measurement error, and for the purposes of this study, SEM was used to differentiate meaningful differences between scores (i.e., scores that differ by more than the expected SEM) versus negligible differences between scores (i.e., scores that do not differ by more than the expected SEM). The expected SEM was determined based upon a study by Christ and Silberglitt (2007). In that study, the researchers calculated approximate SEM for CBM-R passages based on eight years of data with over 8,000 first through fifth graders. The SEM reported by Christ and Silberglitt varied by grade level, sample type, and reliability of measurement. For a typical sample, and with expected levels of reliability at 0.93, the SEM estimates by grade are as follows: first grade (8 WCPM), second grade (9 WCPM), third grade (10 WCPM), fourth grade (10 WCPM), and fifth grade (10 WCPM). These SEM estimates are likely the best representation of error ranges to date, and were used to define meaningful differences between WCPM scores at different grade levels.

**Grade Leveling and Same Grade-Level Groups.** As evidenced by the readability estimates listed in Table 2, there is marked variability in the grade level assigned to a given text based on the readability formula used—a finding noted elsewhere in readability research of this nature (e.g., Ardoi et al., 2005; 2010). As a result, passages were collapsed into a “Same Grade-Level Group” when passages have the same grade level according to a particular readability formula. For example, passage A is the only first grade passage as calculated with the Fry formula, so this is the only first grade passage in the Fry, first grade, Same Grade-Level Group. However, passages D and E are calculated by the Fry formula as fourth grade passages, so both of these passages are in the Fry, fourth grade, Same Grade-Level Group. Table 3 shows the list of passages according to each
readability formula’s Same Grade-Level Group. As shown in Table 3, because Same Grade-Level Groups vary according to each specific readability formula (e.g., Fry, Spache, Dale-Chall), there is natural variability in leveling created by each formula, and this was a focus of the research questions in this study.

Same Grade-Level Groups were developed to represent grade levels that would be used by consumers, such as teachers, using readability levels. Most other researchers have looked for relative differences in readability estimates without regard for absolute grade levels (e.g., Ardoin et al., 2005; Compton et al., 2004; Powell-Smith & Bradley-Klug, 2001), yet as stated previously, teachers are more likely to consider the readability estimate of a passage or book that corresponds to a specific absolute grade level (e.g., first, second, third). Therefore, testing the passages according to the absolute grade level offers an aspect of practical significance that was not evaluated in earlier studies.

Data Analysis

To determine whether the patterns follow an anticipated or “expected” pattern, comparisons of WCPM for Same Grade-Level Groups were made. As an example, a second grade student read passages A-F, and according to Flesch-Kincaid, passage A is categorized as first grade, passage B is second grade, passages C, D, E are third grade, and passage F is fifth grade. For each represented Same Grade-Level Group (1st, 2nd, 3rd, 5th), a WCPM score was assigned based on either the WCPM score for a single passage or an average of WCPM scores for all of the passages in that group. Once all data were categorized, comparisons across consecutive grade levels were made to determine if the patterns match an expected pattern. In other words, the study evaluated whether WCPM scores are lower for the higher Same Grade-Level Group (e.g., is the WCPM score for the student’s third grade level group lower, as would be expected, than the student’s WCPM score for the second grade group?). Thus, grade level comparisons were only made between Same Grade Level Groups that vary by one grade (e.g., first grade compared to second grade, fourth grade compared to fifth grade, second grade compared to third grade). Last, because the oldest students who participated in the study were fifth graders, any comparisons above seventh grade were eliminated from the analyses.

To further clarify the analytic framework for this study, consider the following case illustration. Jimmy, a third grade student, reads at the following ORF rates: passage C (112 WCPM), passage D (108 WCPM), passage E (110 WCPM), passage F (107 WCPM), passage G (103 WCPM), passage H (92 WCPM). According to the Lexile, grade level categories are as follows: passage C (third grade), passage D (fourth grade), passage E (fourth grade), passage F (fourth grade), passage G (fourth grade), passage H (fifth grade). As a result, comparisons are made between third and fourth grade and also between fourth and fifth grade, and this will ascertain two different “expected” or “unexpected” results. For the third to fourth grade comparison, Jimmy’s 112 WCPM score (the third grade score) is compared to 107 WCPM (i.e., his average WCPM scores of passages D, E, F, and G since each of these four passages are defined by the Lexile as fourth grade passages). The difference between these two numbers is 5 WCPM. Because the SEM for third grade is 10 WCPM, there is not a meaningful difference and this is therefore coded as “unexpected” (i.e., we would expect a difference in difficulty between third and fourth grade material, as defined by the Lexile formula, but in this case Jimmy’s scores did not meet the criterion for a meaningful difference). However, the fourth to fifth grade comparison is considered “expected” because 92 WCPM on the fifth grade passage (passage H) is 15 WCPM less than his 107 WCPM on the fourth grade passage category. It is expected that there would be a difference this large (i.e., beyond the corresponding SEM) because the fifth grade passage should be significantly more difficult than the fourth grade passages, according to the Lexile formula.
Table 4
Grade Level Comparisons Following Expected Patterns

<table>
<thead>
<tr>
<th>Comparisons</th>
<th>Dale-Chall</th>
<th>F-K</th>
<th>FOG</th>
<th>Fry</th>
<th>Lexile</th>
<th>PSK</th>
<th>SMOG</th>
<th>Spache</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st vs. 2nd</td>
<td>– 41% (87)</td>
<td>–</td>
<td>41% (87)</td>
<td>–</td>
<td>41% (87)</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>2nd vs. 3rd</td>
<td>– 22% (87)</td>
<td>–</td>
<td>5% (87)</td>
<td>5% (87)</td>
<td>–</td>
<td>–</td>
<td>– 65% (170)**</td>
<td></td>
</tr>
<tr>
<td>3rd vs. 4th</td>
<td>70% (170)**</td>
<td>–</td>
<td>41% (87)</td>
<td>36% (170)</td>
<td>73% (170)**</td>
<td>46% (87)</td>
<td>–</td>
<td>18% (93)</td>
</tr>
<tr>
<td>4th vs. 5th</td>
<td>85% (272)**</td>
<td>–</td>
<td>15% (87)</td>
<td>–</td>
<td>4% (272)</td>
<td>44% (266)</td>
<td>58% (170)* –</td>
<td></td>
</tr>
<tr>
<td>5th vs. 6th</td>
<td>– 8% (272)</td>
<td>62% (170)**</td>
<td>–</td>
<td>48% (189)</td>
<td>59% (93)</td>
<td>49% (170)</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>6th vs. 7th</td>
<td>–</td>
<td>–</td>
<td>15% (266)</td>
<td>–</td>
<td>–</td>
<td>8% (179)</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>79% (442)**</td>
<td>17% (446)</td>
<td>32% (610)</td>
<td>29% (344)</td>
<td>33% (805)</td>
<td>48% (446)</td>
<td>38% (519)</td>
<td>48% (263)</td>
</tr>
</tbody>
</table>

Note. The number in parentheses is the total number of comparisons for that particular grade level and readability formula. An asterisk denotes that significantly more than 50% were expected results. An underline denotes that significantly more than 50% of the comparisons followed the unexpected pattern based on the readability formulas and grade levels compared. *p < .05, **p < .01.

In summary, codes of expected versus unexpected patterns of ORF rates were determined using WCPM, SEM, and Same Grade-Level Groups, and this was determined for each individual student in the study. Individual analyses were then converted into specific frequency counts of students who, within the specific analysis (e.g., the Lexile analysis for fourth versus fifth grade passages) obtained an expected or unexpected result. Table 4 shows the frequency counts reported for each formula and grade level comparison.

To determine a possible statistically significant difference (setting alpha at $p < .05$) between the expected values obtained for each of the readability formulas, the respective total frequency of expected versus unexpected results for each readability index was tested using binomial exact tests, setting a 50/50 split of expected versus unexpected results as the theoretical criteria. Individual results were conducted for the 25 available grade level comparisons along with a total comparison for each of the eight readability formulas (the Forcast formula was eliminated because no passages had a readability level below seventh grade).

To examine the possible effects of student reading ability, a series of Fishers’ exact tests were conducted. Student ability groups were assigned based on the oral reading fluency rates of their grade level material (according to the DIBELS leveling system). The top 33.3% of students (i.e., those who read the most WCPM) were in the “high ability” group and the bottom 33.3% of students (those who read the least WCPM) were considered in the “low ability” reading group. The middle 33.3% of students were in the “average” group and were eliminated from the comparison of ability levels. Students were grouped in the categories listed above so that two relatively distinct groups of students could be compared. The reason to compare distinct groups is that, in practice, teachers are unlikely to differentiate instruction (e.g., select higher or lower grade level texts) for the average reader, and are more likely to do so for the highest and lowest groups of readers. However, only the middle 33.3% was eliminated as the “average” group because an adequate sampling of students is necessary to make valid comparisons. Fishers’ exact tests based on high versus low ability and expected versus unexpected results were compared for the total available comparisons for each of the readability formulas.

**RESULTS**

Correspondence between readability formula levels and student reading rates (i.e., WCPM) was tested based on the frequency of expected results according to the coding scheme described previously. The percentage of expected (i.e., correct) classifications based on the total available comparisons is presented in Table 4. Results of exact binomial tests for each grade level comparison
Table 5
Readability Formula Comparisons Following Expected Patterns by Student Ability Level

<table>
<thead>
<tr>
<th>Risk Level</th>
<th>Readability Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dale-Chall</td>
</tr>
<tr>
<td>Lower Ability</td>
<td>66% (148)</td>
</tr>
<tr>
<td>Higher Ability</td>
<td>88% (146)</td>
</tr>
<tr>
<td>Difference</td>
<td>22%**</td>
</tr>
<tr>
<td>Total</td>
<td>77% (294)</td>
</tr>
</tbody>
</table>

Note. The number in parentheses is the total number of comparisons for that particular ability level and readability formula. The Higher Ability group is classified as the highest 33.3% of the readers on grade level text. The Lower Ability group is comprised of the lowest 33.3% of readers as classified by their reading rates on grade level texts.

*(p < .05, **p < .01).

(and a total comparison) per formula are shown, and asterisks denote whether the results differed significantly from chance (i.e., a 50/50 split of expected versus unexpected). It should be noted that some comparisons were significantly different from chance, but in the opposite direction (i.e., the unexpected results were much more frequent than expected results). These significant differences are denoted with an underline. For clarity, only comparisons in which the percentage of expected results is significantly higher than unexpected (with p < .05) provide support for a formula as a valid index of text difficulty, as measured by ORF.

With regard to individual grade level comparisons, only 6 of the 25 possible comparisons had significantly more expected results (3rd vs. 4th Dale-Chall, 4th vs. 5th Dale-Chall, 5th vs. 6th FOG, 3rd vs. 4th Lexile, 4th vs. 5th SMOG, 2nd vs. 3rd Spache). Only the Dale-Chall formula was significant for the total comparisons, with almost 80% of all of the comparisons falling within the expected category. Inspection of Table 4 reveals that the majority of comparisons were below 50%, with some as low as 5%, meaning that very few students had reading outcomes that followed the patterns that would be predicted by the grade level of text assigned by the particular readability formula.

Table 4 also reveals a fairly large degree of variation within a particular readability formula and between the different grade level comparisons of that formula. For example, with the Lexile formula, expected percentages ranged from 5% (2nd vs. 3rd) to 73% (3rd vs. 4th), and nearly all readability formulas evidenced ranges of more than 30 percentage points. The only exceptions to these wide ranges in grade comparison values were the Dale-Chall and PSK formulas, which had ranges of only 15 percentage points.

To test whether the patterns of expected results depended on student ability (Lower Ability and Higher Ability), a series of Fisher’s Exact Tests were conducted. Table 5 shows the percentage of comparisons following the expected pattern for the two ability groups and the percentage difference between the expected values. With four of the eight readability formulas (Dale-Chall, Fry, SMOG, and Spache), Higher Ability students showed significantly better “expected patterns” than Lower Ability students. Table 5 also shows that for every formula, the number of comparisons following an expected pattern is lower for Lower Ability students, and with the exception of the Dale-Chall formula, all other percentages (for Higher Ability and Lower Ability students) revealed relatively low expected patterns (i.e., all less than 55%).

**Discussion**

The purpose of the current study was to identify which readability formulas (if any) showed an expected correspondence between grade level (as determined by the readability formula) and

*Psychology in the Schools*  DOI: 10.1002/pits
difficulty level (as determined by students’ ORF performance). Our goal was to answer the following two questions: (a) which readability formulas, if any, are the most valid in discriminating grade level differences as measured by ORF; and (b) does the validity of a readability formula depend on a student’s reading ability level? Of the readability formulas examined in this study, only one (the Dale-Chall formula) was identified as a valid measure of text difficulty level for each of its comparisons. Specifically, findings showed that the Dale-Chall formula successfully discriminated between both of the grade level comparisons that were evaluated with the formula (i.e., 3rd vs. 4th grade materials and 4th vs. 5th grade materials). This finding is encouraging because the Dale-Chall formula is used in readability research (e.g., Hintze & Christ, 2004; Kotula, 2003) and has been used for decades to help educators determine which reading materials are appropriate for students. Thus, as one of the oldest and most commonly used formulas, the Dale-Chall appears to be a relatively reliable formula for gauging general text difficulty across grades 3–5. In fact, this finding is consistent with the intended purpose of the formula, which is to gauge text difficulty around the fourth grade level and above.

Other than the Dale-Chall formula, some formulas (i.e., FOG for 5th vs. 6th grades, Lexile for 3rd vs. 4th grades, and Spache for 2nd vs. 3rd grades) showed promise as valid indicators for one specific grade level comparison, but these three formulas were inconsistent and not found to be valid indicators for other grade-level comparisons. In fact, an overall examination of all formulas showed that most evidenced greater numbers of unexpected results. Overall findings from this study therefore suggest that most common readability formulas may be inappropriate to use across a range of grade levels, or for any grade level, when an educator is trying to discriminate general difficulty level among reading materials.

Our study is the first to test several different readability formulas on basic but practical distinctions among text difficulty (i.e., distinctions based solely on absolute grade levels), but the findings are in many ways consistent with previous studies. For example, there has been at least some evidence documenting the utility of readability formulas such as the Spache and Dale-Chall (Hintze & Christ, 2004). Also, a high frequency (i.e., familiar) word list—specifically, the Dale-Chall list of 3,000 familiar words—is a primary component of the Dale-Chall formula; and other studies have found that the percentage of high frequency words correlates well with student reading rates (Ardoin et al., 2005; Compton et al., 2004). Therefore, the results of this study not only highlight at least one potentially useful formula, but also relate to previous findings showing that percentage of high frequency words may be a good gauge of text difficulty, particularly when difficulty is assessed according to students’ ORF. However, like our findings suggesting that most readability formulas are unable to successfully discriminate general text difficulty (or at best, do so inconsistently across grade levels), earlier research has shown poor correlations between readability levels and student reading rates (Ardoin et al., 2010; Powell-Smith & Bradley-Klug, 2001).

Another key finding of this study was that readability formulas appear to be more accurate when predicting text difficulty for students who are better readers. In fact, for nearly half of the formulas (i.e., Dale-Chall, Fry, SMOG, and Spache) there was a significant difference between Higher and Lower Ability students when examining the percentage of expected difficulty-level outcomes. One previous study (Compton et al., 2004) tested the correlation between students’ ORF with readability levels and found little to no differences between students who are average readers compared to below average readers. However, in this study, expected difficulty-level outcomes were compared among a different type of sample (i.e., the highest 33% of readers and the lowest 33% of readers), which may help to explain the differential findings between the two studies. Also important to reiterate, this study examined the highest and lowest readers because, in practice, teachers would be most likely to seek text of higher or lower difficulty for above average and below average readers.
Although these findings are interesting and shed light on another variable that might be considered when understanding the validity of readability formulas, the results should be interpreted while keeping in mind that, with the exception of the Dale-Chall formula, expected patterns for each formula and each category (Higher and Lower Ability students) were all generally at or below 50%. Therefore, for many readability formulas, although there appears to be a significant difference in the expected patterns when comparing the formulas to above and below average readers, the overall predictive ability—regardless of the student’s ability level—was usually still no better than chance.

**Implications for Researchers and Practitioners**

Overall findings of this study suggested that only one of the readability formulas (Dale-Chall) demonstrated to be a valid and consistent indicator of text difficulty when compared to a commonly used proxy of elementary-aged students’ reading ability (i.e., ORF). In other words, nearly all formulas did not follow expected patterns (e.g., that second grade material would be easier than third grade material) and several formulas actually produced many unexpected patterns (e.g., second and third grade materials were equally difficult or third grade materials were actually easier for students to read than second grade materials). This finding suggests that most readability formulas may not assist teachers very well with selecting text that is of greater or lesser difficulty, whether the purpose of text selection is for instructional or assessment purposes. Further, although findings showed that several readability formulas seem to be better at differentiating text that is read by students with higher versus lower reading abilities, even with higher performing students nearly all formulas do not appear to be valid indicators of text that is more or less difficult. Teachers should therefore be cautious of using most readability formulas (or leveling systems based on readability formulas) for elementary-aged students, and for this reason the reported “leveling mania” (Dzaldov & Peterson, 2005) within schools must be tempered by understanding that there appears to be little evidence that leveling with readability formulas is a completely valid and consistent way of differentiating text difficulty.

Despite these seemingly pessimistic findings about using readability formulas, there are some approaches that teachers (and educational researchers) may take when trying to identify text of varying difficulty for instructional or intervention purposes. First, this study found that the Dale-Chall formula could be used to discriminate differing levels of text in over 70% of comparisons for grades 3, 4, and 5. As a result, in the older elementary grades it appears reasonable to use this formula as an indicator of text difficulty. The Spache formula (developed for early elementary texts) also showed evidence for validity in the comparison of second and third grade texts for 65% of the comparisons and could reasonably be used in earlier elementary grades as an initial measure of difficulty level. It is promising that two of the most commonly used and well-established formulas were useful in discriminating difficulty levels and may be appropriate to continue using as an initial guide to text levels for educators. However, further study is necessary to test whether Dale-Chall could be accurate for other grade levels.

Because many readability formulas may not be appropriate for teachers when understanding and possibly selecting text of varying difficulty levels, teachers are in need of alternatives when trying to understand text difficulty for their students. One approach, if time and resources permit, might be for the teacher to have the student (or a few representative students) read a small portion of the potentially assigned text aloud for one minute. Because the student’s ORF provides a meaningful estimate of the text difficulty for the individual student, such an assessment would probably be the most valid indicator of text difficulty. For example, if the text is too difficult for the student, the ORF is likely to be less than desirable (e.g., below the average for their grade or below their...
personal average) and/or they are likely to misread several words (e.g., higher than average words read incorrectly). An evaluation like this would allow teachers to rule out texts that are either too easy or too difficult for the student. Yet, it is also important to consider the reason the student is reading a particular text. For instance, if the goal of the reading material is to expose the student to new concepts or vocabulary, then reading the text may be appropriate even if it appears too difficult for the student. Overall, students should regularly read texts that are at or slightly above their current reading level in order to help them improve their basic reading proficiency; however, occasionally reading materials that are otherwise more difficult or easy is acceptable if it fits with strategic instructional decision-making.

Another option for teachers is to simply use a presequenced curriculum of materials that is integrated as part of an instructional program with empirical evidence of effectiveness for the students the teacher aims to instruct. Using this approach would eliminate the need for teachers to be concerned with text levels, and it should eliminate or minimize the need to specifically select or create text materials to use for intervention or instruction because the sequenced text would have already been validated as part of the overall instructional program. As described by Begeny, Ross, Greene, Mitchell, and Whitehouse (2012), having instructional materials available as part of a specific reading program—rather than, for example, simply using research-based strategies that do not include specific reading materials—may be an important advantage for teachers when selecting instructional programs and the corresponding materials.

With regard to selecting text of a particular difficulty for assessment purposes, this may be more complicated than selecting text for everyday instruction because assessment results are typically used for instructional decision-making. Therefore, teachers and researchers are simply encouraged to use norm-referenced and psychometrically validated assessment materials, and should not try to use readability formulas for selecting assessment materials. As noted by Ardoin et al. (2010), assessments leveled according to student reading data are more predictive of student reading outcomes than readability levels. Thus, assessments with known psychometric properties should be used, because even if the text is too easy or too difficult for the student, normative comparisons can still be made to accurately assess the student’s reading ability.

Overall, the “leveling mania” (Dzaldov & Peterson, 2005) that has been sweeping classrooms over the past years has highlighted the need to evaluate how we think about appropriate reading materials for students. It is important for educators to recognize that grade levels of published reading materials are often defined by readability formulas, but there is very little evidence that these measures are valid measures of text difficulty when compared to students’ actual reading performance. For this reason, teachers might consider the levels of books as one metric for understanding text difficulty, but they should also consider how the books were leveled (e.g., with readability formulas, and if so, which formulas were used) and whether other textual components such as vocabulary and conceptual content are relevant. Additionally, when necessary, educators can assess text difficulty by having some students in the class (e.g., an above average, average, and below average reader) read a passage from the text aloud for a minute and use reading rates to better determine the difficulty level of the text. Teachers are also well advised to use instructional materials that are integrated as part of an instructional program that has research-based evidence of effectiveness.

Limitations and Future Directions

One limitation of this study was that grade-level groups were not equally represented for each formula, which limited the extensiveness of our data analyses. For instance, the PSK formula classified 6 of 12 passages at the fifth grade level, but no passages at first, second, or seventh grade levels. Similarly only grades 2, 3, and 4 were represented by the Spache readability formula. This
limitation may not be surprising because previous research also found that a particular passage is often classified by grade level very differently across varying readability formulas (Ardoin et al., 2005; 2010; Compton et al., 2004). However, future research that can include numerous (probably hundreds) different passages and have the time and resources to administer many of these passages to a large sample of students would help to control for equal representation of grade level groups among the different formulas. Similarly, future research might also consider how such passages are developed (e.g., with readability formulas, and if so, how they were used in development), and whether passages developed (at least in some part) with readability formulas could impact study results.

The present study is also limited because only reading fluency was assessed to determine reading difficulty, and there are many important elements, such as comprehension and vocabulary, which may be important to consider when discussing text difficulty levels. Although reading fluency has been well established as a strong indicator of students’ overall reading skills (e.g., Fuchs et al., 2001), future research evaluating a wider range of reading skills may contribute to understanding the relationship between readability formulas and actual text difficulty levels.

Later research might also explore in greater depth the extent to which specific student reading skills could influence a formula’s ability to predict the expected outcome. As stated previously, we had particular interest in examining each formula’s expected patterns for low-performing readers; and with this subgroup we found one formula to perform better than chance (i.e., > 50%). However, because readability formulas differ (see Table 1), it is possible that one student’s reading difficulty (e.g., trouble reading multisyllabic words) would or would not be “captured” by a specific formula, compared to a student who, for example, mostly had difficulty with basic decoding. This detailed level of analysis likely extends beyond the general applicability hoped for by most readability developers, but such research could improve how teachers use readability to make instructional decisions for struggling readers.

Finally, future research might be designed to replicate the general methodology of the current study, but use different reading materials. If subsequent research finds patterns of results like those from this study (e.g., the valid use of Dale-Chall for grades 3–5, Spache for grades 2 and 3, FOG for grades 5 and 6), it would provide valuable information for educational researchers and practitioners who are trying to determine whether they can rely and/or use particular readability formulas. For example, replicated data may inform educators that, for example, the Spache formula can be used successfully as a gauge of text difficulty in grades 2 and 3 and Dale-Chall can be used in grades 3–5. Given the unique calculations of each readability formula, and the many variables that influence text difficulty for a particular student, it is not surprising that some formulas may be more or less appropriate at certain grade levels. Thus, substantiating which formulas can be used for a given grade level should add significant practical value for educational researchers and practitioners.

REFERENCES


Readability Formulas


