Read Writ (2018) 31:813–833
https://doi.org/10.1007/s11145-017-9815-4

Investigating the validity of two widely used quantitative text tools

James W. Cunningham[1] · Elfrieda H. Hiebert[2] · Heidi Anne Mesmer[3]

Published online: 6 January 2018

Abstract In recent years, readability formulas have gained new prominence as a basis for selecting texts for learning and assessment. Variables that quantitative tools count (e.g., word frequency, sentence length) provide valid measures of text complexity insofar as they accurately predict representative and high-quality criteria. The longstanding consensus of text researchers has been that such criteria will measure readers' comprehension of sample texts. This study used Bormuth's (1969) rigorously developed criterion measure to investigate two of today's most widely used quantitative text tools—the Lexile Framework and the Flesch–Kincaid Grade-Level formula. Correlations between the two tools' complexity scores and Bormuth's measured difficulties of criterion passages were only moderately high in light of the literature and new high stakes uses for such tools. These correlations declined a small amount when passages from the University grade band of use were removed. The ability of these tools to predict measured text difficulties within any single grade band below University was low. Analyses showed that word complexity made a larger contribution relative to sentence complexity when each tool's predictors were regressed on the Bormuth criterion rather than their original criteria. When the criterion was texts' grade band of use instead of mean cloze scores, neither tool classified texts well and errors disproportionally placed texts from higher grade bands into lower ones. Results suggest these two text tools may lack adequate validity for their current uses in educational settings.

✉ James W. Cunningham
jwcunnin@email.unc.edu

[1] University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[2] TextProject, Santa Cruz, CA, USA

[3] Virginia Tech, Blacksburg, VA, USA

Keywords Quantitative text analysis · Readability · Text complexity · Text difficulty

Introduction

In recent years, educators and researchers have shown a heightened interest in quantitative and qualitative approaches to the analysis of text complexity and the prediction of text difficulty (Cunningham & Mesmer, 2014). In this study, we concentrate on quantitative measures (historically called readability formulas) because they currently receive the bulk of attention in the everyday lives of schools.

Quantitative measures of text complexity have enjoyed greater prominence than qualitative measures for both assessment and instruction. They are more straightforward to use than qualitative ones, which typically require considerable training of raters or evaluators (Pearson & Hiebert, 2014). Further, the ease of using quantitative measures has increased in the digital era. Unlike previous generations of readability formulas that required manual computation of features of printed texts, the new generation of text complexity systems allows for rapid analysis of even book-length digitized texts (Mesmer, 2008).

Broad application of quantitative measures of text complexity appears to have benefitted from the Common Core State Standards (CCSS) writers' specification of text complexity bands for particular grades in Appendix A of the Standards (National Governors Association Center for Best Practices (NGA) & Council of Chief State School Officers (CCSSO) 2010a) and in a later study elicited by Student Achievement Partners (Nelson, Perfetti, Liben, & Liben, 2012). The Lexile tool is applied nearly universally in the test programs of all 50 states (Metametrics, 2017a) and in "over 115,000 books, 80 million articles, and 60,000 websites" (Metametrics, 2017b, "Target instruction for all learners," para. 1). The Consortia for the Common Core Assessments (PARCC, n.d.; Smarter Balanced, 2016) also use Lexiles.

The present study was prompted by questions about the validity of the current generation of readability formulas. Further, availability of a dataset based on an extensive sample of students on a carefully delineated set of texts (Bormuth, 1969) made it possible for us to objectively and independently examine the validity of two of the most widely used quantitative text tools.

Two dimensions of validity of quantitative text tools

Almost all readability formulas, past and present, are regression equations. Predictor variables in an equation represent countable features of a text (e.g., mean sentence length in words). A dependent (criterion) variable is selected to serve as a measure of text difficulty (Mesmer, Cunningham, & Hiebert, 2012). The parameters of an equation (its constant and beta weights) are generated by statistical modeling. A sample of texts is analyzed to produce a score for each one on every predictor and a difficulty level is assigned to each text on the criterion variable. The model is then fitted to the data to yield parameters that maximize the model's prediction of the criterion. The resulting equation is used to predict the difficulty of other texts in the population from which the sample was taken.

The validity of a tool for predicting the difficulty of a text has two dimensions. The first dimension is the criterion validity of the regression equation, typically reported as an R^2. This dimension of validity addresses the percentage of variance in the criterion accounted for by the model (equation). In regression, a common term for the process of model fitting is *validation*, the degree to which the model predicts the criterion variable. Specific to readability formulas, a relatively high R^2 in the data from the sample of texts has typically been expected before an equation can be considered valid enough to be used to predict the difficulty of other texts.

Since a formula is intended to predict the difficulty of texts for readers, it is also necessary to attend to a second dimension of validity: The validity of the criterion variable as a measure of text difficulty. If the criterion lacks validity as a measure of the difficulty of sample texts for readers (dimension 2), it does not matter how well the equation predicts that criterion (dimension 1).

Over the decades, text researchers have studied many text features as predictors, but only a few different criterion variables of text difficulty have been employed in the same literature (Klare, 1984; Nelson et al., 2012). An implication inherent in much of the literature is that it is on the predictor side where progress stands to be made in quantitative text tool development, rather than on the criterion side. Yet, after a career of studying quantitative methods for predicting text difficulty, Klare (1984) concluded that they "can be no more accurate than the criteria on which they are based" (pp. 701–702). Therefore, our focus in this study was on the criterion variable, a relatively neglected, but crucial factor in text research. At the end of the day, for all their speed and ease of use, quantitative text tools still require validation on a criterion (dimension 1) that itself has established validity as a measure of text difficulty for readers (dimension 2).

The criterion variable in historical and recent text complexity research

Several histories of readability have been published (e.g., Klare, 1963, 1974, 1984), but none has focused on the criterion variable. Our historical review shows how an enduring consensus developed that readers' comprehension performance on sample texts should constitute the criterion variable for validating a quantitative text tool.

Evolution of the consensus (1923–1971)

Early criterion variables

For the first readability formula, Lively and Pressey (1923) used 16 texts in rank order of difficulty based on subjective judgment as the criterion but, soon thereafter, an objective criterion variable for validating a readability formula was developed. Vogel and Washburne (1928) assigned levels to books based on median scores on the paragraph-comprehension subtest of the Stanford Achievement Test of students who had read and enjoyed them.

Dale and Tyler (1934) were the first to measure readers' performance on a set of passages for use as their criterion variable. All 74 passages were on the topic of personal health and the measure the researchers gave to 800 adults used a single multiple-choice comprehension task for each passage. Participants were asked to select the best and worst conclusions from five choices. Gray and Leary (1935) soon applied a similar strategy with a somewhat larger sample, 1000 adults, and a wider range of text types.

Norming passages with pre-assigned difficulties

Beginning with Lorge (1939), the publisher's grade placements of the McCall–Crabbs' *Standard Test Lessons in Reading* (1925, 1950, 1961) became the dominant criterion variable in text research. The use of the McCall–Crabbs provided researchers with a criterion variable that was much less expensive in labor and cost than collecting reader performance data on sample or norming passages. The McCall and Crabbs's passages were used as the criterion variable for widely used readability formulas, including Flesch's (1948) original formula.

The 25-year dominance of the McCall–Crabbs' (1925, 1950, 1961) texts as the criterion variable for readability research must, in retrospect, be questioned. Stevens (1980) analyzed all extant documentation for the lessons and interviewed William A. McCall. According to McCall, whatever data may have been collected to assign the grade placement levels to the lessons had been neither extensive nor evaluated for reliability. McCall claimed to have been unaware that the lessons had been used to develop or validate readability formulas.

Even with the almost exclusive use of the McCall–Crabbs' test lessons as a criterion variable from 1939 until the mid-1960s, their dominance was not complete; a few researchers employed other criterion variables. However, none of these alternatives marked a return to aggregated reader comprehension performance, but relied on norming passages with pre-assigned difficulties (e.g., Dolch, 1948; Spache, 1953).

Readers' comprehension performance as the criterion measure

The failure to validate readability equations on a measure of reader performance drew heavy criticisms by the mid-1960s (Klare, 1984), leading researchers to look for effective ways to assess students' comprehension of sample passages. A promising new means of assessing comprehension for this purpose was the cloze procedure (Taylor, 1953). Cloze systematically deletes noncontiguous words from a passage, replacing them with blanks that readers fill in. Traditionally, exact replacement (except for spelling) is required for a correct response. Coleman (1965) was the first to use cloze test performance as the criterion variable for validating readability formulas, while Bormuth (1969) employed the technique soon thereafter.

The consensus

After a quarter century of reliance on publishers' judgments of the difficulty of their texts for criterion variables, most readability researchers returned to reader performance. Although a measure of any type of reader performance (oral reading accuracy or fluency, critical reading, stamina, etc.) could conceivably provide the criterion variable on which a set of text feature counts would be regressed, most criterion variables based on reader performance after the mid-1960s employed some measure of reading comprehension (Klare, 1984).

Criterion variables used to develop two widely used text tools (1975–present)

The consensus that text difficulty of sample passages should be determined using a measure of readers' comprehension has generally been adhered to since it was reached. In this section, we review the specific criterion variables used to develop the two text tools examined in this study and evaluate the current status of the consensus.

Revision of the Flesch reading ease formula

When revising Flesch's (1948) Reading Ease formula, Kincaid, Fishburne, Rogers, and Chissom (1975) kept the same predictors (mean syllables per word and mean sentence length) but estimated new parameters using a criterion variable that relied on both cloze and standardized multiple-choice test performance. For their revision, Navy enlisted personnel took either the middle- or high-school form of the comprehension subtest of the Gates–MacGinitie Reading Test (GMRT). The researchers then made a cloze test for each of 18 GMRT passages and administered them to the participants. Fifty percent of participants with GMRT performance at a particular grade needed to score 35% or better on cloze for a passage to be assigned that grade level. A multiple regression analysis was conducted using these graded passages as the criterion variable to validate the revised grade-level version of the Flesch formula.

The Lexile framework

The development of the Lexile Framework (Stenner, Smith, & Burdick, 1983) began when the researchers determined that the common (base 10) logarithm of a word's frequency from Carroll, Davies, & Richman's (1971) *The American Heritage Word Frequency Book* database was the best predictor of a word's logit difficulty on the Peabody Picture Vocabulary Test-Revised (PPVT; Dunn & Dunn, 1981).

According to Stenner (1996), an additional analysis was conducted using the mean log word frequency (MLWF) for the 66 reading comprehension test items (each consisting of a single sentence accompanied by four pictures) from the *Peabody Individual Achievement Test* (PIAT; Dunn & Markwardt, 1970). The rank order of the 66 test items reported by the publisher served as the observed item difficulty (criterion variable). The mean of the log word frequencies for each sentence provided the highest correlation of any of their semantic variables with the item rank order criterion. The log of the mean sentence length was the best syntactic predictor of PIAT item difficulty rank.

The resulting provisional regression equation was used to assign predicted difficulties to 400 norming passages. The investigators created a one-sentence summary of each passage with one deletion for which four choices were provided. This unique measure of passage comprehension was referred to as the *native item type*. These 400 items were administered to approximately 3000 students in grades 2–12.

Based on students' performances, 138 items were removed because of "misfitting" (Stenner & Burdick, 1997, p. 11), leaving 262 items. The observed logit scores for the sentence length and word frequency variables for the passages were entered into a regression analysis with the Rasch scale score based on the 3000 students' performances on the 262 items as the criterion variable. The beta coefficients in the equation became the parameters for the Lexile Framework.

Current status of the consensus

The nearly 50-year consensus for using a measure of readers' comprehension of sample passages as the criterion variable in text research was largely maintained until Nelson et al. (2012). Of their seven reference (criterion) variables, three were based on aggregated student comprehension performance, but four represented a return to relying on publishers' or experts' judgments of the difficulty of sample texts. It remains to be seen whether the Nelson et al. (2012) study was an aberration or an indication that the consensus no longer holds. If the consensus does end, the problem will not be when readers' measured ability to read different texts agrees with the judgments of publishers or experts, but when they disagree. If judgments of text difficulty are preferred over readers' performances with the texts, how can those judgments ever be validated or even questioned?

Measuring readers' comprehension for criterion variables

The first readability researchers to use readers' performance on norming passages as the criterion variable (Dale & Tyler, 1934; Gray & Leary, 1935) created multiple-choice comprehension-assessment tasks. However, by the time the consensus was reached, cloze had become the comprehension measure of choice (Klare, 1984).

The use of cloze as a criterion measure

Bormuth (1971) makes it clear it was not cloze per se, but a measurement principle that lay behind the preference for cloze at the time the consensus was reached. He argues that "tests [composed of comprehension questions] are subject to unpredictable variations in the size of the mean scores, variations that are due to the uncontrolled ways test writers select and phrase the items included in the tests" (p. 3). He goes on to state that a criterion measure should be affected only "by the characteristics of the passage itself and not by any other source of systematic variance" (p. 26). Such a measurement approach, Bormuth acknowledges, would not necessarily be the one chosen when making a reading comprehension achievement test, but is best for the criterion variable in readability research because there "the variance of interest [is] the between passage variance" (p. 27). Cloze was selected because it was the available approach most consistent with the principle. Since cloze test items are selected systematically (every *n*th word), do not provide distractors, and are scored for exact replacement rather than allowing synonyms, there is no test-constructor or test-scorer variance. Consequently, all dependable variance is due to variations in the demands texts place on readers.

The decisions by Coleman (1965) and Bormuth (1969) to use cloze in their research received independent empirical support from Miller (1975) who found that criterion validities (R^2s; dimension 1) of the same readability equations were higher when the criterion measure was cloze than when it was multiple-choice comprehension questions.

Comprehension theory, the consensus, and cloze as a criterion variable

The consensus that criteria for readability research should be (a) readers' comprehension of sample passages, and (b) use a comprehension measure that minimizes test-constructor and test-scorer variance, occurred prior to the theoretical work and research on text comprehension that began in the mid-1970s. Nevertheless, the consensus seems consistent with any theory positing an essential and separable contribution to comprehension made by the text. For example, Kintsch's (1988) construction-integration model of discourse comprehension theorizes that comprehenders form "concepts and propositions directly corresponding to the linguistic input" (p. 166), and that "The words and phrases that make up a discourse are the raw material from which a mental representation of the meaning of that discourse is constructed" (p. 180). For Kintsch, knowledge of language and the world is essential for comprehension, but so is the text: "Initial processing is strictly bottom-up" (p. 163).

An indispensable role for the text in comprehension is consistent with another characteristic of cloze as a measure for criterion variables: every item is highly passage dependent (i.e., having the text available is necessary but not sufficient for scoring well). Were there no text accompanying the blanks, it is very improbable a particular student would correctly fill in any particular item. Cloze item difficulties vary with the knowledge required of readers, but no amount of knowledge can compensate for inability to read the text. Of course, any other highly passage dependent measure would serve as well as cloze to meet this principle.

Cloze as a criterion measure since the consensus was reached

Although the criterion variable used to validate the Flesh–Kincaid tool relied on both cloze and standardized multiple-choice test performance, cloze fell into disuse after 1975. The criterion for the Degrees of Reading Power quantitative text tool (DRP) was modified-cloze scores, with selected rather than systematic deletions and a multiple-choice response format (Koslin, Zeno, & Koslin, 1987). As discussed earlier, the criterion variable used in the development of Lexile was another kind of deletion item with multiple choices. In one sense, these alternative measures are consistent with the consensus in that they do not include comprehension questions. Their use is also understandable because of the economy and ease of scoring multiple-choice tests. However, because they provided distractors, the alternatives may have reintroduced test-constructor variance into criterion variables and reduced the passage dependence of items.

Traditional cloze may also have been abandoned because of studies questioning whether it measures comprehension beyond the sentence level (Shanahan, Kamil, and Tobin, 1982). Their study compared cloze scores for original texts versus ones with randomly reordered sentences and found no significant difference. However, this design may be flawed. McNamara and Kintsch (1996) found that local cohesion (which should be reduced by scrambling sentences) benefits low-knowledge readers, but that *reducing* local cohesion benefits high-knowledge readers. If so, this design could fail to find cloze's sensitivity to local cohesion because scrambling sentences may increase the cloze scores of high-knowledge readers, but decrease those of low-knowledge readers without changing the overall mean significantly.

Recent research on the construct(s) measured by reading comprehension tests

The consensus among text researchers since the 1960s has been that reading comprehension performance on norming passages should serve as the criterion variable when validating a text tool. However, recent studies provide both encouragement and discouragement for this idea. For example, Keenan, Betjemann, and Olson (2008) and Muijselaar, Kendeou, de Jong, & van den Broek (2017) found that some widely used comprehension tests differ from each other on how much variance is accounted for by measures of word reading versus listening comprehension. This finding suggests the discouraging possibility that any criterion variable—including that used by Bormuth, the Lexile Framework, the Flesch–Kincaid, or any other tool—may be more a measure of word reading than comprehension. Future theorizing is needed to consider the ideal range of balance of dependence on word reading versus language comprehension in a comprehension criterion measure. Perhaps in future, text researchers will be expected to evaluate existing and potential criterion variables for whether they achieve such a balance. Or, conversely, text researchers may be able to demonstrate that text-dependent test items measure word reading to a greater extent than text-independent items, but without necessarily lessening their validity as tests of comprehension, i.e., that it is not a zero-sum game.

More encouraging is the finding that different texts and question types do not result in different types of comprehension being assessed (Muijselaar, Swart, Steenbeek-Planting, Droop, Verhoeven, & de Jong, 2017). Their finding suggests that reliance on a single item type such as cloze does not necessarily mean comprehension scores are only generalizable to that task.

Rationale for the present study

In this study, we used Bormuth's (1969) criterion variable to investigate the validity of two of the most widely used quantitative text tools, Lexile and the Flesch–Kincaid. We chose Bormuth's criterion for two reasons. First, Bormuth used very rigorous methods to develop his criterion variable (as will become apparent in the description that follows). We could find no recent criterion with comparable strengths to the same degree. Second, his traditional cloze criterion measure exemplified the measurement principles of minimizing test-constructor and test-scorer variance and maximizing text dependency.

Research questions

1. How well do two text tools predict Bormuth's criterion variable?
2. What change in parameter estimates occurs when each text tool predicts Bormuth's criterion variable?
3. How well does each text tool predict text difficulty within individual grade bands?
4. How accurately does each text tool predict the grade bands of Bormuth's sample passages?

Method

The criterion variable

Bormuth's 330 passages and mean student performance data associated with each passage were obtained from the developers of the DRP—Touchstone Applied Science Associates (TASA) (now part of Questar Assessment) and converted into digital form.

The 330 texts from which the sample passages were selected came from 10 subject areas spanning the curriculum and five levels of school use: "grades 1–3, 4–6, 7–9, 10–12, and college" (Bormuth, 1969, p. 11). A passage of approximately 110 words was selected from each of the 330 texts. The sample came from a randomly selected page and paragraph number.

Because Bormuth's (1969) sample passages ($X = 110$ words) were relatively short, we compared the cloze means from Bormuth (1969) and Bormuth (1971) to examine their stability. The 1971 study used 32 of the 330 passages but that were longer ($X = 250$ words) in length. We identified the 32 corresponding passages from the two studies with different participants and then correlated their cloze means from both studies. Results supported the stability of the cloze means for the shorter passages ($r = .98$; $r^2 = .95$).

Bormuth's (1969) sample consisted of approximately 2600 students from schools in Minneapolis suburbs (predominantly Caucasian, middle class at the time) with students distributed as follows: approximately 500 in grades 4 through 6, 1000 in grades 7 through 9, and 1000 in grades 10 through 12. These students were divided into 50 matched groups using their scores on the *California Reading Achievement Test* (1963 edition). Because there were about half as many students in grades 4–6 as there were in grades 7–9 or grades 10–12, the younger students were each assigned to two of the 50 matched groups. The final 50 matched groups had 57 students each.

Bormuth (1969) created five forms of each passage, resulting in 1650 cloze tests (i.e., 330 passages × 5 forms). In the first form of a passage, words 1, 6, 11, and so on were deleted, words 2, 7, 12, and so on in the second, and so forth. Fifty test booklets were created, one for every matched group of participants. Each test booklet had 33 randomly arranged, unique test passages. In all, 94,050 cloze test protocols were scored for exact replacement except for spelling errors. Bormuth computed a difficulty score for each of the 330 passages by calculating the proportion of students' correct responses per item and then averaging these responses across all words to achieve an aggregated cloze score (CM) for each passage.

Predictor/independent variables

We used both the individual variables that make up the Flesch–Kincaid equation—mean word length in syllables (Sylls/Wds) and mean sentence length in words (Wds/Sent)—as well as the overall Flesch–Kincaid grade level estimate for each passage as predictors of Bormuth's criterion variable. The predictor variables for Lexile were, first, the estimated difficulty in Lexiles of each sample passage, then alternatively, the value for every passage on each of the two output variables for each passage from the Lexile tool: mean log word frequency (MLWF) and mean sentence length (MSL).

Procedure

The CM from each passage formed the criterion variable for this study. We also entered Bormuth's reported grade band for each passage. Data on predictors were obtained by using the Flesch–Kincaid and Lexile tools to computationally analyze each of the 330 sample passages.

Text difficulty, on the Flesch–Kincaid, is reported as an overall grade level on Microsoft Word as well as in other programs, which made it necessary to find a means of calculating Sylls/Wd and Wds/Sent within the 330 texts. The analyzer at Readability.com was used to obtain the Flesch–Kincaid level, Sylls/Wd, and Wds/Sent for each passage.

The MetaMetrics site, lexile.com, provides free access to the Lexile Analyzer. Each of the 330 passages was independently entered into the Lexile Analyzer, providing the Lexile, MSL, MLWF, and number of words.

Results

Research question 1: How well do two text tools predict Bormuth's criterion variable?

To answer this question, we performed both a Spearman's *rho* and a simple regression analysis of each tool's scores for the 330 sample passages with Bormuth's mean cloze scores as criterion. For the Flesch–Kincaid, the Spearman's rank order correlation was $rho = -.78$. Unfortunately, the result of the simple regression ($\beta = -.02$; $r = -.76$; $r^2 = .57$; $p < .001$) could not be compared with how well the tool had predicted its criterion variable when its equation was developed, since Kincaid et al. (1975) did not report the correlation coefficient from their multiple regression analysis.

For Lexile, the Spearman's rank order correlation was $rho = -.70$. The result of the simple regression analysis ($\beta = .00$; $r = -.70$; $r^2 = .49$; $p < .001$) could be indirectly compared with the correlation from a study by the developers of the Lexile equation using its original criterion measure (items from the PIAT). Stenner (1996) reported that, in the earlier study, the Lexile predictors had accounted for 85% of the variance in PIAT item rankings (equivalent to a multiple correlation of approximately .92). It could also be compared with the correlation of .97 from the study that set the parameters for the Lexile equation (Stenner & Burdick, 1997).

We were also able to compare our results with those reported for the Lexile tool in the Nelson et al. (2012) study, since it was one of the seven quantitative text tools they investigated. Three of Nelson et al.'s seven criterion variables were based on aggregated student comprehension performance. Of these, two covered the full range of texts from grade 1-adult. The Spearman's rank order correlations between Lexile estimates and those two criterion variables were .74 (Gates–MacGinite) and .95 (Oasis Empirical). The Pearson product correlations between Lexile estimates and those two criterion variables were .75 (Gates–MacGinite) and .95 (Oasis Empirical). The Pearson correlation between Lexile estimates and Bormuth's criterion variable (− .70) in our study was near in size (.75) to Lexile's correlation with Rasch testlet scores from the Gates–MacGinitie in the Nelson et al. (2012) study. These correlations are noticeably lower than the approximately .92 correlation reported in the Stenner (1996) study and the .95 correlation with Oasis Empirical reported in the Nelson et al. (2012) study.

Research question 2: What change in parameter estimates occurs when each text tool predicts Bormuth's criterion variable?

As discussed earlier, the regression equation in the Flesch–Kincaid was modeled and validated using a new criterion variable (Kincaid et al., 1975) that replaced the McCall–Crabbs passages originally used by Flesch (1948). The unstandardized regression coefficients in the Flesch–Kincaid equation are:

$$\begin{aligned}\text{Constant}: &\quad \beta = -15.59\\ \text{Sylls/Wd}: &\quad \beta = 11.8\\ \text{Wds/Sent}: &\quad \beta = 0.39\end{aligned}$$

To examine the difference using Bormuth's criterion variable would make to the parameters in the Flesch–Kincaid equation, we conducted a multiple regression analysis using Flesch–Kincaid's two predictors, Sylls/Wd and Wds/Sent, to predict Bormuth's mean cloze scores for the 330 passages. Bormuth's mean cloze scores needed to be transformed to the Flesch–Kincaid metric for this multiple regression analysis since the cloze scores lie on a proportional metric from 0 to 1 and Flesch–Kincaid's criterion lay on a grade-level metric.

To accomplish the transformation, we first used the Flesch–Kincaid equation to compute a predicted grade level for each of the 330 passages. Then, we entered those predicted grade level scores as the dependent variable in a simple regression analysis where Bormuth's mean cloze scores were the predictor. Finally, we used the resulting regression equation to transform each passage's mean cloze score to an equivalent Flesch–Kincaid grade level. The Pearson correlation between Bormuth's mean cloze scores and transformed grade level scores for the 330 passages was $r = -1.00$.

The multiple regression analysis to answer question 2 for the Flesch–Kincaid used its two predictors, Sylls/Wd and Wds/Sent, to predict Bormuth's mean cloze scores transformed to the Flesch–Kincaid grade level scale, as shown in Table 1. All three unstandardized regression coefficients in the original Flesch–Kincaid equation were outside their respective confidence intervals in the remodeled equation, suggesting that each would change significantly if Bormuth's sample passages and mean cloze scores became the criterion variable for a revised Flesch–Kincaid text-assessment tool. Estimating new parameters for the Flesch–Kincaid equation in order to predict Bormuth's criterion variable reduced the unstandardized regression coefficients for both Sylls/Wd and Wds/Sent, but more for Wds/Sent. The coefficient for Sylls/Wd when predicting Bormuth's cloze scores was 82% of what it had been in Kincaid et al.'s equation; the coefficient for Wds/Sent was 41% of what it had been. When predicting Bormuth's criterion variable, word complexity makes a larger relative contribution than sentence complexity compared with the relative contributions of the two factors in the original Flesch–Kincaid statistical modeling.

For the Lexile tool, we started with the published partial equation (Stenner & Fisher, 2013), which provides the unstandardized regression coefficients for the two predictors, mean log word frequency (MLWF) and log mean sentence length (LMSL), but not for the constant (except for its sign):

$$\begin{aligned}\text{Constant:} &\quad \beta = -?\\ \text{MLWF}: &\quad \beta = -2.14634\\ \text{LMSL}: &\quad \beta = 9.82247\end{aligned}$$

Furthermore, we are given the formula for converting the predicted text difficulty in logits as yielded by the equation to the Lexile scale: ((logit + 3.3) * 180) + 200.

Table 1 Predictors of Bormuth's mean cloze scores transformed to the Flesch–Kincaid grade level scale

Variable	B	95% CI
Constant	− 8.41*	[− 10.22, − 6.60]
Mean syllables per word	9.62*	[8.24, 11.00]
Mean words per sentence	0.16*	[0.13, 0.19]
R^2	.60	
F	241.70*	

$N = 330$. CI, confidence interval

*$p < .001$

To examine the difference that using Bormuth's criterion variable would make to parameter estimates in the Lexile equation, we conducted a multiple regression analysis using Lexile's two predictors, MLWF and LMSL, to predict Bormuth's mean cloze scores for his 330 sample passages. This decision required three data transformations.

First, the Lexile Analyzer provides two output scores for each passage: MLWF and MSL. Because the Lexile equation requires LMSL instead of MSL as a predictor, we first transformed the MSL scores for all passages to their natural (base *e*) log equivalents. For the second transformation, we used the formula provided to convert the Lexile scores for Bormuth's 330 passages back to logits. Third, we transformed Bormuth's mean cloze scores for all passages to their natural log equivalents to place them on the logit scale and added the same logit amount to each passage logit score so that the mean logits would be the same for both the Lexile logits and our transformed mean cloze scores.

The multiple regression analysis to answer this question used Lexile's two predictors, MLWF and LMSL, to predict Bormuth's mean cloze scores transformed to logits, as shown in Table 2. Both unstandardized regression coefficients for the predictors in the Lexile equation are outside their respective confidence intervals in the remodeled equation, suggesting that each would change significantly if Bormuth's sample passages and mean cloze scores became the criterion variable for a revised Lexile text-assessment tool. The coefficient for MLWF when predicting Bormuth's cloze scores transformed to logits was 63% of what it had been in Lexile's equation; the coefficient for LMSL was 6% of what it had been. When predicting Bormuth's criterion variable, word complexity makes a larger relative contribution than sentence complexity compared with the relative contributions of the two factors in the original Lexile statistical modeling.

This change in parameter estimates for the Lexile Framework could be compared more directly than for the Flesch–Kincaid because both Lexile predictors lay on the

Table 2 Predictors of Bormuth's mean cloze scores transformed to logits

Variable	B	95% CI
Constant	− 1.99*	[− 2.75, − 1.23]
Mean log word frequency	1.35*	[1.16, 1.54]
Log mean sentence length	− 0.60*	[− 0.69, − 0.52]
R^2	.61	
F	250.22*	

$N = 330$. CI, confidence interval

*$p < .001$

same metric (i.e., the natural log scale). This means that a ratio could be taken of the two unstandardized regression coefficients in both the original and remodeled equations. In the Lexile equation, the ratio of the MLWF and LMSL coefficients was − .22; it became − 2.25 in the remodeled equation.

Research question 3: How well does each text tool predict text difficulty within individual grade bands?

For this question, we compared how well each text tool predicted Bormuth's mean cloze scores within each of the five grade bands of use of the 330 texts. The Flesch–Kincaid worked best when predicting Bormuth's mean cloze scores for passages from the University grade band (see Table 3.) The correlation of − .76 within this grade band was the same as it had been across the full range of grade bands (see question 1). The Flesh–Kincaid tool did not approach its ability to predict across the full range of grades (1-University) within any other grade band of use. The grade band with the second highest correlation ($r = -.54$; $r^2 = .29$) was the 1–3 grade band (see Table 3).

The Lexile tool also worked best for predicting Bormuth's mean cloze scores within the University level (see Table 4). Its Pearson correlation of − .78 within that grade band was comparable with its correlation of − .70 across all the grade bands (see question 1). As with the Flesch–Kincaid, the Lexile tool did not approach its ability to predict across the full range of grades within any other grade band of use. The grade band with the second highest correlation ($r = -.51$; $r^2 = .26$) was the 1–3 grade band (see Table 4).

Post-hoc analyses

After the analyses to answer question 3 showed that the University grade band was the only one within which either Lexile or the Flesch–Kincaid tool predicted mean cloze scores nearly as well as for the full range of grade bands, we conducted a post hoc analysis with implications for both questions 1 and 3.

To answer question 1, we had correlated each tool's predictions with passages' mean cloze scores across the full grade range from 1st grade through University. In this post hoc analysis, we reran the Spearman rank order correlation analysis (*rho* = − .74) and the simple regression between the Flesch–Kincaid tool and Bormuth's mean cloze scores without the University passages. The correlation for

Table 3 Correlations between Flesch–Kincaid and mean cloze scores within each grade band (means and standard deviations included)

Grade band of use	r	r^2	n	Mean	SD
1–3	− .54	.29	75	4.1	2.7
4–6	− .42	.18	72	6.6	2.7
7–9	− .44	.19	80	8.5	2.7
10–12	− .45	.21	70	11.8	3.6
University	− .76	.57	33	12.9	4.0

Table 4 Correlations between Lexile and mean cloze scores within each grade band (means and standard deviations included)

Grade band of use	r	r^2	n	Mean	SD
1–3	− .51	.26	75	691	291
4–6	− .41	.17	72	932	217
7–9	− .26	.07	80	1034	208
10–12	− .29	.08	70	1235	223
University	− .78	.61	33	1281	308

the 297 passages from grades 1–12 ($r = - .71$; $r^2 = .51$) only declined a small amount from what it had been with all 330 passages ($r = - .76$; $r^2 = .57$).

Second, we reran the Spearman rank order correlation (*rho* = − .65) and the simple regression between Lexile and mean cloze scores without the University passages. The correlation for the 297 Grade 1–12 passages ($r = - .65$; $r^2 = .43$) declined a small amount from what it had been with all 330 passages ($r = - .70$; $r^2 = .49$).

With respect to question 3, these post hoc analyses provide two additional but provisional findings. First, each tool appears to gain significant predictive power when parameters are estimated over a wide range of grades. The correlation for all 297 grade 1–12 passages was much higher than for any individual grade band from 1 to 12 (see Tables 3, 4). Second, the markedly stronger predictive power of both tools within the University grade band is largely specific to that grade band. The removal of the 33 passages from that grade band did not greatly diminish the ability of either tool to predict the difficulty of the remaining 297 passages.

Research question 4: How accurately does each text tool predict the grade bands of Bormuth's sample passages?

Unlike the first three questions that used the mean cloze scores for sample passages, the criterion variable for the analyses for question four was the grade band of use from which Bormuth (1969) had originally sampled each of the 330 texts. While our focus was on aggregated student comprehension performance as the gold standard for a criterion variable, Bormuth's original grade bands of use for the 330 passages provided an opportunity to investigate that conclusion.

We first performed a simple regression of each tool's estimates for the 330 passages onto the original grade-bands of use, numbered 1–5. Then, we rounded each tool's predicted grade-band of use (Y-hat symbol) to its whole grade band. Finally, we computed the accuracy of each tool at placing each sample passage within its original grade band of use.

The Flesch–Kincaid tool did best with passages from grade bands 1–3 and 4–6, classifying two thirds of them correctly. Its accuracy dropped significantly—to a quarter or less—with passages from grade bands 7–9 and 10–12 At the University level, 97% of the passages were classified into a lower grade band. In all three grade bands above grades 4–6, the errors overwhelmingly resulted in texts from higher grade-bands being classified into lower ones (see Table 5).

Table 5 Percent accuracy of the Flesch–Kincaid's classifications of Bormuth's sample passages into their grade bands of use

Grade band of use	Predicted grade band of use		
	Same (%)	Lower (%)	Higher (%)
1–3	66.7	0.0	33.3
4–6	66.7	20.8	12.5
7–9	25.0	68.8	6.3
10–12	21.4	75.7	2.9
University	3.0	97.0	0.0
Overall	40.6	47.0	12.4

N = 330. Rounding errors prevent some row totals from equaling 100.0%

Table 6 Percent accuracy of Lexile's classifications of Bormuth's sample passages into their grade bands of use

Grade band of use	Predicted grade band of use		
	Same (%)	Lower (%)	Higher (%)
1–3	53.3	8.0	38.7
4–6	55.6	16.7	27.8
7–9	40.0	58.8	1.3
10–12	14.3	85.7	0.0
University	3.0	96.9	0.0
Overall	37.3	47.6	15.2

N = 330. Rounding errors prevent some row totals from equaling 100.0%

The Lexile tool also did best with passages from grade bands 1–3 and 4–6, classifying between half and three-fifths of them correctly. Its accuracy dropped somewhat when classifying passages from the 7–9 grade band (two-fifths) and much more when classifying passages from the 10–12 grade band (about 14%). Like the Flesch–Kincaid tool, Lexile classified 97% of the passages from the University grade band into a lower one. In all grade bands above 4–6, the errors Lexile made overwhelmingly resulted in texts from higher grade-bands being classified into lower ones (see Table 6).

Discussion

Limitations

This study used Bormuth's (1969) criterion variable to examine the validity of two widely used quantitative text assessment tools. An ideal criterion would have significantly longer passages than Bormuth's. However, the finding of an independent correlation of .98 between mean cloze scores on 32 of his passages and their longer versions may mean that passage length was not a limiting factor.

Another possible limitation of Bormuth's (1969) criterion is the absolute difficulty of its passages. School texts, at least those in grades one and three (Gamson, Lu, & Eckert, 2013), appear to have increased in complexity since the 1960s. Therefore, first- through third-grade texts in Bormuth's sample may not represent current primary-level texts. However, since text complexity beyond the primary grades appears to have remained fairly stable over past decades (Gamson et al.), higher-level texts in the Bormuth criterion may be relatively representative of current texts. It is also the case that the texts Bormuth sampled from the 1–3 grade band were relatively easier than texts from his 4–6 grade band.

Statistical procedures for constructing reading comprehension assessments have changed since Bormuth (1969) developed his criterion variable. Rasch modeling was used in constructing the tests of reading comprehension for the DRP and Lexile Framework, a procedure not available to Bormuth who relied on classical test theory. Conversely, Briggs (2013) and Domingue (2014) have argued that comprehension assessment data, especially the data underlying the Lexile Framework, may not fit the Rasch model, because the latent variable lacks the quantitative structure necessary for equal-interval scaling. The cautions of Briggs and Domingue suggest that classical test theory may satisfy quantitative data structure requirements for reading assessments more aptly than Rasch modeling. Within these potential limitations, the results of our study led us to draw several conclusions.

Validity needed to use quantitative text tools for today's high-stakes purposes

Since implementation of the CCSS (NGACBP & CCSSO, 2010b), text tools now serve as sources for interpretation of student growth, construction of assessments, selection of texts for instructional programs, and creation of text sets. These changes reflect recommendations by CCSS writers: (a) a call for texts to increase in difficulty in Appendix A (NGA & CCSSO, 2010a) and (b) the indexing of other quantitative text tools to the higher Lexile grade bands (Nelson et al., 2012). The combination of redefining text grade level higher for all grade bands and most text tools and reversing the purpose of text tools from setting a ceiling of difficulty for students to setting a floor means that using a quantitative tool to analyze a text has become a decision with high stakes for students and their teachers.

Logically, validation coefficients of quantitative text tools would be expected to be substantially higher than previously to justify their current use for high-stakes decisions. Historically, prior to the mid-1960s, the correlation of .70 (Klare, 1963) between predictors in readability formulas and their criterion seemed sufficient to justify their use to aid teachers in reader-text and class-text match decisions. When readability researchers began to use traditional cloze rather than multiple-choice questions as the criterion measure, correlations between predictors and criterion increased (Miller, 1975), usually into the .80–.89 range (Klare, 1974). For these two reasons, in this study we expected each text tool to predict Bormuth's criterion variable with a correlation of at least .80. This anticipated level was not attained by either the Lexile Framework ($r = .70;\ r^2 = .49$) or the Flesch–Kincaid ($r = -\ .76$; $r^2 = .57$). These results suggest that the two widely used tools, especially Lexile, may lack adequate validity for their current high-stakes uses in schools.

This conclusion was reinforced by the finding that each tool's accuracy in assigning a text to its original grade band declined in higher grade bands and, when a tool missed, a higher-level text was typically assigned to a lower grade band. If current texts have the same relative differences between grade bands as in Bormuth's 330 passages, validation coefficients that fall below .80 may understate the problem. Misses may disproportionally increase the challenge of texts at all grade bands other than the highest one. These findings lead us to caution users against over-relying on these text tools' estimates of text difficulty.

Validity of quantitative tools for analyzing texts within a single grade band

Results from research question 3 in this study caution against uses of quantitative tools that involve assigning texts to a specific grade or an order for teaching or testing within any grade band below University. The best performance for either the Flesch–Kincaid or the Lexile tool below University level was within Bormuth's grade 1–3 band, accounting for an unimpressive 29 and 26% of variance in mean cloze scores, respectively. Performance for each tool declined from that point through the higher grade bands below University. When teachers in grades 1–12 need to select books for students at their precise reading levels or rank texts from easiest to hardest for use, these two text tools appear to provide little accuracy.

Historically, readability formulas gained much of their predictive power from having their parameters estimated using a criterion variable with texts sampled from a wide range of difficulty (Klare, 1974). In a study of the contrary case, Rodriguez and Hansen (1975) re-estimated the parameters in Bormuth's readability formulas using criterion data only from seventh graders reading seventh-grade materials. The validation coefficients for the formulas dropped to around .45 ($r^2 = .20$). These patterns suggest that different quantitative text tools are needed for specific grade bands. In the most obvious historical example, Spache (1953) developed a formula for application only to primary-grade texts.

Effects of criterion measures with a multiple-choice response mode on the validation of text tools

The results of this study suggest that, when developing or selecting a criterion variable to validate a quantitative text tool, the participants' response mode may matter. As Klare (1974) and Miller (1975) reported, criterion variables consisting of traditional cloze scores yield higher validation coefficients (usually above .80) than criterion variables consisting of multiple-choice comprehension test scores (usually near .70). Since that time, modified cloze with a multiple-choice response mode has replaced traditional cloze.

There are two potential problems with a multiple-choice response mode as a criterion variable in text complexity research. First, unless distractors are chosen randomly from a well-defined set of possibilities, multiple choice introduces test-constructor variance into a measure. Second, distractors—including randomly chosen ones—can systematically reduce the rigor with which text comprehension is measured. For example, if even a single distractor in most items on a multiple-choice assessment with deletions is inconsistent with the syntax of the text, odds are increased that participants will get those items right solely based on syntactic information. Similarly, if the correct answer is the only choice that collocates with words surrounding the blank in the text, odds are increased that participants may get items right based on meaning vocabulary knowledge rather than text comprehension. The relatively weak validation coefficients of Bormuth's criterion with the Flesch–Kincaid and Lexile tools in this study may be partly due to such false positives in their original criteria.

Word complexity may be underweighted relative to sentence complexity

The focus in this study on the criterion side of the text complexity/difficulty relationship meant that we did not modify the predictors in either tool. However, when we used Bormuth's criterion for predicting text levels, both the Lexile and Flesch–Kincaid formulas required substantial changes in parameter estimates to optimize predictions. In both cases, these changes increased the weighting of the word complexity factor relative to the sentence complexity factor. This finding of our study has a clear implication: Either the original criteria for the two tools were more valid than Bormuth's as a measure reflecting the modal balance of semantic and syntactic factors in text comprehension or they were less valid. Historic readability research (Klare, 1974) as well as research on the role of meaning vocabulary knowledge in word reading and comprehension (e.g., Ouellette, 2006) would support a greater weighting of the semantic, but additional research on the issue is needed for today's readers and texts.

Need for a new generation of criterion variables

Much is currently known about reading comprehension, educational measurement, and text complexity. Yet, the criterion variables used to estimate the parameters of the three most widely used quantitative text tools (Kincaid et al., 1975; Koslin et al., 1987; Stenner, 1996; Stenner & Burdick, 1997) were all developed several decades ago. The most meticulously developed criterion variable was developed almost 50 years ago (Bormuth, 1969). For text research to move meaningfully beyond the state of the art achieved last century, we need to harness the advantages of our digital age to craft a new generation of criterion variables based on today's questions and uses of text tools. For example, it seems probable that a technology could be developed to score a test with a traditional cloze or other deletion format reasonably accurately, making the ease and economy of scoring that has favored multiple-choice formats less relevant.

Hopefully, the impressive rigor and precision that Bormuth employed 5 decades ago to develop his criterion variable will inspire and guide today's text researchers. Perhaps the criterion variables in future text research will progress to match current goals, measurement tools, and knowledge, and in so doing, help significantly improve understanding of the texts students read in school.

References

Bormuth, J. R. (1969). *Development of readability analyses* (Final Report, Project No. 7-0052, Contract No. OEC-3-7-070052-0326). Retrieved from ERIC database. (ED029166).

Bormuth, J. R. (1971). *Development of standards of readability: Toward a rational criterion of passage performance* (Final Report, Project No. 9-0237, Contract No. OEG-0-9-230237-4125(010)). Retrieved from ERIC database. (ED054233).

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement, 50,* 204–226. https://doi.org/10.1111/jedm.12011.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American Heritage Word Frequency Book*. Boston, MA: Houghton Mifflin.

Coleman, E. B. (1965). *ON understanding prose: Some determiners of its complexity* (NSF Final Report GB-2604). Washington, DC: National Science Foundation.

Cunningham, J. W., & Mesmer, H. A. (2014). Quantitative measurement of text difficulty: What's the use? *The Elementary School Journal*, *115*, 255–269.

Dale, E., & Tyler, R. W. (1934). A study of the factors influencing the difficulty of reading materials for adults of limited reading ability. *The Library Quarterly: Information, Community, Policy, 4,* 384–412.

Dolch, E. W. (1948). Graded reading difficulty. *Problems in Reading* (pp. 229–255). Champaign, IL: The Garrard Press.

Domingue, B. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika, 79,* 1–19.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test, Revised: Forms L and M*. Circle Pines, MN: American Guidance Service.

Dunn, L. M., & Markwardt, F. C. (1970). *Peabody Individual Achievement Test*. Circle Pines, MN: American Guidance Service.

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

Gamson, D. A., Lu, X., & Eckert, S. A. (2013). Challenging the research base of the common core state standards a historical reanalysis of text complexity. *Educational Researcher, 42,* 381–391.

Gray, W. S., & Leary, B. E. (1935). *What Makes a Book Readable: With Special Reference to Adults of Limited Reading Ability*. Chicago, IL: University of Chicago Press.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12,* 281–300.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel* (No. RBR-8-75). Millington TN: Naval Technical Training Command Research Branch.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95,* 163–182.

Klare, G. R. (1963). *The Measurement of Readability*. Ames, IA: Iowa State University Press.

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly, 10,* 62–102.

Klare, G. R. (1984). Readability. In P. D. Pearson, R. Barr, M. L. Kamil, & P. Mosenthal (Eds.), *Handbook of Reading Research* (Vol. 1, pp. 681–744). New York, NY: Longman.

Koslin, B. I., Zeno, S., & Koslin, S. (1987). *The DRP: An Effective Measure in Reading*. New York, NY: College Entrance Examination Board.

Lively, B. A., & Pressey, S. L. (1923). A method for measuring the vocabulary burden of textbooks. *Educational Administration and Supervision, 9,* 389–398.

Lorge, I. (1939). Predicting reading difficulty of selections for children. *Elementary English Review, 16,* 229–233.

McCall, W. A., & Crabbs, L. M. (1925, 1950, 1961). *Standard Test Lessons in Reading: Teacher's Manual for All Books*. New York, NY: Teachers College Press.

McNamara, D. S., & Kintsch, W. (1996). Learning from text: Effect of prior knowledge and text coherence. *Discourse Processes, 22,* 247–288.

Mesmer, H. A. E. (2008). *Tools for matching readers to texts: Research-based practices*. New York, NY: Guilford Press.

Mesmer, H. A., Cunningham, J. W., & Hiebert, E. H. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading Research Quarterly*, *47*, 235–258.

Metametrics. (2017a). *Statewide assessments*. Retrieved from https://lexile.com/about-lexile/How-to-get-lexile-measures/states/.

Metametrics. (2017b). *State consortia*. Retrieved from https://lexile.com/using-lexile/lexile-measures-and-the-ccssi/state-consortia/.

Miller, L. R. (1975). Predictive powers of multiple-choice and cloze-derived readability formulas. *Reading Improvement, 12,* 52–58.

Muijselaar, M. L., Kendeou, P., de Jong, P. F., & van den Broek, P. W. (2017). What does the CBM-Maze test measure? *Scientific Studies of Reading, 21,* 120–132. https://doi.org/10.1080/10888438.2016.1263994.

Muijselaar, M. M. L., Swart, N. M., Steenbeek-Planting, E. G., Droop, M., Verhoeven, L., & de Jong, P. F. (2017). The dimensions of reading comprehension in Dutch children: Is differentiation by text and question type necessary? *Journal of Educational Psychology, 109*(1), 70–83. https://doi.org/10.1037/edu0000120.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010a). *Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects*, Appendix A. Washington, DC: Author. Retrieved from www.corestandards.org/assets/Appendix_A.pdf.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010b). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects.* Washington, DC: Authors. Retrieved from www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf.

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of Text Difficulty: TESTING their Predictive Value for Grade Levels and Student Performance*. New York, NY: Student Achievement Partners.

Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology, 98,* 554.

PARCC (n.d.). *ELA test specifications documents: Understanding summative assessment design.* Retrieved from: http://www.parcconline.org/assessments/test-design/ela-literacy/ela-performance-level-descriptors.

Pearson, P. D., & Hiebert, E. H. (2014). The state of the field: Qualitative analyses of text complexity. *The Elementary School Journal*, *115*, 161–183.

Rodriguez, N., & Hansen, L. H. (1975). Performance of readability formulas under conditions of restricted ability level and restricted difficulty of materials. *Journal of Experimental Education, 44,* 8–14.

Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly, 17,* 229–255.

Smarter Balanced (November 17, 2016). *Smarter Balanced, MetaMetrics partner to provide more specific info about students' reading abilities.* Retrieved from: http://www.smarterbalanced.org/smarter-balanced-metametrics-partner-provide-specific-info-students-reading-abilities/.

Spache, G. (1953). A new readability formula for primary grade reading materials. *The Elementary School Journal, 53,* 410–413.

Stenner, A. J. (1996). *Measuring reading comprehension with the Lexile Framework*. Paper presented at the North American Conference Adolescent/Adult Literacy (4th, Washington, DC, February). Retrieved from ERIC database (ED435977).

Stenner, A. J., & Burdick, D. S. (1997). *The objective measurement of reading comprehension: In response to technical questions raised by the California Department of Education Technical Study Group*. Washington, DC: IES/ERIC. Retrieved from ERIC database (ED435978).

Stenner, A. J., & Fisher, W. P., Jr. (2013). Metrological traceability in the social sciences: A model from reading measurement. *Journal of Physics: Conference Series, 459*, 1–6. Retrieved from http://iopscience.iop.org/1742-6596/459/1/012025.

Stenner, A. J., Smith, M., & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement, 20,* 305–316.

Stevens, K. C. (1980). Readability formulae and McCall-Crabbs standard test lessons in reading. *The Reading Teacher, 33,* 413–415.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Bulletin, 30,* 415–433.

Vogel, M., & Washburne, C. (1928). An objective method of determining grade placement of children's reading material. *The Elementary School Journal, 28,* 373–381.

Terms and Conditions