Reading in a Foreign Language October 2023, Volume 35, No. 2
ISSN 1539-0578 pp. 106–137

A Unitary Measure of L2 Silent Reading Fluency Accounting for Comprehension

Steven J. Carter
Brigham Young University-Hawaii
United States

Matthew P. Wilcox
Brigham Young University
United States

Neil J. Anderson
Brigham Young University-Hawaii
United States

Abstract

This research presents a novel reading fluency (*rf*) measurement formula that accounts for both reading rate and comprehension. Possible formulas were investigated with 68 participants in a strategic reading course in an IEP at a small Pacific Island university. The selected formula's scores demonstrated concurrent validity through strong correlation ($r[66] = .680$, $p < .001$) with the Adaptive Reading Test (ART), an assessment aligned with ACTFL's proficiency levels. Furthermore, when ART scores were regressed onto formula scores, formula scores accounted for 49% of the variance in ART scores ($R^2 = .488$, $F[1, 66] = 62.88$, $p < .001$); these results were comparable to a model in which comprehension and rate were the independent variables ($R^2 = .514$, $F[2, 65] = 34.38$, $p < .001$). The formula appears preferable to currently available alternatives and ensures that high performance in reading rate cannot compensate for low performance in comprehension nor vice versa. An Excel workbook for exploring formula variants and tracking learners' fluency is provided to readers of *Reading in a Foreign Language*.

Keywords: reading measurement, silent reading fluency, reading rate, reading speed, reading assessment, reading comprehension, second language reading

L2 learners need to attain some degree of English silent reading fluency in academic contexts. Reading fluently directly influences all reading activities and much of academic performance. As Grabe (2009, p. 290) stated:

> Fluency is what allows a reader to experience a much larger amount of L2 input, to expand the breadth and depth of vocabulary knowledge beyond direct instruction, to develop automatic word-recognition skills, to read for additional learning, to build reading motivation, and, in L2 university contexts, to read the large amounts of material that might be assigned every week. Moreover, fluency is one of the keys to L2 learning outside the classroom. Students who have some degree of reading fluency and who are motivated to develop fluency further will most likely be engaged in a continual L2 learning environment.

Fluent reading skills are clearly favorable to L2 students' academic success. Consequently, monitoring and assessing fluency should have some priority in programs' efforts to monitor and assess students' academic English development. However, current reading fluency measurement practices that reasonably account for comprehension remain somewhat rudimentary.

In a recent study, Kramer and McLean (2019, p. 202) argued for the need for "increased accuracy and precision in reading fluency measurement within second language (L2) reading research". The current research aims to help facilitate this ideal—this report presents the rationale and method that resulted in a proposed measure of L2 silent reading fluency, a measure which:

(a) accounts for comprehension,
(b) reflects the complexity of the interaction between rate and comprehension, and
(c) does not allow high performance in rate to compensate for low performance in comprehension.

Literature Review

Defining Silent Reading Fluency

Silent reading fluency can be partially described in terms of the component skills at its core: word-reading efficiency, vocabulary development, text-reading ease, and reading with comprehension, amongst others (Grabe, 2009). According to L1 researchers, many concurrent processes characterize fluent reading; when bottom-up processes (e.g., word and semantic decoding) are more automatized and accurate, cognitive resources are freed for top-down, integrative comprehension processes (Fuchs et al, 2001, p. 242; see also Breznitz, 2006; LaBerge & Samuels, 1974; Stanovich, 2000). To be concise, silent fluent reading could be described as efficient and accurate processing of text resulting in tenable inferences and comprehension (Breznitz, 2006; Grabe & Stoller, 2020; LaBerge & Samuels, 1974). Put more simply, it is the ability "to read at an appropriate rate with adequate comprehension" (Anderson, 2018, p. 2; see also Anderson, 1999a, 2008).

Comprehension and Fluency

Various hypotheses have been made about the nature of the relationship between reading comprehension and fluent or rapid reading. Whereas some past L2-focused scholars have viewed fluent reading and comprehension as "competing factors in L2 performance," others have concluded the opposite, that fluent reading builds chunking, promoting accuracy and

comprehension (Grabe, 2010, p. 76). In recent L1 literature on both silent and oral reading fluency for varying age groups, there is repeated concurrence on a reciprocal relationship between fluency (or indicators of automaticity) and comprehension (Fuchs et al, 2001; Klauda & Guthrie, 2008; National Reading Panel, 2000; Paige, 2011; Pikulski & Chard, 2005). Paige et al (2014) posited that "indicators of fluent reading" work conjointly with comprehension during effective reading. They proposed a Tandem Theory of reading, where "automaticity is optimized by the reader. … [working] on a bi-directional basis in tandem with, and for the purpose of, maximizing comprehension" (p. 145). This relates to Carver's (2000) theory of rauding, where learners comprehend most efficiently at their optimal reading rates.

The literature and the definitions cited to this point all emphasize that comprehension is requisite evidence of fluent reading (Grabe & Stoller, 2020). Indeed, "the simultaneity of decoding and comprehension … is the essential characteristic of reading fluency" (Samuels, 2007, p. 564). Reading rapidly with little understanding, on the other hand, is contradictory and meaningless (Samuels, 2007; Zwick, 2018); it is no better than very slow reading with high comprehension. Indeed, ideal development in reading fluency necessitates increasing in rate while maintaining or improving comprehension (Zwick, 2018). This is pertinent to the L2 context because high reading rates can sometimes come "at the expense of comprehension or the other way around" (Chang & Millet, 2013, p. 128; see also Chang, 2014; Gorsuch & Taguchi, 2008, 2010).

Despite the importance of accounting for adequate comprehension when measuring fluency (McLean, 2014), sometimes it is simply glossed over or ignored (Chung & Nation, 2006; Lynn, 2021; Macalister, 2008, 2010; Taguchi, 1997). Fortunately, some researchers and practitioners do control for comprehension by ensuring readers meet a minimum criterion score on a post-hoc comprehension check (Kramer & McLean, 2019; Quinn et al, 2007). Yet, this approach may be impractical when monitoring students' progress in real settings where classes with mixed ability levels are common. For instance, opting to use texts for which a minimum level of comprehension is achievable by all perhaps caters to the low end of the ability spectrum and inflates students' rate by presenting them with relatively easy material[1]. Even when the minimum criterion is achievable by most, it is still problematic to evaluate performance for those who fall below that standard. For instance, in a study on the effect of Extensive Reading on rate, McLean and Rouault (2017) used the mean score of participants (72.1%) as a control variable to establish adequate comprehension for pre- and post-treatment Timed Reading (TR) passages. Though the mean performance exceeded the 70% threshold suggested in the literature (Anderson, 1999a, 1999b, 2003, 2006, 2008, 2018; Kramer & McLean, 2019; Nation, 2005; Nuttall, 2007; Quinn et al, 2007), the standard deviation was 10.9%. In a separate study, Beglar et al (2012) used 75% as their criterion; the mean score for their participants was roughly 85%, but the standard deviation was roughly 9%. Clearly, some participants' scores fell below the standard in these studies. The legitimacy of these students' fluency gains is questionable. As explained above, processing with understanding gives credibility to reading rate, whereas poor processing and poor understanding renders readers' rate something of a façade or pretense.

[1] We acknowledge that the content of timed- or speed-reading programs should not be overly difficult and that comprehension should not be overemphasized as it can lead learners to fixate on comprehension and lose sight of their purpose—to increase their rate (Quinn et al, 2007). Our primary concern is accurate and defensible measurement.

Accounting for Comprehension in a Measure of L2 Silent Reading Fluency

A measure of silent reading fluency that more accurately accounts for comprehension could address problems with current practices, allowing for more effective monitoring of individual students' progress in mixed-level classes. Furthermore, it would allow for both teachers and learners to more easily quantify individual learners' various TR performances relative to their past performances. This research attempts to improve on limited current practices that already attempt to account for comprehension when measuring reading fluency.

Surprisingly, there is little precedent for a more precise measure. In some studies, measures factoring in reading accuracy or understanding with word-reading speed are used (e.g., see Aro et al, 2018). Two better-known examples are the fourth edition of the Gray Oral Reading Test (GORT-4; Bryant et al, 2009) and the Woodcock-Johnson Sentence Reading Fluency Test (WJ-IV; Schrank et al, 2014). For these measures, errors (in word reading accuracy or understanding) factor into scores through simple addition and subtraction. Yet this approach is unlikely to reflect the complex interaction between comprehension and rate (Carver, 2000; Paige et al, 2014; Zwick, 2018) that occurs during silent reading. Furthermore, these measures were largely designed for children and adolescents. For instance, the WJ-IV Sentence Reading Fluency Test asks readers to indicate whether simple sentences (i.e., "A cow can dance") are true or false. This comprehension task does not require the layered understanding necessary with more complex academic texts.

Principles to consider when constructing a measure. Before proceeding, we emphasize that consideration of the context and purpose for reading is paramount in any attempt to quantify silent reading fluency (Carver, 1990). This research is focused on L2 learners in a university setting.

With this context in mind, it is helpful to consider constructing a formula for which performances are valued according to their proximity to theoretically justified optimal thresholds for rate and comprehension; these are performance levels beyond which gains are less consequential or likely. There is support for this in extant L1 literature on rate with repeated references to an asymptote or peak in performance (Breznitz, 2006; Carver, 1990; Hudson et al, 2008; Kuhn et al, 2010; Paige et al, 2014). L2 literature also repeatedly references certain target rates for different contexts, generally ranging from 200 to 300 words per minute (wpm) (Anderson, 1999a, 1999b, 2003, 2006, 2008, 2018; Grabe & Stoller, 2020; Nation, 2009). The threshold could either be described as a rate: (a) around which most skilled readers tend to level off, having reached a functionally successful level of reading fluency, or (b) beyond which increases in rate while maximizing comprehension become difficult to attain. There is also reference to a comprehension threshold, primarily in L2 literature: Generally, 70% is considered sufficient in a TR exercise or fluency assessment scenario (Anderson, 1999a, 1999b, 2003, 2006, 2008, 2018; Nuttall, 2007). Though this threshold has not been empirically proven, logic lends some credibility to it: initial reading purpose moderates comprehension by guiding what readers attend to within a text (De Hoyos & David, 2018; Erten, 2018; Grabe, 2009), but when the purpose is to practice or demonstrate fluency, readers are not primed to attend to specific details and it is unsurprising for them to recall content imperfectly.

Another important consideration is that rate and comprehension's impact on one another and formula output scores should align with two logical principles of L2 silent reading fluency. The self-monitoring chart presented in Figure 1 is helpful in understanding these principles. Values on the *y*-axis represent wpm scores or learners' rate, whereas values on the *x*-axis represent TR comprehension scores. Anderson (2018) designates Quadrant #4 on the chart as the goal—when L2 learners' rate and comprehension are sufficient to situate their performance consistently in Quadrant #4, their reading fluency meets expectations as defined by Anderson. The quadrants on the chart are illustrative as we delineate principles with which a valid measure of silent reading fluency should align. See Figure 1.

	Quadrant #2					Quadrant #4			
400									
380									
360									
340									
320									
300									
280									
260									
240									
220									
200									
180									
160									
140									
120									
100									
80									
60									
40									
20									
wpm	Quadrant #1					Quadrant #3			
	0–29%	30%	40%	50%	60%	70%	80%	90%	100%
	comprehension								

Figure 1
Reading Fluency Chart adapted from Anderson (2018)

We might sum up the discussion in the following way:

Principle 1. For any measure used, changes in performance that represent direct movements toward the area of adequate silent reading fluency (i.e., Quadrant #4, Figure 1) should consistently translate to increases in measured fluency regardless of where readers' past performances fall on the chart.

Principle 2. Presuming a reader starts with both low rate and low comprehension (i.e., in Quadrant #1, Figure 1), for any measure used, increases in one variable alone would initially translate to increases in fluency. However, (once again, presuming both initial rate and

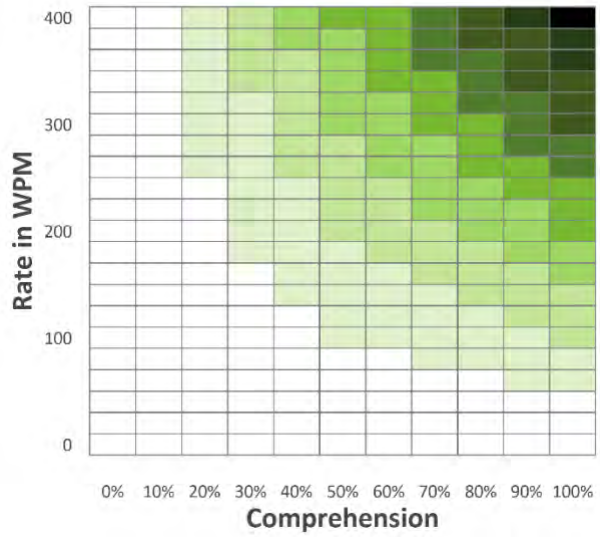
comprehension are low) continued increases in one variable without corresponding increases in the other would yield diminishing returns in fluency, eventually yielding no further increase in measured fluency.

If these two principles hold, then differing levels of fluency performance would collectively resemble a target with the bull's eye superimposed over Quadrant #4 in Figure 1. Successively lower levels would radiate away (down and to the left) from the area of adequate fluency similar to rings radiating away from a bull's eye. Regardless of where the exact thresholds on the chart are positioned, these principles should hold true.

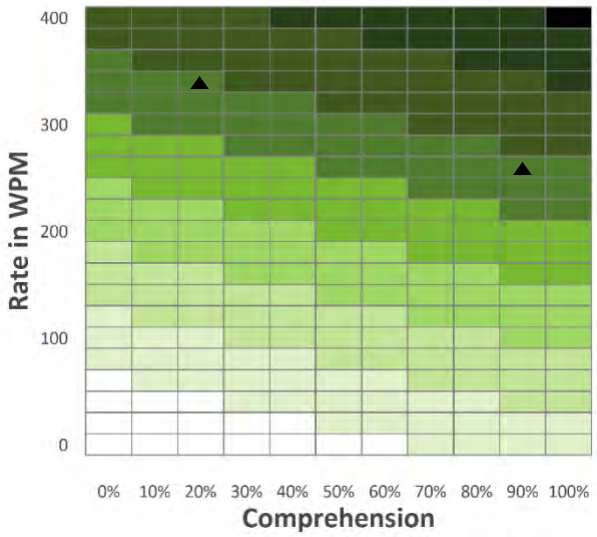
To expand on this and to clarify, we can consider the output of three relatively simple reading fluency formula possibilities in Figure 2. In each of the three examples below comprehension (*x*) is input as a decimal (e.g., 60% is input as .6) and reading rate (*y*) is input as a whole number (e.g., 100 wpm is input as 100). Using the format of the chart presented in Figure 1, Figure 2 shows how different combinations of comprehension and reading rates would be scored according to the different formulas. Subsequent darker green areas indicate successively higher learners scores similar to that of a heat map. The range of learner reading fluency score output values differs for each reading fluency formula, but rather than focusing on exact score values, we would ask you to note the patterns in scores exhibited by each reading fluency formula used.

On each chart, same-colored cells receive similar reading fluency scores. Obviously, reading fluency formulas (2, which is $100x + y$) and (3, which is $zscore_x + zscore_y$) violate principles #1 and #2 named above. For example, the two performances marked with triangles on the chart for formula (2), in Figure 2 would receive roughly the same score—i.e., using this formula, a reader who reads quickly (350 wpm) without comprehension (below 20%) would score as well as a reader who reads quickly (280 wpm) with high comprehension (95%). Of the three, reading fluency formula (1) aligns best with the principles, but for all three formulas, readers could simply read very quickly with lower understanding and score well (this is less apparent on formula [1]'s chart, but becomes obvious if the chart is extended upwards with higher y axis values—higher wpm values).

$$xy \quad (1)$$



$$100x + y \quad (2)$$



$$zscore_x + zscore_y \quad (3)$$

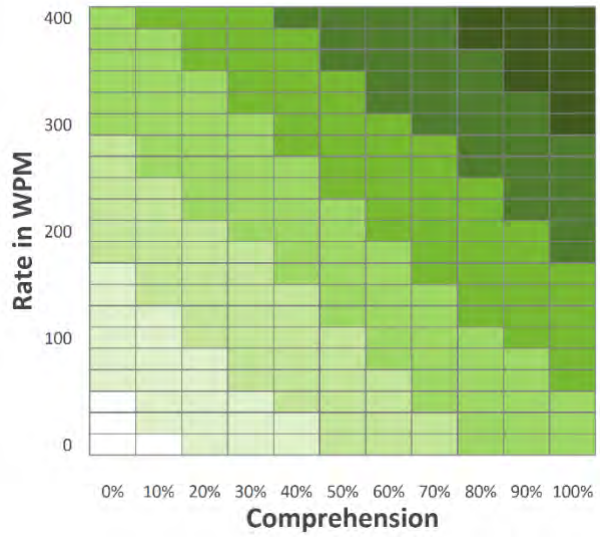


Figure 2
Simple Formula Possibilities

To potentially improve on formula (1) above and on current practice, we propose a new measure of silent reading fluency, *rf* (see [6]) below. The measure was designed to be flexible while aligning with the theoretical and measurement principles presented earlier. The primary variables in the formula are equated comprehension $e_{(x)}$ (noted as [4]) and equated rate $g_{(y)}$ noted as [5]). These variables each incorporate a respective threshold (t_x and t_y). See Appendix A for a detailed explanation of the construction and rationale for the formula. The balance of this report is dedicated to identifying a variant of formula (6) that produces scores well aligned with the principles, and with maximized correlation with scores from the Adaptive Reading Test (ART;

Clifford & Cox, 2013; Wilcox, 2020), a well-known reading proficiency test aligned with ACTFL's proficiency levels (ACTFL, 2012). High correlation with ART scores would give evidence of the formula's validity because silent reading fluency is a large contributor to general reading proficiency. The two should correlate highly with each other.

$$e_{(x)} = \begin{cases} x/t_x; x < t_x \\ 1 \ ; x \geq t_x \end{cases} \quad (4)$$

$$g_{(y)} = \begin{cases} y/t_y; y < t_y \\ 1 \ ; y \geq t_y \end{cases} \quad (5)$$

$$rf = \frac{e_{(x)} + g_{(y)}}{2} \quad (6)$$

Research Questions

This leads to the research questions for the current study:

1) For an optimal variant of formula (6):

    a) What target thresholds for comprehension and reading rate (t_x and t_y) are best for maximizing correlation between readers' averaged L2 silent reading fluency scores (*rf*) and their ART scores while aligning with theoretical principles?
    b) How should equated comprehension ($e_{(x)}$) and rate ($g_{(y)}$) mathematically impact one another to maximize correlation between readers' averaged L2 silent reading fluency scores (*rf*) and their ART scores while aligning with theoretical principles?

2) When averaged L2 silent reading fluency scores (*rf*) are regressed onto ART scores, how does the explained variance compare to that in a model where rate and comprehension scores are regressed onto ART scores?

3) Do *rf* scores grouped by ACTFL level (ART scores) form a unidimensional hierarchy?

4) How does *rf* compare to other possible measures of reading fluency including rate alone?

Method

Context and Participants

This research was conducted at a small university in the Hawaiian Islands with a diverse student population. Historically, the university has enrolled a high percentage of international students. In 2020, 44 percent of students at the institution were non-resident aliens. Matriculating students

with lower English proficiency are required to enroll in university Intensive English Program (IEP) courses until demonstrating a requisite level of academic English ability.

Of the 68 participants (48 females and 20 males) in this research, 63 were international students from multiple countries in Asia, Southeast Asia, the Pacific, and Central and South America. Also included were five native English-speaking students from the United States. All were convenience sampled from one TESOL Principles and Methods course and five intact sections of a Strategic Reading course. The Strategic Reading course was designed to prepare English language learners (ELLs) for university reading demands and was one of many electives available to IEP students. Some were assigned to the course because of poor performance in reading while others enrolled of their own accord. Tables 1 and 2 give detailed sampling and demographic information.

Table 1
Sampling Information

Course	Point of data collection	*n*	(*n* removed) Reasons for removal from study
Strategic Reading			
Section 1 (Fall)	end	16	(1) Suspicions of cheating
Section 2 (Fall)	end	15	(1) Incomplete data
Section 3 (Winter)	beginning	15	(2) Incomplete data
Section 4 (Fall)	beginning	11	(4) Incomplete data and incorrect procedure
Section 5 (Fall)	beginning	10	(2) Incomplete data and ambiguous marking of answer sheet
TESOL Course (Fall)	midway	14	(3) Incomplete data
		N = 81	13 total removed

Table 2
Demographic Information for Participants

Region of the world	Countries represented	*n*	% of sample
Asia	Japan (8), China (8), Mongolia (6), Taiwan (5), Hong Kong (4), South Korea (2)	33	48.5%
Pacific	Samoa (5), Tahiti (4), Western Samoa (1), Tonga (1), Solomon Islands (1), Kiribati (1), Fiji (1)	14	20.6%
Southeast Asia	Philippines (6), Papua New Guinea (2), Cambodia (1), Thailand (1), Myanmar (1), Indonesia (1)	12	17.6%
North America and South America	United States (5), Mexico (1), Guatemala (1), Honduras (1), Brazil (1)	9	13.2%

Note: *N* = 68

Materials

Timed Readings (TRs). Three TRs were administered to participants to assess their reading fluency. The TRs came from the *Advanced Reading Power 4 Test Booklet* (ARP4 TB; Jeffries & Mikulecky, 2014b). The readings were titled *Katherine Boo*, *Dads at Work*, and *Wikipedia*. Each focused on a thematically different academic topic and was roughly 1000 words in length. To adjust for variation in word length, we calculated text lengths and participants' reading rates in wpm using a standardized, 6-character based word count (Carver, 1990; Kramer & McLean, 2019). Flesch Reading Ease, Flesch-Kincaid Grade Level, and Coh-Metrix (Graesser et al, 2004) scores along with corpus frequency data for the three texts are presented in Table 3. The frequency analysis was conducted with Microsoft Excel and the Analyze Text tool available from the Corpus of Contemporary American English (Davies, 2015) at: english-corpora.org/coca. Though Wikipedia appeared to be somewhat more difficult than the other two texts based on these data, we did not feel the difference was great enough to compromise our intent.

No changes were made to the TR texts. However, we revised the multiple choice (MC) comprehension questions. From prior experience using the TRs, we were concerned about the reliability and validity of the unaltered MC items. For instance, some items were too easy and could be answered without attending to the reading. To improve them, we first determined that each set of items would consist of six items targeting understanding of details, two targeting global comprehension, and two requiring basic inference. This increased the number of items from the original eight to ten per TR. Next, we changed the number of MC options for each item from three to four. We then piloted the readings and accompanying MC items multiple times with various colleagues, all of whom were teachers in the university IEP. Changes were made to items based on their feedback.

Table 3
Readability and Lexical Complexity Data for TR Texts

Reading characteristics	TR Texts		
	Katherine Boo	*Dads at Work*	*Wikipedia*
Word count	1012	1011	1010
Standard word count	984.7	975.2	1025.5
Corpus frequency analysis			
High 1-500	72.4%	75.4%	65.4%
Mid 501-3000	11.0%	12.7%	16.2%
Low >3000	7.7%	8.2%	10.0%
Other	8.9%	3.8%	8.3%
Words per sentence (mean)	18.2	17.6	18.4
Flesch Reading Ease	59.8	62.5	46.4
Flesch-Kincaid Grade Level	9.4	8.9	11.4
Coh-Metrix	18.9	17.0	13.4

TR administration procedure. Participants who received the TRs at the beginning of the semester (Table 1) were told that the purpose was to determine their current level of reading fluency. They were instructed to read at a comfortable rate while maintaining comprehension.

Those who received the TRs at the end of the semester had completed weekly TRs throughout the semester. They had been told in advance that a small amount of extra credit would be applied to their grades if they demonstrated gains in reading fluency. The TESOL Principles and Methods course students received the TRs as part of a unit about teaching reading.

For beginning or mid-semester participants, to familiarize them with the procedure, they were given one practice TR prior to doing any of the three readings used for the assessment. Following the practice session, each assessment TR and its comprehension questions were completed at different points during a two-week period. Participants who received the TRs at the end of the semester also completed them in two weeks. They were already familiar with the procedure and did not receive a separate practice TR.

The small size of classes allowed for the TRs to be carefully proctored. The researchers distributed the TRs to participants in class, ensuring that no participant began reading until the timer was started. To track the time, an internet stopwatch was projected on a screen at the front of the class. We used Vclock (Vclock.com, 2022), a clean, unobtrusive internet stopwatch. As participants finished, they set the reading aside and recorded their time in minutes and seconds on an answer sheet. They then answered the MC questions which were on the front and back of a separate sheet of paper. The researchers observed to ensure participants did not read the questions before completing the reading and did not refer back to the reading while answering the questions. Any referring back to the reading would have required shuffling and flipping over sheets of paper which would have been easily noticed.

Adaptive Reading Test (ART). The ART is a computer adaptive, criterion-referenced test of reading proficiency used in institutions of higher education for placement, learning gains, and program evaluation (Wilcox, 2020). Results are aligned with the ACTFL Proficiency Guidelines (ACTFL, 2012), for Novice, Intermediate, Advanced, and Superior language abilities. Results from this test can easily be converted to proficiency levels on the Interagency Language Roundtable (ILR) Scale used by United States governmental agencies. Developed and maintained at Brigham Young University, the ART has been shown to be a reliable instrument for measuring English reading comprehension ($\alpha = 0.86$). More information concerning the evidence of validity for the ART can be found in its Technical Report (Wilcox, 2020). During the same two-week period in which participants received the TRs, they completed the ART in the university testing center.

Data Analysis

Due to the imbalance in male and female participants, we wanted to ensure there were no notable differences in performance related to gender. Table 4 presents descriptive statistics for comprehension and rate for each TR by gender, confirming that their performances were similar.

Table 4
Descriptive Statistics for all TRs by Participant Gender

TR	Male $n = 20$				Female $n = 48$			
	Comprehension		Rate		Comprehension		Rate	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Katherine Boo	72.5%	22.1%	165.4	51.0	72.9%	22.8%	156.2	38.1
Dads at Work	72.5%	16.0%	161.0	37.5	71.9%	22.7%	157.5	39.1
Wikipedia	56.3%	24.2%	186.2	36.2	58.6%	25.9%	163.7	47.2

Reliability. There were two ways of calculating reliability for the TRs. The first was to treat the three distinct sets of MC items collectively as a single test, calculating reliability for all 30 items together. There is some precedent for this approach in Beglar et al (2012). The second approach was to calculate reliability for each set of 10 MC items separately as a series of three testlets. We did both, but argue for the latter though it presents some disadvantages.

In the first scenario, where all 30 items were treated as one test, after three items were excluded from our data (one from each set of 10) due to poor item discrimination ($r_{\text{p-bis}} < .2$), Cronbach's α for the remaining 27 items was .77. Alternatively, when alpha was calculated for each set of 10 items separately, individual reliability for each set was notably lower, unsurprisingly. The values were .62, .52, and .60 for *Katherine* Boo, *Dads at Work*, and *Wikipedia* items respectively. For each of the three readings, two items were excluded[2] prior to calculating reliability due to poor discrimination ($r_{\text{p-bis}} < .2$).

One reason for favoring the second approach (treating each testlet separately) is that comprehension is heavily influenced by reading rate. When readers exceed a comfortable rate, comprehension can suffer. Because readers may intentionally vary their rate for different TRs (attempting to read faster for instance), it makes more sense to consider *rf* (factoring in both comprehension and rate) for each separate TR than it does to calculate *rf* from an average of three comprehension scores and an average of three rates.

In part, the low reliability of the distinct testlets was likely due to small numbers of items. However, it is also likely that items could have been further improved despite efforts to ensure their quality. In any case, the reliability of items intended to assess comprehension for TRs is not always investigated or reported and is seemingly poor in many cases (McLean, 2014) with rare exceptions (Huffman, 2014). To illustrate, prior to conducting this research we informally collected data on nine other TRs and their accompanying MC items from ARP4 textbook (Jeffries & Mikulecky, 2014a); data came from the same population from which participants were drawn. The number of test takers for each set of items ranged from 30 to 45. The average reliability for these eight-item assessments was .38 even after poorly discriminating items were excluded and all values for α were below .50 except for one ($\alpha = .58$). We have consistently found generally poor reliability on all other TR materials we have investigated and often times questionable validity for other reasons.

[2] When we re-calculated the reliability for the three sets of items as a collective set (one test), deleting the same six items that were deleted from the individual sets for individual reliability calculations (two items were deleted from each set making six items in total deleted rather than just three), alpha for the collective group of 24 items was still .77.

Refining and validating the measure. We intended to construct a measure in which equated comprehension ($e_{(x)}$) and rate would ($g_{(y)}$) mathematically impact one another such that correlation between produced *rf* scores and ART results would be maximized (they would demonstrate concurrent validity—see Research Questions).

First, to prevent high performance in rate from veiling poor performance in comprehension (or vice versa), we explored multiplying the equated values for both reading rate and comprehension by exponential terms, enabling both reading rate and comprehension to influence both terms in the formula. With the exponential terms (in gray), the formula is:

$$rf = \frac{e_{(x)}{g_{(y)}}^{w} + g_{(y)}{e_{(x)}}^{z}}{2} \tag{7}$$

We explored numerous options, but deemed (7) the most parsimonious.

To achieve a maximized correlation between participants' ART scores and their distinct *rf* scores, using Microsoft Excel and the cor.test function in the statistical program R (R Core Team, 2021) we systematically investigated iterations of measure (7), conducting Spearman rank-order correlations (r_s) between ART and each iteration's output scores. For variants with the strongest correlations, we generated bootstrapped confidence intervals (100,000 samples with replacement each) of r_s. Spearman correlations were chosen because participants' ART scores (ACTFL levels) could best be described as ordinal rather than interval data.

For each iteration of (7), we first averaged together individual participants' three distinct *rf* output scores (one for each TR) to produce one *rf* score for each participant based on the three testlets. Then, a correlation between these averaged *rf* scores and ART scores was calculated for each iteration.

A matrix showing all the combinations of exponential terms we explored is in Appendix B. For the pair of terms in each cell of the matrix, we tested every possible combination from a fixed range of thresholds for both comprehension and rate. These were from 70–100% at 5% intervals for comprehension (i.e., 70%, 75%, 80%, etc.) and from 175–350 wpm at 25 wpm intervals for rate (i.e., 175, 200, 225, etc.). This method resulted in 2688 distinct variants of the formula. The fundamental questions were: which comprehension and rate threshold would be most appropriate for t_x and t_y for (4)(above) and (5)(above), and what exponents would be most appropriate for w and z for (7)(above)?

Regression analyses. In order to conduct regression analyses to get answers we needed for Research Question #2, we generated interval scores from ART response data through Rasch analysis. Despite limitations and concerns that arise with a small sample size ($N = 68$) with Rasch analysis, using interval scores as the dependent variable (DV) seemed more appropriate than using overly broad ACTFL reading levels converted into a numeric ordinal sequence. The ART presented items at three levels of difficulty. Because the ART was adaptive, not all participants responded to test items from all three levels. For instance, participants who could not successfully answer a certain percentage of the lowest-level items, never received items from the two successively more difficult levels. To avoid inaccurately large standard errors, we followed

recommendations from Linacre (2006), creating zero-valued responses for lower ability participants and randomly assigning those zero-valued responses to a predetermined, randomly-selected number of the higher-level items that they did not receive [3]. Following this procedure, using Bond&FoxSteps (Linacre, 2006) we conducted Rasch analysis on participants' ART responses. We then used participants' resulting logit scores as the DV for regression analyses.

We first conducted simultaneous multiple regression analysis in R Version 4.1.1, regressing participants' ART logit scores on their averaged comprehension scores and averaged rate (derived from the three TRs). This model was compared to a model in which ART logit scores were regressed onto averaged *rf* scores from an optimal variant.

Both models were checked for violations of assumptions. The linearity assumption was confirmed by plotting residuals against fitted values. Qqplots demonstrated that distributions of residuals were reasonably normal. To check for independence of errors, we generated boxplots of residuals, clustered by the distinct classes in which participants had been enrolled. The results confirmed reasonable independence. Residual plots for both models showed possible violations of homoscedasticity, but Studentized Breusch-Pagan tests (Breusch & Pagan, 1979), using the lmtest package in R (Zeileis & Hothorn, 2002), confirmed heteroscedasticity was not significant in both cases. Lastly, we inspected the correlation (.425) between the two independent variables (IVs) in the first model and confirmed that multicollinearity was not a concern.

Results

Research Question 1a and 1b: Optimizing the Formula

For RQ #1a we explored various threshold values for t_x (comprehension) and t_y (reading rate), aiming to maximize correlation between readers' averaged L2 silent *rf* scores and their ART test scores (while considering alignment with theoretical principles). We arrived at several possible combinations of comprehension and rate thresholds. However, these are not fully meaningful unless considered with the best-fitting values for the exponent terms introduced in formula (7). The values selected for these terms address our RQ #1b: How should equated comprehension ($e_{(x)}$) and reading rate ($g_{(y)}$) mathematically impact one another to maximize the correlation coefficient? Because of their codependence, we will consider answers to both parts of RQ #1 simultaneously. See Table 5.

[3] For example, for all those participants who did receive level-two items, we averaged the number of level-two items they received. After arriving at a whole-number average (in this case, 13 items), all those participants who did not complete level-two items were assigned zero-value responses to an individually unique set of 13 randomly selected level-two items. The same process was followed for level-three items.

Table 5
Values for Terms of the Three Optimal Formulas

Formula	Thresholds		Exponent values		Spearman Rho	Bootstrapped 95% CI
	comp (*x*)	rate (*y*)	$e_{(x)}{}^{z}$, *z*=	$g_{(y)}{}^{w}$, *w*=		
a	70%	225	3	.5	.688	[.525, 807]
b	70%	250	4	.5	.688	[.522, .810]
c	70%	275	4	.5	.689	[.527, .809]
d	70%	300	4	.5	.689	[.527, .808]
e	70%	325	4	.5	.687	[.523, .807]
f	70%	350	4	.5	.687	[.524, .806]
g	70%	225	5	.75	.687	[.521, .808]
h	70%	250	5	.75	.689	[.523, .810]
i	70%	250	7	.75	.687	[.522, .808]
j (weakest)	100%	175	8	.25	.633	[.453, .769]
simple	--	--	--	--	.654	[.481, .785]
rate only	--	--	--	--	.471	[.247, .651]

Of all 2688 distinct possible variants, Table 5 presents data from our analysis for those combinations that maximized the correlation. For readers interested in how all 2688 correlations compared by possible thresholds and possible exponential-term values see Appendix C. For the sake of comparison to the combinations that produced the highest correlations with ART, Table 5 also gives data for a combination that produced one of the weakest correlations (*j*), data for the simple approach (formula [1]), and data for reading rate alone.

From Table 5 above, formula *d* in its complete form is:

$$e_{(x)} = \begin{cases} x/.7; x < .7 \\ 1 \ \ ; x \geq .7 \end{cases} \tag{8}$$

$$g_{(y)} = \begin{cases} y/300; y < 300 \\ 1 \quad ; y \geq 300 \end{cases} \tag{9}$$

$$rf = \frac{e_{(x)}{g_{(y)}}^{.5} + g_{(y)}{e_{(x)}}^{4}}{2} \tag{10}$$

We elected to continue our analysis with this iteration (variant *d*) because (a) the exponent value of 4 for *z* penalizes the formula less than a higher value would for inferior comprehension, balancing the influence of both variables; and (b) The sensitivity of the formula is increased by the higher threshold of 300 wpm (this will be further explained in the Discussion section later). However, notably, no correlation fell below .633, indicating that all variants of the formula had somewhat similar effects on the rank ordering of participants' three-TR average.

Research Question 2: Explained Variance and Regression Analyses

The purpose of conducting regression analyses was to investigate how explained variance would compare in two distinct models, one including both Comprehension and Reading Rate as IVs and another including only Averaged *rf* Scores generated from an optimal formula as the single IV.

Our primary interest was whether or not *rf* Scores alone could function as a stand-in for both Comprehension and Rate in the model.

For the first model, ART Logit Scores (from the Rasch analysis) were regressed onto the two separate IVs of Averaged Comprehension and Averaged Reading Rate. The overall multiple regression was statistically significant ($R^2 = .514$, $F[2, 65] = 34.38$, $p < .001$); the two variables accounted for 51% of the variance in Logit Scores, and each of the two variables had a statistically significant effect. The standardized regression coefficient (β) for Comprehension was .577 ($t[65] = 6.038$, $p < .001$). For Reading Rate, the β was .247 ($t[65] = 2.582$, $p = .012$).

The regression for the second model was also statistically significant ($R^2 = .488$, $F[1, 66] = 62.88$, $p < .001$). Averaged *rf* Scores as a single IV accounted for 49% of the variance in ART Logit Scores ($\beta = .698$, $t[66] = 7.929$, $p < .001$). How, then to select a model?

Akaike information criterion (AIC). One means of comparing the two models was through AIC model selection (Burnham & Anderson, 2004). The best-fit model was the first of the two (including both Comprehension and Reading Rate as IVs), carrying 66% of the cumulative model weight. However, a value of 1.31 for Δ_2 (i.e., less than 2) indicated "substantial support (evidence)" (p. 271) for the second model.

Research Question 3: Unidimensionality

Did *rf* measures occur in a scaled pattern (i.e., repeated co-occurrence of certain approximate comprehension values with certain approximate rates) sufficiently correlated with overall academic reading performance for *rf* output to fit a unidimensional hierarchy? See Figure 3.

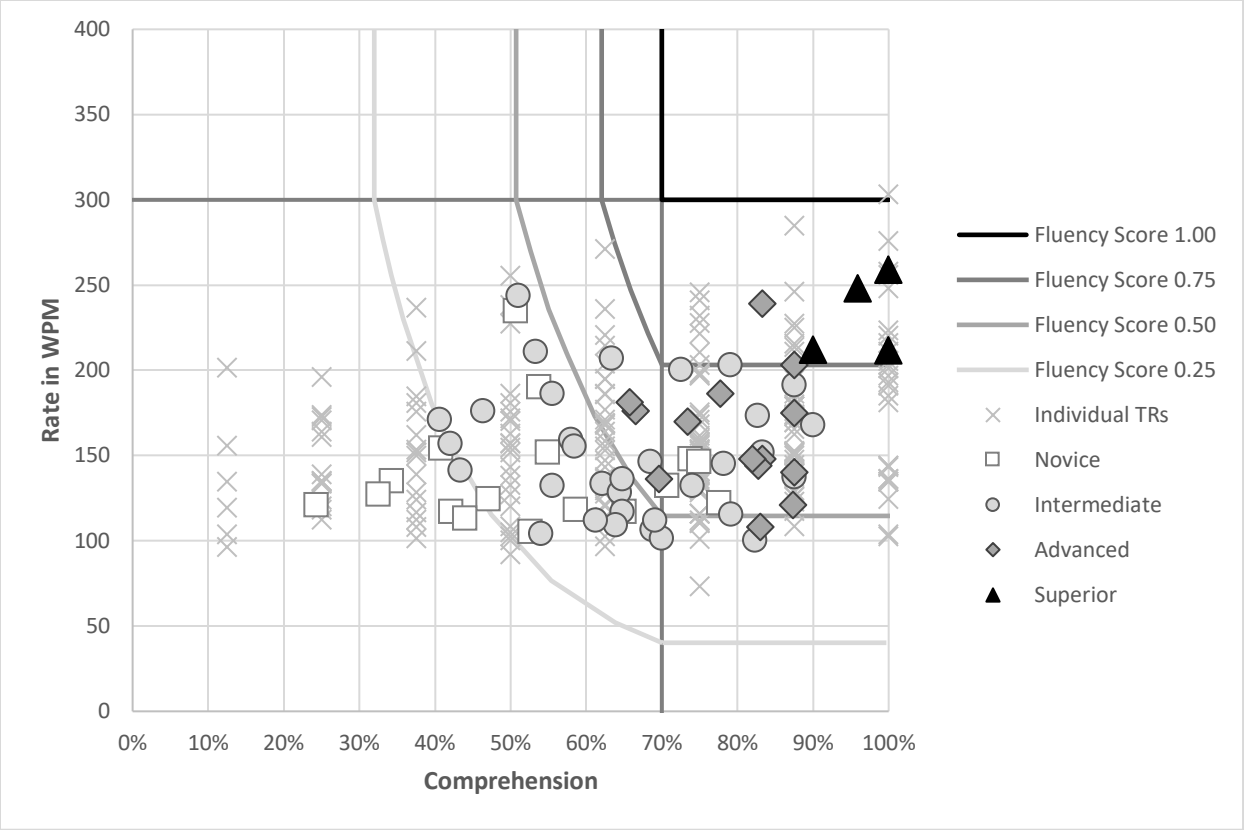


Figure 3
Three-TR Average by ACTFL Reading Level

We have already established high correlation with the ART, satisfying part of this question (see Table 5). Figures 3 and 4 present additional evidence addressing the question. Though there was overlap in performance between adjacent ACTFL levels (e.g., Novice and Intermediate), there was little overlap in performance when Novice readers were compared to Advanced readers, and when Intermediate readers were compared to Superior readers. Furthermore, after generating bootstrapped confidence intervals for mean differences (100,000 samples each), for variant *d* all mean *rf* differences between adjacent ACTFL levels were significantly different. This was not true for the weakly correlated variant *j*, the simple formula (1), or reading rate alone (see Appendix D).

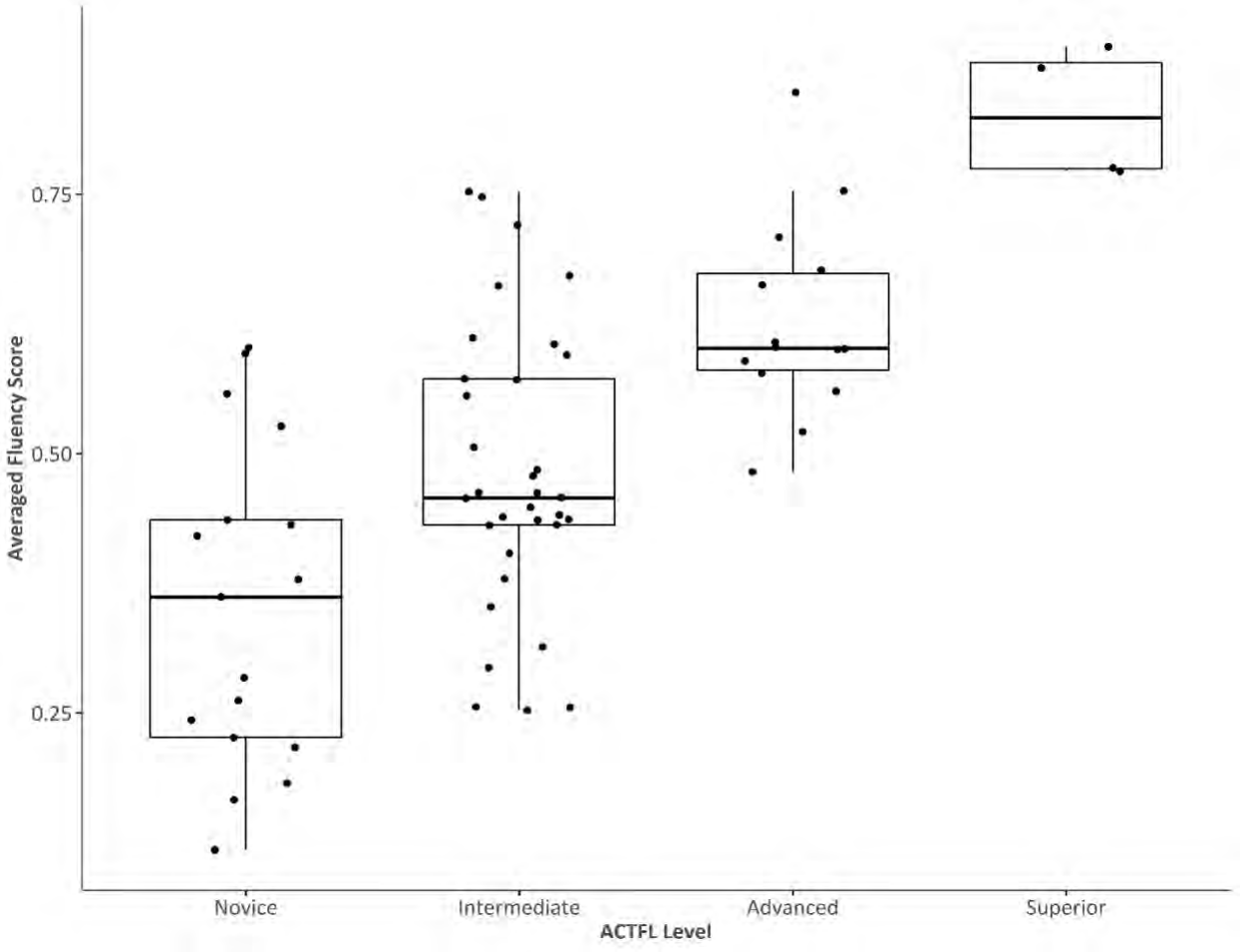


Figure 4
Boxplot of Three-TR Average rf Scores by ACTFL Level

Research Question 4: Comparing Measures

Our last RQ asked: How does *rf* compare to other possible measures of reading fluency including reading rate alone? Viewing the results of the quantitative methods, thus far *rf* shows only modest advantage compared to other measures. For instance, all confidence intervals presented in Table 5 overlap with one another indicating no significant differences in correlation coefficients. On the other hand, *rf* scores for variant *d* demonstrated strong unidimensionality when grouped by ACTFL level (Figures 3 and 4), whereas variant *j*, the simple formula (1), and rate all exhibited less unidimensionality.

The evidence that shows the greatest difference between a more strongly correlated variant of *rf* (variant *d*) and other variants and possible measures (formula [1] and reading rate alone) is presented in Figure 5. This chart shows simplified output patterns for variants *d* and *j* (scores have been constrained to seven hierarchical levels). Variant *d* clearly aligns with the principles of measuring reading fluency discussed in the Introduction whereas variant *j* does not. While variant *d* rewards adequate comprehension, it encourages increases in rate because increased comprehension beyond 70% has no effect on scores. Variant *j*, on the other hand, emphasizes comprehension, so much so that scores of 90% comprehension at 200 wpm and above only yield

a score of .5 (the maximum score being 1). Despite correlating relatively strongly with the ART, it is quite clear that theoretically and pedagogically, variant *j* is highly problematic.

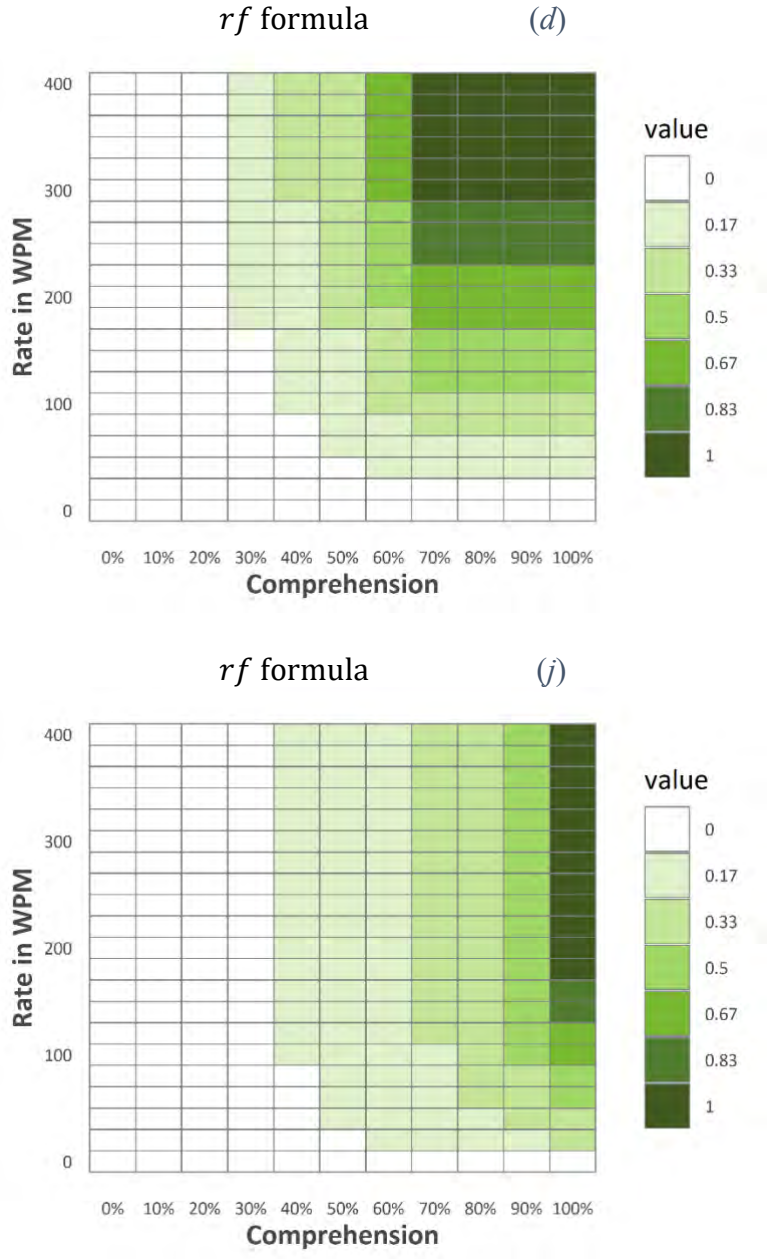


Figure 5
Score Pattern for Variant d Versus j

Discussion and Implications

Variants *a* through *i* presented in Table 5 are all similar. They give unambiguous support for a 70% threshold for adequate comprehension (also, see Figure 7 in Appendix C), confirming prevalent thinking in the literature. However, although data suggest a threshold between 225 and 350 wpm for rate, an optimal threshold remains less clear. Nonetheless, rates in excess of 200 wpm are clearly favorable. We endorse the 300-wpm threshold for the formula because this allows *rf* to be sensitive to changes in performance from 0 all the way to 300 wpm, strengthening correlation with ART scores. Including more participants with higher ability could have clarified and strengthened findings in this regard. Suffice it to say, generally, existing theory supports our chosen thresholds.

Though results demonstrate that the inclusion of both Comprehension and Reading Rate as IVs produced a slightly better-fitted model, there is evidence for the defensibility of a model with Averaged *rf* Scores as a single IV. The minimal difference in explained variance between the two models and the small difference between AIC_{min} and AIC_2 indicate that averaged *rf* score data could plausibly stand in for comprehension and reading rate data. Further studies with larger sample sizes would be insightful in verifying the tenability of this conclusion. Its high correlation with the ART test, the defensibility of the regression model, the unidimensional nature of formula output, and the theoretical support for our formula all provide some evidence in favor of *rf*. Though the formula may appear daunting, it can be calculated simply through a function in Microsoft Excel. If comprehension is found in cell A1 (as a decimal value) and rate in cell B1, the formula is

```
=(MIN(1,A1/.7)*MIN(1,B1/300)^.5+MIN(1,B1/300)*MIN(1,A1/.7)^4)/2
```

Perhaps the greatest potential value of *rf* is that it offers a practical means of quantifying developing L2 readers' fluency progress, enabling students and their teachers to make more straightforward comparisons between their various TRs. This is especially true for those learners with lower comprehension. An excellent way of utilizing the strengths of *rf* in the classroom is to encourage students to monitor their own progress with a chart similar to the top one presented in Figure 5 (see Appendix E). The clear proximal goals on the chart provide a way of discerning incremental growth especially for lower ability students who may need substantial practice to reach high levels of fluency.

The benefit of the clarity of measurement that is provided by *rf* could potentially serve other populations of readers as well. Other threshold values could easily be plugged into the formula resulting in variants that could be explored with different populations. For instance, lower reading rate thresholds may be appropriate for younger age groups. This research could also benefit reading fluency research in general as it provides an alternative means of controlling for comprehension when investigating rate. However, we fully acknowledge that this report represents an initial step, inviting further study and exploration.

When compared to the much simpler formula (1), *rf* was in some ways quite similar when considering quantitative analyses. However, *rf* is likely preferable when considering Figures 2 and 5, and the unwanted problem of rate compensating for poor comprehension. Furthermore,

differing *rf* formulas, despite having seemingly minor differences in correlation with the ART in this study, have dramatic practical consequences because of the differences in the relative value they assign to a given performance on the chart (Figure 5). This has implications for students' perceptions of the objective and their progress toward it as well as how well formula output aligns with accepted theory.

Limitations

One major limitation of this study is the reliability estimates for the three TRs, which fall below the generally accepted standard of 0.80. While reliability is an important aspect of test development and research, it is but one of several factors to be considered in relation to TLU domain, as described by Bachman and Palmer's framework for test usefulness (1996). Content and procedures for TRs take significant time and effort to develop; because the TR content and procedures were the best fit for this particular language testing case, lower reliability was tolerated. Future studies may use other TR content and procedures that align with the TLU domain and have higher reliability.

Further, this study was conducted with a small sample size with unequal representation of genders, which may not generalize to other populations. While a differential item or test functioning analysis between genders was not the purpose of this study, replication studies and further examination of this particular approach to reading fluency are needed.

Conclusion

It is incumbent upon practitioners to control for comprehension when assessing reading fluency, because as Pressley (2006) stated, "Nobody should be interested in or promoting fast reading with low comprehension" (p. 209). We add that when measuring reading fluency is a loose process, the measure becomes less credible. Unless principles of measurement are followed, variance in students' progress can be heavily influenced by error stemming from various sources. In the absence of sound practices, it is difficult to assess where learners stand or how much progress they have made.

This report introduces a formula that accounts for comprehension in a measure of fluency, exhibits unidimensional characteristics, and has some empirical and theoretical support. Our analysis of numerous possible variants of the formula provides relatively unambiguous empirical support for a 70% threshold—i.e., comprehension scores at or exceeding 70% should be the target for university English language learners' timed reading activities. It is also clear that reading rates at or exceeding 225 wpm are preferred. While the strongest variants of the formula were not statistically better in terms of correlation with the ART than the weakest variants or the simple formula (1), practical implications clearly favor the stronger variants. We encourage other researchers and practitioners to explore the formula and other possible variants using the versatile reading fluency tracking workbook provided at this url: https://tinyurl.com/rdngfluencytools.

An appropriate formula combined with sound measurement principles promises to produce more precise, defensible, and telling measurements of reading fluency.

References

ACTFL. (2012). *ACTFL proficiency guidelines 2012*. Retrieved August 9, 2023, from https://www.actfl.org/educator-resources/actfl-proficiency-guidelines

Anderson, N.J. (1999a). *Exploring second language reading: Issues and strategies.* Heinle & Heinle.

Anderson, N.J. (1999b). Improving reading speed: Activities for the classroom. *English Teaching Forum, 37*(2), 2–5.

Anderson, N.J. (2003). Teaching reading. In D. Nunan (Ed.), *Practical English language teaching* (pp. 67–86). McGraw Hill Publishers.

Anderson, N.J. (2006). *ELT Advantage: Teaching ESL/EFL Reading.* Thompson ELT. Retrieved from http://www.ed2go.com/eltadvantage/

Anderson, N.J. (2008). *Practical English language teaching: Reading.* McGraw Hill.

Anderson, N.J. (2018). Silent reading fluency. In J.I. Liontas, TESOL International Association, & M. DelliCarpini (Eds.), *The TESOL encyclopedia of English language teaching*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118784235.eelt0464

Aro, T., Viholainen, H., Koponen, T., Peura, P., Räikkönen, E., Salmi, P., Sorvo, R., & Aro, M. (2018). Can reading fluency and self-efficacy of reading fluency be enhanced with an intervention targeting the sources of self-efficacy? *Learning and Individual Differences, 67*, 53–66. https://doi.org/10.1016/j.lindif.2018.06.009

Bachman, L.F., & Palmer. A.S. (1996). *Language testing in practice*. Oxford University Press.

Beglar, D., Hunt, A., & Kite, Y. (2012). The effect of pleasure reading on Japanese university EFL learners' reading rates. *Language Learning, 62*(3), 665–703. https://doi.org/10.1111/j.1467-9922.2011.00651.x

Breusch, T.S., & Pagan, A.R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica, 47*(5), 1287–1294. https://doi.org/10.2307/1911963

Breznitz, Z. (2006). *Fluency in reading: Synchronization of processes.* Erlbaum.

Bryant, B.R., Shih, M., & Bryant, D.P. (2009). The Gray Oral Reading Test—fourth edition (GORT-4). In J.A. Naglieri, & S. Goldstein (Eds.), *Practitioner's guide to assessing intelligence and achievement* (pp. 417–447). John Wiley & Sons, Inc.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods Research, 33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Carver, R.P. (1990). *Reading rate: A review of research and theory.* Academic Press, Inc.

Carver, R.P. (2000). *The causes of high and low reading achievement*. Erlbaum.

Chang, C-S. (2014). Measuring reading comprehension in an L2 speed reading course: Response to McLean. *Reading in a Foreign Language, 26*(1), 192–194.

Chang, C-S., & Millet, S. (2013). Improving reading rates and comprehension through timed repeated reading. *Reading in a Foreign Language, 25*(2), 126–148.

Chung, M., & Nation, I.S.P. (2006). The effect of a speed reading course. *English Teaching, 61*(4), 181–204.

Clifford, R., & Cox, T.L. (2013). Empirical validation of reading proficiency guidelines. *Foreign Language Annals, 46*(1), 45–61. https://doi.org/10.1111/flan.12033

Data USA. (n.d.). *Brigham Young University-Hawaii.* Retrieved June 8, 2022, from https://datausa.io/profile/university/brigham-young-university-hawaii#about

Davies, M. (2015). The corpus of contemporary American English (COCA): 520 million words, 1990–present. Available from https://www.english-corpora.org/coca/

De Hoyos, D., & David, N. (2018). Understanding reading purpose. In J.I. Liontas, TESOL International Association, & M. DelliCarpini (Eds.), *The TESOL encyclopedia of English language teaching*. John Wiley & Sons, Inc. https://doi-org.byuh.idm.oclc.org/10.1002/9781118784235.eelt0501

Erten, I.E. (2018). Activation of prior knowledge. In J.I. Liontas, TESOL International Association, & M. DelliCarpini (Eds.), *The TESOL encyclopedia of English language teaching*. John Wiley & Sons, Inc. https://doi-org.byuh.idm.oclc.org/10.1002/9781118784235.eelt0801

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

Gorsuch, G., & Taguchi, E., (2008). Repeated reading for developing reading fluency and reading comprehension: The case of EFL learners in Vietnam. *System, 36*, 253–278. https://doi.org/10.1016/j.system.2007.09.009

Gorsuch, G., & Taguchi, E., (2010). Developing reading fluency and comprehension using repeated reading: Evidence from longitudinal student reports. *Language Teaching Research, 14*(1), 27–59. https://doi.org/10.1177/1362168809346494

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice.* Cambridge University Press.

Grabe, W. (2010). Fluency in reading—Thirty-five years later. *Reading in a Foreign Language, 22*(1), 71–83.

Grabe, W., & Stoller, F.L. (2020). *Teaching and researching reading* (3rd ed.). Routledge.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193–202. https://doi.org/10.3758/BF03195564

Hudson, R.F., Pullen, P.C., Lane, H.B., & Torgesen, J.K. (2008). The complex nature of reading fluency: A multidimensional view. *Reading & Writing Quarterly, 25*(1), 4–32. https://doi.org/10.1080/10573560802491208

Huffman, J. (2014). Reading rate gains during a one-semester extensive reading course. *Reading in a Foreign Language, 26*(2), 17–33.

Jeffries, L., & Mikulecky, B.S. (2014a). *Advanced reading power 4* (2nd ed.). Pearson Education, Inc.

Jeffries, L., & Mikulecky, B.S. (2014b). *Advanced reading power 4 test booklet* (2nd ed.). Pearson Education, Inc.

Klauda, S.L., & Guthrie, J.T. (2008). Relationships of three components of reading fluency to reading comprehension. *Journal of Educational Psychology, 100*(2), 310–321. https://doi.org/10.1037/0022-0663.100.2.310

Kramer, B., & McLean, S. (2019). L2 reading rate and word length: The necessity of character-based measurement. *Reading in a Foreign Language, 31*(2), 201–225.

Kuhn, M.R., Schwanenflugel, P.J., & Meisinger, E.B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly, 45*(2), 230–251. https://doi.org/10.1598/RRQ.45.2.4

LaBerge, D., & Samuels, J.S. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*(2), 293–323. https://doi.org/10.1016/0010-0285(74)90015-2

Linacre, J.M. (2006). *Bond&FoxSteps Rasch measurement computer program* [Computer software]. Winsteps.com.

Lynn, E.M. (2021). Unassisted repeated reading: Exploring the effects of intensity, treatment duration, background knowledge, individual variation, and text variation on reading rate. *Reading in a Foreign Language, 33*(1), 30–54.

Macalister, J. (2008). The effect of a speed reading course in an English as a second language environment. *TESOLANZ Journal, 16*, 23–33.

Macalister, J. (2010). Speed reading courses and their effect on reading authentic texts: A preliminary investigation. *Reading in a Foreign Language, 22*(1), 104–116.

McLean, S. (2014). Addressing the importance of comprehension to reading: Learning lessons from Chang (2012). *Reading in a Foreign Language, 26*(1), 186–191.

McLean, S., & Rouault, G. (2017). The effectiveness and efficiency of extensive reading at developing reading rates. *System, 70*, 92–106. https://doi.org/10.1016/j.system.2017.09.003

Nation, I.S.P. (2005). Reading faster. *PASAA, 36*, 21–35.

Nation, I.S.P. (2009). *Teaching ESL/EFL reading and writing*. Routledge.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (National Institute of Health Pub. No. 00-4769). Washington, DC: National Institute of Child Health and Human Development.

Nuttall, C. (2007). *Teaching reading skills in a foreign language* (2nd ed.). Heinemann.

Paige, D.D. (2011). Engaging struggling adolescent readers through situational interest: A model proposing the relationships among extrinsic motivation, oral reading fluency, comprehension, and academic achievement. *Reading Psychology, 32*(5), 395–425. https://doi.org/10.1080/02702711.2010.495633

Paige, D.D., Rasinski, T., Magpuri-Lavell, T., & Smith, G.S. (2014). Interpreting the relationships among prosody, automaticity, accuracy, and silent reading comprehension in secondary students. *Journal of Literacy Research, 46*(2), 123–156. https://doi.org/10.1177/1086296X14535170

Pikulski, J.J., & Chard, D.J. (2005). Fluency: Bridge between decoding and reading comprehension. *The Reading Teacher*, *58*(6), 510–519. https://doi.org/10.1598/RT.58.6.2

Pressley, M. (2006). *Reading instruction that works* (3rd ed.). Guilford Press.

Quinn, E., Nation, I.S.P., & Millet, S. (2007). *Asian and Pacific speed readings for ESL learners.* English Language Institute Occasional Publication.

R Core Team. (2021). *R: A language and environment for statistical computing* (Version 4.1.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Samuels, S.J. (2007). The DIBELS tests: Is speed of barking at print what we mean by reading fluency? *Reading Research Quarterly, 42*(4), 563–566. https://doi-org.byuh.idm.oclc.org/10.1598/RRQ.42.4.5

Schrank, F.A., McGrew, K.S., & Mather, N. (2014). *Woodcock-Johnson IV Tests of Cognitive Abilities.* Riverside.

Stanovich, K. (2000). *Progress in understanding reading: Scientific foundations and new frontiers.* Guilford Press.

Taguchi, E. (1997). The effects of repeated readings on the development of lower identification skills of FL readers. *Reading in a Foreign Language, 11*(1), 97–119.

Vclock.com. (2022). *Vclock Stopwatch.* https://vclock.com/stopwatch/

Wilcox, M. (2020). *Brigham Young University Adaptive Listening (ALT) and Adaptive Reading Test (ART) Technical Report* (p. 120) [Technical Report]. Center for Language Studies, Brigham Young University.

Zeileis, A., & Hothorn, T. (2002). *"Diagnostic Checking in Regression Relationships." R News, 2(3), 7–10* (Version 0.9-39) [Computer software]. https://CRAN.R-project.org/package=lmtest

Zwick, M.J. (2018). Measuring reading fluency. In J. I. Liontas, TESOL International Association, & M. DelliCarpini (Eds.), *The TESOL encyclopedia of English language teaching*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118784235.eelt0495

Appendix A

Turning our attention to the construction of a measure, proposed limits function as thresholds of achievement beyond which measurement is no longer sensitive to any increase. If 200 wpm is selected as a threshold for rate, any increase beyond 200 wpm would no longer add to the measure. If 70% is selected as a threshold for comprehension, then any increase in comprehension beyond this point would not add to a fluency score.

First, rate and comprehension must be equated to a common scale. Comprehension is often measured on a scale of 0 to 100% whereas rate is measured in wpm, ranging anywhere from 0 to in excess of 300. Considering comprehension first, the range from 0 to 70% can be equated to a 100-point proportion scale (0 to 100%). For the equation to operate effectively (i.e., 70% yielding a maximum score), thresholds need to be incorporated. Where x = *raw comprehension* (a decimal proportion score), t_x = *threshold for comprehension*, and $e_{(x)}$ = *equated comprehension*, then

$$e_{(x)} = \frac{x}{t_x}$$

The algebra is straightforward. For a threshold of success defined as 70%

$$\frac{x}{.7} = e_{(x)} \quad \text{and} \quad \frac{.7}{.7} = 1$$

By including this constant, the following comprehension scores are modified as follows:
65% = .929
40% = .571
75% = 1.071[4]
For rate we can adopt a similar approach; where y = *rate (in wpm)*, t_y = *threshold for rate*, $g_{(y)}$ = *equated rate*, then

$$g_{(y)} = \frac{y}{t_y}$$

[4] In actuality, for the equation to function such that comprehension cannot compensate for rate and vice versa, anything in excess of 1 should simply be truncated at 1

If 200 wpm is set as the threshold, this equates 200 wpm to 1 (a maximum score):

$$\frac{y}{200} = g_{(y)} \quad \text{and} \quad \frac{200}{200} = 1$$

Using this formula, the following reading rate scores (in wpm) are modified as follows:
200 = 1
150 = .75
80 = .4
The respective contributions of each equated variable ($e_{(x)}$ and $g_{(y)}$) to the fluency calculation must be specified. Being that rate has no value without comprehension, logically, comprehension should have a sizeable impact on the measure. As a starting point for investigation, we specified that $e_{(x)}$ or comprehension accounted for 50% of the measure, and $g_{(y)}$ or rate accounted for the remaining 50%. The two equated variables can be combined to create a measure of fluency (*rf*) that is sensitive to both rate and comprehension as follows:

$$rf = \left(e_{(x)} + g_{(y)}\right)/2 \quad \text{or} \quad \left(\frac{x}{.7} + \frac{y}{200}\right)/2 \tag{i}$$

However, the problem remains that without some modification, one part of formula (i) could veil poor performance in the other. An additional element that truncates both parts of (i) at their designated limits partially solves the problem. With this element added, the two parts of (i) become:

$$e_{(x)} = \frac{x}{t_x} - \left(\frac{\sqrt{\left(1 - \frac{x}{t_x}\right)^2} - \left(1 - \frac{x}{t_x}\right)}{2}\right) \tag{ii}$$

$$g_{(y)} = \frac{y}{t_y} - \left(\frac{\sqrt{\left(1 - \frac{y}{t_y}\right)^2} - \left(1 - \frac{y}{t_y}\right)}{2}\right) \tag{iii}$$

Effectively, these added elements limit $e_{(x)}$ and $g_{(y)}$ respectively to a maximum value of 1. For the plausible thresholds of interest, the two parts would be

$$e_{(x)} = \frac{x}{.7} - \left(\frac{\sqrt{\left(1 - \frac{x}{.7}\right)^2} - \left(1 - \frac{x}{.7}\right)}{2}\right)$$

$$g_{(y)} = \frac{y}{200} - \left(\frac{\sqrt{\left(1 - \frac{y}{200}\right)^2} - \left(1 - \frac{y}{200}\right)}{2}\right)$$

Rather than representing these two parts algebraically, it is perhaps simpler and more concise to represent them as piecewise functions. Therefore, the full formula would be

$$rf = \frac{e_{(x)} + g_{(y)}}{2} \tag{i}$$

$$e_{(x)} = \begin{cases} x/t_x; x < t_x \\ 1 \ ; x \geq t_x \end{cases} \tag{iv}$$

$$g_{(y)} = \begin{cases} y/t_y; y < t_y \\ 1 \ ; y \geq t_y \end{cases} \tag{v}$$

Inputting the example threshold values, the two parts would be

$$e_{(x)} = \begin{cases} x/.7; x < .7 \\ 1 \ ; x \geq .7 \end{cases}$$

$$g_{(y)} = \begin{cases} y/200; y < 200 \\ 1 \quad ; y \geq 200 \end{cases}$$

The complexity of these calculations may be off-putting to the average practitioner, but the entirety of the formula can be executed rather simply in Microsoft Excel or Google Sheets.

Appendix B

$e_{(x)}{}^{9}$ $g_{(y)}{}^{.25}$	$e_{(x)}{}^{8}$ $g_{(y)}{}^{.25}$	$e_{(x)}{}^{7}$ $g_{(y)}{}^{.25}$	$e_{(x)}{}^{6}$ $g_{(y)}{}^{.25}$	$e_{(x)}{}^{5}$ $g_{(y)}{}^{.25}$	$e_{(x)}{}^{4}$ $g_{(y)}{}^{.25}$	$e_{(x)}{}^{3}$ $g_{(y)}{}^{.25}$	$e_{(x)}{}^{2}$ $g_{(y)}{}^{.25}$
$e_{(x)}{}^{9}$ $g_{(y)}{}^{.5}$	$e_{(x)}{}^{8}$ $g_{(y)}{}^{.5}$	$e_{(x)}{}^{7}$ $g_{(y)}{}^{.5}$	$e_{(x)}{}^{6}$ $g_{(y)}{}^{.5}$	$e_{(x)}{}^{5}$ $g_{(y)}{}^{.5}$	$e_{(x)}{}^{4}$ $g_{(y)}{}^{.5}$	$e_{(x)}{}^{3}$ $g_{(y)}{}^{.5}$	$e_{(x)}{}^{2}$ $g_{(y)}{}^{.5}$
$e_{(x)}{}^{9}$ $g_{(y)}{}^{.75}$	$e_{(x)}{}^{8}$ $g_{(y)}{}^{.75}$	$e_{(x)}{}^{7}$ $g_{(y)}{}^{.75}$	$e_{(x)}{}^{6}$ $g_{(y)}{}^{.75}$	$e_{(x)}{}^{5}$ $g_{(y)}{}^{.75}$	$e_{(x)}{}^{4}$ $g_{(y)}{}^{.75}$	$e_{(x)}{}^{3}$ $g_{(y)}{}^{.75}$	$e_{(x)}{}^{2}$ $g_{(y)}{}^{.75}$
$e_{(x)}{}^{9}$ $g_{(y)}$	$e_{(x)}{}^{8}$ $g_{(y)}$	$e_{(x)}{}^{7}$ $g_{(y)}$	$e_{(x)}{}^{6}$ $g_{(y)}$	$e_{(x)}{}^{5}$ $g_{(y)}$	$e_{(x)}{}^{4}$ $g_{(y)}$	$e_{(x)}{}^{3}$ $g_{(y)}$	$e_{(x)}{}^{2}$ $g_{(y)}$
$e_{(x)}{}^{9}$ $g_{(y)}{}^{1.25}$	$e_{(x)}{}^{8}$ $g_{(y)}{}^{1.25}$	$e_{(x)}{}^{7}$ $g_{(y)}{}^{1.25}$	$e_{(x)}{}^{6}$ $g_{(y)}{}^{1.25}$	$e_{(x)}{}^{5}$ $g_{(y)}{}^{1.25}$	$e_{(x)}{}^{4}$ $g_{(y)}{}^{1.25}$	$e_{(x)}{}^{3}$ $g_{(y)}{}^{1.25}$	$e_{(x)}{}^{2}$ $g_{(y)}{}^{1.25}$
$e_{(x)}{}^{9}$ $g_{(y)}{}^{1.5}$	$e_{(x)}{}^{8}$ $g_{(y)}{}^{1.5}$	$e_{(x)}{}^{7}$ $g_{(y)}{}^{1.5}$	$e_{(x)}{}^{6}$ $g_{(y)}{}^{1.5}$	$e_{(x)}{}^{5}$ $g_{(y)}{}^{1.5}$	$e_{(x)}{}^{4}$ $g_{(y)}{}^{1.5}$	$e_{(x)}{}^{3}$ $g_{(y)}{}^{1.5}$	$e_{(x)}{}^{2}$ $g_{(y)}{}^{1.5}$

Figure 6
All Pairs of Exponential Terms Tested

Appendix C

Figures 7 and 8 each show the average spearman correlation between formula output scores and ART scores for all formula variants according to their incorporated combination of thresholds or exponential terms. These graphs demonstrate clearly which combinations of thresholds and exponential terms resulted in higher correlations on average. For instance, referring to Figure 7, the average of all correlations for all 48 formula variants that incorporated thresholds of 175 wpm for rate and 70% for comprehension was .662 (keep in mind that each variant had a different combination of exponential terms). It is fairly clear that the combination of any threshold for rate exceeding 200 wpm (i.e., 225 wpm and above) and a comprehension threshold of 70% resulted in the higher correlations. Similarly, certain combinations of exponent values for respective terms clearly resulted in higher average correlations of output scores with ART scores (see the grey and dark orange lines in Figure 8).

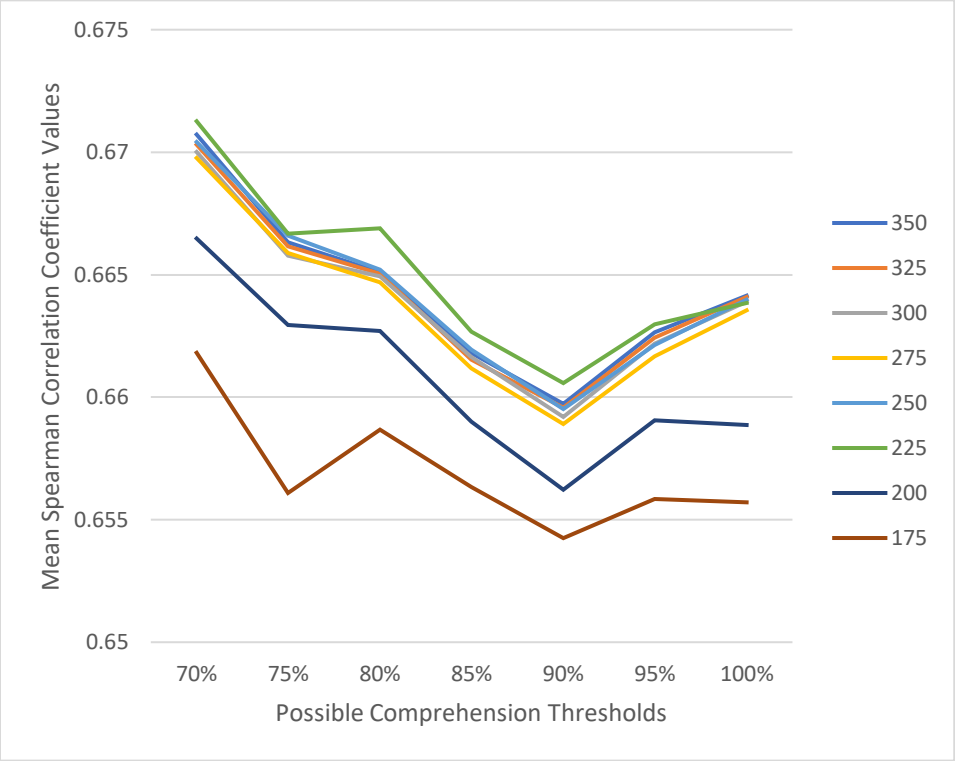


Figure 7
Mean Correlation Coefficient Values by All Formula Variants According to Incorporated Comprehension and Rate Thresholds

Note. The colored lines represent different possible rate thresholds in words per minute. Refer to formulas (4) and (5) in the article to see how these thresholds influence the formula.

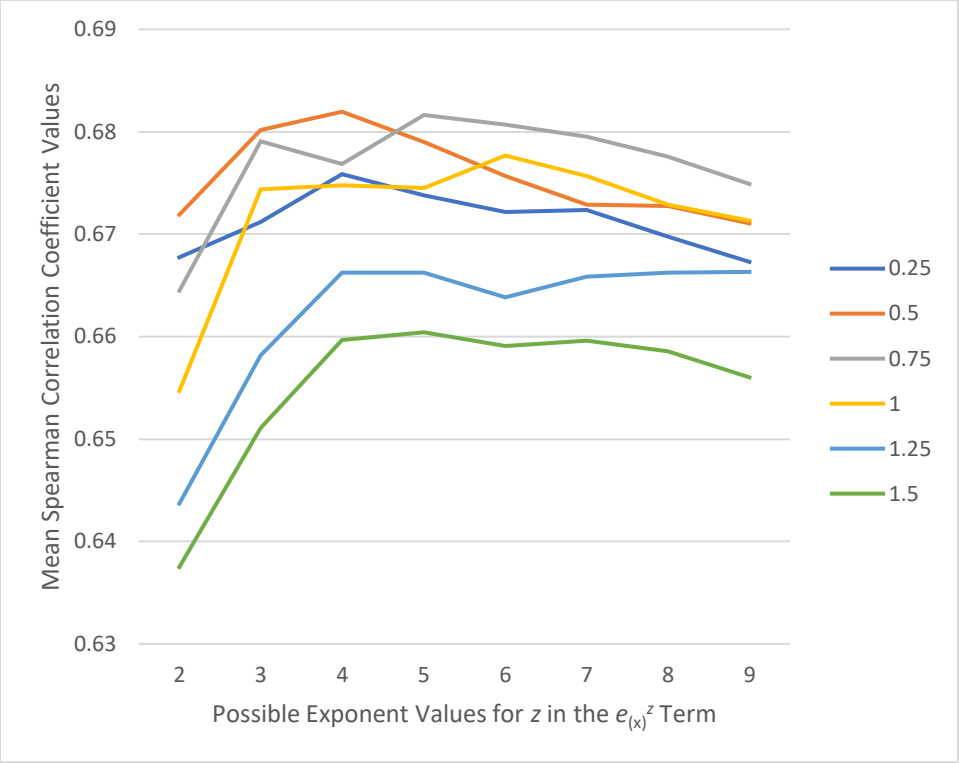


Figure 8
Mean Correlation Coefficient Values by All Formula Variants According to Incorporated Exponential Terms
Note. The colored lines represent different possible exponent values for w in the $g_{(y)}{}^{w}$ term. Refer to formula (7) in the article to see how these exponential terms influence the formula.

Appendix D

To investigate the effect of ACTFL level on reading fluency for each distinct measure (i.e., whether mean differences between adjacent ACTFL levels were significant for each distinct measure), we followed a bootstrapping methodology. Each group of participants' average reading fluency scores (the average of their three TR scores for each formula or the average of their three rates for the 'rate only' comparisons) for each specific ACTFL level were repeatedly sampled with replacement. Separate sets of 100,000 bootstrap samples were generated to create confidence intervals. Mean differences were then calculated for all samples using the two distinct sets of 100,000 in each instance. This resulted in 100,000 possible mean differences in fluency scores between ACTFL levels for each comparison from which 95% confidence intervals were created using the percentile method. Once confidence intervals were generated, *p*-values were calculated from the confidence intervals following a method described by Rousselet et al. (2021). The resulting *p*-values were compared to a Bonferroni-adjusted alpha level of .004 (.05/12 = .004) because 12 statistical tests were conducted. Variant *d* was the only variant of the formula for which all comparisons between scores by ACTFL level proved to be significantly different.

Table 6
Results of Bootstrapped Pairwise Comparisons

Formula	Compared ACTFL Levels		Difference in Mean Scores				
			M1	*M2*	*M1-M2* =	Bootstrapped *p*-value =	Bootstrapped *95% CI*
d	Intermediate	Novice	0.48	0.35	0.13	.003*	[0.04, 0.22]
	Advanced	Intermediate	0.63	0.48	0.15	.000*	[0.08, 0.21]
	Superior	Advanced	0.83	0.63	0.20	.000*	[0.13, 0.27]
j	Intermediate	Novice	0.39	0.29	0.10	.020	[0.02, 0.18]
	Advanced	Intermediate	0.56	0.39	0.17	.000*	[0.09, 0.24]
	Superior	Advanced	0.90	0.56	0.34	.000*	[0.21, 0.46]
simple	Intermediate	Novice	0.35	0.26	0.09	.008	[0.02, 0.16]
	Advanced	Intermediate	0.47	0.35	0.13	.000*	[0.07, 0.19]
	Superior	Advanced	0.76	0.47	0.28	.000*	[0.19, 0.38]
rate only	Intermediate	Novice	158.1	143.1	15.0	.122	[-4.2, 33.0]
	Advanced	Intermediate	176.7	158.1	18.6	.078	[-2.1, 39.2]
	Superior	Advanced	233.6	176.7	57.0	.000*	[30.1, 83.8]

Note. For Superior, $n = 4$; for Advanced, $n = 14$; for Intermediate, $n = 33$; and for Novice, $n = 17$. A Bonferroni-adjusted alpha level of .004 was used.
*$p < .004$.

Reference

Rousselet, G. A., Pernet, C. R., & Wilcox, R. R. (2021). The percentile bootstrap: A primer with step-by-step instructions in R. *Advances in Methods and Practices in Psychological Science, 4*(1), 1–10. https://doi.org/10.1177/2515245920911881

Appendix E

Modified Reading Fluency Chart

Rate (WPM)	0-19%	20%	30%	40%	50%	60%	70%	80%	90%	100%
300	0	1	1	2	2	3	5	5	5	5
290	0	1	1	2	2	3	5	5	5	5
280	0	1	1	2	2	3	5	5	5	5
270	0	1	1	2	2	3	4	4	4	4
260	0	1	1	1	2	3	4	4	4	4
250	0	1	1	1	2	3	4	4	4	4
240	0	1	1	1	2	3	4	4	4	4
230	0	1	1	1	2	3	4	4	4	4
220	0	1	1	1	2	3	4	4	4	4
210	0	1	1	1	2	3	4	4	4	4
200	0	1	1	1	2	3	4	4	4	4
190	0	1	1	1	2	2	3	3	3	3
180	0	1	1	1	2	2	3	3	3	3
170	0	1	1	1	2	2	3	3	3	3
160	0	1	1	1	2	2	3	3	3	3
150	0	0	1	1	2	2	3	3	3	3
140	0	0	1	1	1	2	3	3	3	3
130	0	0	1	1	1	2	3	3	3	3
120	0	0	1	1	1	2	2	2	2	2
110	0	0	1	1	1	2	2	2	2	2
100	0	0	1	1	1	2	2	2	2	2
90	0	0	1	1	1	1	2	2	2	2
80	0	0	1	1	1	1	2	2	2	2
70	0	0	1	1	1	1	2	2	2	2
60	0	0	0	1	1	1	2	2	2	2
50	0	0	0	1	1	1	1	1	1	1
40	0	0	0	1	1	1	1	1	1	1
30	0	0	0	0	1	1	1	1	1	1
20	0	0	0	0	0	1	1	1	1	1
10	0	0	0	0	0	0	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0

Comprehension

About the Authors

Steven J. Carter (corresponding author), MA TESOL, is an Assistant Professor in the English Language Teaching and Learning faculty at Brigham Young University–Hawaii. He has taught English language courses in both intensive English and community English programs. He has worked on three major curriculum development projects and has been involved in the creation, analysis, and revision of numerous assessments. His research interests include second language reading and assessment.
https://orcid.org/0000-0003-1031-9624; steven.carter@byuh.edu

Matthew P. Wilcox, Ph.D., is the Associate Director for Measurement and Evaluation at the Center for Language Studies at Brigham Young University. He is responsible for the development and maintenance of proficiency-based and achievement language tests at BYU for both major and less-commonly taught languages. His interests include measurement, psychometrics, and data science. https://orcid.org/0000-0002-6020-529X; wilcoxmp@byu.edu

Neil J. Anderson, Ph.D., is a former Professor in the English Language Teaching and Learning at Brigham Young University–Hawaii and the former Director of the Center for Learning and Teaching. Professor Anderson is the author or co-editor of over 50 books, book chapters, and journal articles. His research interests include second language reading, language learner strategies, learner self-assessment, motivation in language teaching and learning, and ELT leadership development. neil.anderson@byuh.edu