Investigating the Validity of Two Widely Used Quantitative Text Tools

James W. Cunningham
University of North Carolina at Chapel Hill, USA

Elfrieda H. Hiebert
TextProject &
University of California, Santa Cruz

Heidi Anne Mesmer
Virginia Tech, Blacksburg, USA

TextProject Article Series
July 2018

TextProject, Inc.
SANTA CRUZ, CALIFORNIA

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.

Investigating the Validity of Two Widely Used Quantitative Text Tools

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" (p. 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, 1,000 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, and 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 3,000 students in grades 2 to 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 3,000 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)