Vocabulary Learning and Instruction
Volume 8, Issue 1, March 2019
http://vli-journal.org

Validating the Construct of Readability in EFL Contexts: A Proposal for Criteria

Geoffrey G. Pinchbeck
Carleton University

Abstract

This article examines how English as a foreign language learners might be better matched to reading texts using automatic readability analysis. Specifically, I examine how the lexical decoding component of readability might be validated. In Japan, readability has been mostly determined by publishers or by professional reading organizations who only occasionally publish their lists of readability ratings for specific texts. Without transparent readability methods, candidate texts cannot be independently evaluated by practitioners. Moreover, the reliance on centralized organizations to curate from commercially available texts precludes the evaluation of the multitudes of free texts that are increasingly available on the Internet. Previous studies that have attempted to develop automatic readability formulas for Japanese learners have used surface textual features of texts, such as word and/or sentence length, and/or they have used word-frequency lists derived from large multi-register corpora. In this article, I draw upon on the findings of a study that examines how such word-lists might be validated for use in matching Japanese learners to texts (Pinchbeck, manuscript in preparation). Finally, I propose a list of general criteria that might be used to evaluate the components of readability formulas in general.

Keywords: Readability, vocabulary, matching texts, Japan, cloze test

1 Background

A key dilemma in English as a foreign language (EFL) program design is how to choose course texts that are meaningful and interesting, and that also provide a source of useful new language input. This dilemma lies in the fact that texts – written or spoken – that are *over*-loaded with new (i.e., unknown) language forms will not be meaningful or interesting. Matching texts based on their content, interest, and language level to specific groups of learners and/or individuals is, therefore, a prerequisite for effective language pedagogy. Extensive reading has been proposed as a key learning strategy by which learners can make progress toward advanced proficiency in general and/or academic English (e.g., Nation, 2014). However, the dilemma inherent in matching learners to texts becomes more apparent when research that has examined the lexical difficulty thresholds is considered. These studies have converged toward the recommendation that at least 95% of the words in instructional texts be already known to the

Please cite this article as: Pinchbeck, G. G. (2019). Validating the Construct of Readability in EFL Contexts: A Proposal for Criteria. *Vocabulary Learning and Instruction, 8* (1), 8–16. https://doi.org/10.7820/vli.v08.1.pinchbeck

learner, and this threshold increases to 98% for independent reading (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Schmitt, Jiang, & Grabe, 2011). Therefore, practitioners and learners are inevitably limited by the availability of level-appropriate reading texts that would allow learners to continually encounter new language without being overwhelmed by unknown words in the texts with which they attempt to engage (see Cobb, 2007, 2008).

1.1 Readability formulas

Automatic readability formulas have been used for decades by publishers and practitioners in K-12 education. Theoretically, they allow candidate texts to be evaluated for the difficulty they will likely present to learners at different levels. Although commonly used formulas such as the Flesch–Kincaid grade level (Flesch, 1948; Kincaid et al., 1975) – which are packaged with MS-Word – have been shown to correlate with cloze test scores, the effect size of this relationship in EFL contexts varies dramatically depending on the study (r = 0.48 in Brown, 1998, and r = 0.85 in Greenfield, 1999). Even in K-12 contexts where they were originally conceived, they have been criticized for their lack of construct validity (see below), as they are based on statistical curve fitting, rather than on a theory of reading and/or language development. This problem becomes more obvious when the accepted difficulty of certain types of texts is not accurately predicted by this formula; for example, Shakespeare's *Hamlet* is rated by Flesch–Kincaid analysis as a text appropriate for children in Grade 2. Because these formulas measure only the surface features of texts, rather than the underlying factors that determine text difficulty, they also fail in providing writers any accurate guidance for rewriting texts to be more accessible to readers at lower levels of language proficiency. These types of criticisms of readability formulas have been common in the K-12 reading research literature (Bruce, Rubin, & Starr, 1981; Connatser, 1999; Davison & Kantor, 1982; Duffy, 1985; Lange, 1982; Maxwell, 1978; Selzer, 1981), and in the L2 literature, where several authors have called for the development of more accurate readability formulas that have been optimized for specific L2 contexts (e.g., Carrell, 1987; Greenfield, 1999).

1.2 Readability in Japan

There are a range of approaches to determine text-level appropriateness to EFL practitioners in Japan. Some studies that have examined readability have used surveys that ask students or teachers to rate the difficulty of texts (e.g., Holster, Lake, & Pellowe, 2017). There are also professional practitioner organizations that provide reading list suggestions for learners at different grade levels and/or proficiencies. For example, the *Yomiyasusa Levels* (YL) are published biennially as a list of readability scores for English graded and leveled readers in Japan (Furukawa, 2007). This index depends on the interpretation of factors such as, "illustrations, the backgrounds of the books, the size of the fonts, and different text styles." The details of how such interpretations are operationalized have not been made transparent. Similarly, a second readability system, *The Kyoto Scale* (MReader, 2018), uses a combination of the "headword" numbers that are provided by publishers and information in the YL levels. The practice of using publisher's headword levels is problematic because there are inconsistencies in how headword levels are derived by different publishers (Claridge, 2012). To my knowledge, such readability systems have not been validated directly against the reading difficulty of texts. Furthermore, none of the above approaches provides an individual learner, teacher, or librarian with an ability to choose language-level-appropriate reading materials that have not already been vetted by outside organizations, which precludes the evaluation of the growing multitude of texts that are available for free on the Internet.

1.2.1 Direct measurement of text difficulty

The use of traditional reading comprehension tests to measure text difficulty was recognized as problematic very early on (Lorge, 1939) because the difficulty of the test questions and/or distractors confound the measurement of the difficulty of the text *per se* (for a recent review of this issue, see Cunningham, Hiebert, & Mesmer, 2018). For this reason, cloze tests have been used as the criterion variable for many readability formula studies in academia and in the publishing industry since the 1960s. Greenfield (1999) tested whether readability formulas that had been created for English-as-a-first-language children in the USA were also valid for adult EFL learners in Japan. He administered cloze tests to Japanese university students and then used the mean test scores as a criterion variable for evaluating different readability formulas. He found that the classic formulas did not correlate as well with Japanese readability scores (ranges of r = 0.691 to 0.861) as they did in U.S. K-12 contexts (Bormuth, 1971). He concluded that readability in Japanese EFL contexts should be treated as an overlapping, but unique construct.

1.2.2 The lexical component of text difficulty

Among the readability formulas tested in the Greenfield's (1999) study, only the New Dale–Chall readability formula (Chall & Dale, 1995) included a lexical component, which is calculated as the percentage of words not found on a list of 3,000 words known by 80% of U.S. grade four students. When Greenfield developed his own readability formula, the *Miyazaki EFL Readability Index*, he included the surface-level features of letters-per-word and words-per-sentence. While word and sentence length can be considered as proxies for lexical decoding and syntactic parsing, newer corpus-based and natural-language computational technologies allow the measurement of other factors that might represent such cognitive processes more closely (Baayen, Milin, & Ramscar, 2016). In a more recent study, Crossley, Greenfield, and McNamara (2008) used Greenfield's original cloze score data to develop models of readability using the psycholinguistic indexes available in the Coh-Metrix package (Graesser, McNamara, & Kulikowich, 2011). The Coh-Metrix analysis package has several word frequency indexes, all based on the Centre for Lexical Information (CELEX) corpus (Baayen, Piepenbrock, & Gulikers, 1995). In their final model, the CELEX written word frequency index was shown to correlate r = 0.61 with Greenfield's cloze test scores, and the combined readability model in this study, which included two additional text-indexes, was able to explain 84% of the variance (adjusted R^2).

2 The problem

The readability formulas that do exist for Japanese learners are lacking in several ways. Firstly, readability frameworks currently used in Japan depend on the ratings provided by publishers or by reading organizations. The methods by which publishers determine text reading levels have not been standardized or made available for scientific scrutiny. Similarly, other readability lists, such as the Yomiyasusa, are only published every 2 years and are based on an unpublished evaluation procedure, the reliability of which has not been determined. Secondly, to my knowledge, the few studies that have included a lexical component of readability for Japanese learners have not examined word-lists derived from different source corpora, registers and modalities for the purposes of optimization. Thirdly, commercially available readability frameworks used for the K-12 in the United States (e.g., the Lexile Framework; Stenner, Smith, & Burdick, 1983) are expensive, and they may not be valid for EFL learners. Finally, none of the available methods provides any diagnostic information that might be useful for teachers, learners, and/or material developers.

2.1 Proposal for readability criteria

I propose a set of criteria that the next generation of readability tools should meet. They should (1) be transparent in how they calculate a readability index for a text, (2) be available for use by teachers and learners at minimal or no cost, (3) be easy to use, (4) provide diagnostic information that is of pedagogical benefit to learners, teachers, librarians, and other stakeholders that would allow texts to be reliably rewritten to different readability levels, and (5) be empirically validated on the target population of learners' reading abilities.

One way that readability formulas could be developed toward satisfying these criteria would be to create a scale of linguistic forms (lexis, syntax, morphology, etc.) that could be used not only for text evaluation, but which could also be used to develop diagnostic language tests. In this way, learners could be matched to texts using a common scale of linguistic difficulty. I will now summarize one study that has examined how the lexical component of such a scale might be optimized for Japanese learners.

2.2 Call for a re-examination of the lexical component of readability

In an exploratory study (Pinchbeck, manuscript in preparation), I operationalized readability as the average cloze test scores of Japanese learners ($n = 200$) using previously published data (Bormuth, 1971; Crossley, Greenfield, & McNamara, 2008; Greenfield, 1999). Using Spearman correlations to relate the readability predicted by 26 different lexical indexes with the actual text difficulty as seen by the Japanese students, it was observed that word frequency/distribution rank-lists derived from narrative-texts, conversational-texts, and from TV/movie-captions were the best predictors of text readability (e.g., $\rho = 0.81$) as compared to written informational texts ($\rho < 0.5$). Using a Hotelling–Williams test for correlation differences (Steiger, 1980), the word list ranks derived from conversational and/or narrative registers of English were all significantly better predictors of readability than were published word lists based on general corpora of English, all of which have been promoted as general service lists of English (Brezina & Gablasova, 2013; Browne, Culligan, & Philips, 2013; Nation, 2012). These results are also consistent with a previous study that used vocabulary test-item difficulty as the criterion variable to compare word frequency lists (Pinchbeck, McLean, Brown, & Kramer, 2016).

An additional finding of this study was that the size of the corpus from which word frequency/distribution lists were derived was not found to be an important factor in predicting readability. An analogous finding was also reported by Brysbaert and New (2009), who examined word lists for their ability to predict lexical decision reaction times. They found that increasing the size of corpus samples of the British National Corpus (BNC) beyond 16 million tokens did not result in higher correlations with reaction times.

3 Discussion and Conclusions

The primary goal of this article was to illustrate how pedagogical instruments, such as readability formulas, might be developed, optimized, and validated for EFL learner populations. Many previous studies in readability have promoted cloze tests as a method that can be used to better estimate directly the difficulty of a candidate texts for a given learner population (e.g., Cunningham, Hiebert, & Mesmer, 2018).

The BNC, the Corpus of Contemporary American English (COCA), Ententen12, or the Cambridge English Corpus (CEC) are all large, general corpora of English that have been promoted as being representative of the English language. For these reasons, they have all been used by their proponents in the development of general-purpose word lists (Brezina & Gablasova, 2013; Browne, Culligan, & Philips, 2013; Nation, 2012), each of which have then been used as the basis for the creation of a number of other pedagogical tools such as specialized word lists or vocabulary tests. While the general principles that have driven these developments are sound, empirical validation of the pedagogical tools has been lagging, particularly with respect to how well these tools might work with different learner populations.

The results of the unpublished readability study cited here might be better interpreted in the light of Biber's work (e.g., 1993), who first challenged the notion that a single corpus that has been deliberately "balanced" can truly be used to represent all registers of language. His work and that of his students have detailed the multitude of linguistic differences that exist between different registers and modalities of English. Through this lens, it seems arbitrary that when attempting to measure readability in Japanese EFL contexts, we should choose to use word lists derived from source corpora that includes U.S. or U.K. newspapers and/or academic research articles, which are unlikely sources of language input for Japanese learners. While it is beyond the scope of this article to speculate on the possible psycholinguistic reasons why corpora of fiction or of captions of TV/movies for children might better represent the lexical knowledge of Japanese learners, it nevertheless raises the question of whether a single index of lexical difficulty should be used to fit all pedagogical purposes, for all learner populations, and in all teaching contexts.

The necessary inclusion of a lexical component in any automatic readability system has several benefits, as follows. Firstly, word lists "provide a rational basis for making sure that learners get the best return for their vocabulary learning effort" (Nation & Waring, 1997, p.17). Word lists, based on corpus frequency, or some other psycholinguistic index, can be used as a proxy for a scale of lexical difficulty, which means that word lists can be used as strong predictors of text difficulty. Alternatively, such word lists can also inform how lexical items might be chosen from existing course materials for targeted instruction because they assist developers to avoid teaching words that learners already know. Additionally, word lists also allow materials developers to simplify texts using a principled approach that is likely to produce texts that can be used for a wider range of learner-proficiencies. These types of text manipulations are not possible when only surface textual features are modified. For example, text length (i.e., tokens) was reported to be the major factor of readability in a recent study of Japanese students (Holster, Lake, & Pellowe, 2017). However, we cannot purposefully make the text of a book easier to read by cutting a book in half; as the first half of a typical book is just as difficult to read as the second half. Finally, word lists also represent how useful individual lexical items are likely to be to learners, inside or outside the language classroom. When word lists are included as terms in readability formulas, it allows language programs and/or materials developers to better match learners to texts because the same lexical scales of difficulty that are used for all of the above purposes can also be used to create diagnostic vocabulary tests that measure learner knowledge.

Automatic readability programs are limited because they examine texts separately from their contexts of use and from a complex range of factors related to individual learners, such as learner interest, values, and motivation. Therefore, the experience and intuitions of developers, librarians, teachers, and learners will continue to be necessary in text selection decisions. The next generation of automatic readability technologies will show their value by allowing teachers and materials developers to more efficiently access and evaluate the wealth of materials increasingly available on the Internet. A range of classroom- or learner-suitable texts that match both the thematic content of learning modules and the lexical knowledge of learners might be identified much more efficiently than that which is currently possible. The readability frameworks that are currently in use in Japan would make this a very slow process indeed. The Yomiyasusa readability lists are published only on a biennial basis and rely on the judgment of a committee. Similarly, The Kyoto Scale readability depends on publishers to level their own products and to make that information public. Neither of these approaches allows users to quickly determine the appropriateness of new candidate texts for their own learners.

Using computers, it is not currently possible for any practitioners to easily and accurately analyze the deeper structural and semantic features of texts (i.e., dialect; rhetorical structure; topic background, specificity, and complexity; cohesive devices; allusion; metaphor; the level of inferencing required of the reader; and/or formulaic language; etc.). In order for these textual features to be included in automatic readability estimation, we must await further advances in natural language processing technologies. In the meantime, however, I propose that a re-examination of the lexical component of readability is warranted, particularly in Japan and in other EFL teaching contexts. It is my hope that this article might generate some interest toward the development of large-scale validity studies that would develop new pedagogical tools for more valid and efficient matching of Japanese K-12 and/or adult EFL learners to English texts.

References

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, *30*(11), 1174–1220. doi:10.1080/02687038.2016.1147767

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Biber, D. (1993). Representativeness in corpus design. *Journal of Literary and Linguistic Computing*, *8*(4), 243–257. doi:10.1093/llc/8.4.243

Bormuth, J. R. (1971). *Development of standards of readability: Toward a rational criterion of passage performance (Office of Education, U.S. Department of Health, Education, and Welfare No. 9–0237)*. Chicago, IL: University of Chicago.

Brezina, V., & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, *36*, 1–23. doi:10.1093/applin/amt018

Brown, J. D. (1998). An EFL readability index. *JALT Journal*, *20*(2), 7–36. Retrieved from http://hdl.handle.net/10125/40779

Browne, C., Culligan, B., & Philips, J. (2013). *A new general service list*. Retrieved from http://www.newgeneralservicelist.org

Bruce, B., Rubin, A., & Starr, K. (1981). Why readability formulas fail. IEEE Transactions on Professional Communication, PC-24(1), 50–52. doi:10.1109/TPC.1981.6447826

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. doi:10.3758/BRM.41.4.977

Carrell, P. L. (1987). Readability in ESL. *Reading in a Foreign Language*, *4*(1), 21–40. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl41carrell.pdf

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Cambridge, MA: Brookline Books.

Claridge, G. (2012). Graded readers: How the publishers make the grade. *Reading in a Foreign Language*, *24*(1), 106. Retrieved from http://nflrc.hawaii.edu/rfl/April2012/articles/claridge.pdf

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology*, *11*(3), 38–63. Retrieved from https://scholarspace.manoa.hawaii.edu/bitstream/10125/44117/1/11_03_cobb.pdf

Cobb, T. (2008). Commentary: Response to McQuillan and Krashen (2008). *Language Learning & Technology*, *12*(1), 109–114. Retrieved from https://scholarspace.manoa.hawaii.edu/bitstream/10125/44134/1/12_01_cobb.pdf

Connatser, B. R. (1999). Last rites for readability formulas in technical communication. *Journal of Technical Writing and Communication*, *29*(3), 271–287. doi:10.2190/6EWH-J5C5-AV1X-KDGJ

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using psycholinguistic indices. *TESOL Quarterly*, *42*, 475–493. doi:10.1002/j.1545-7249.2008.tb00142.x

Cunningham, J. W., Hiebert, E. H., & Mesmer, H. A. (2018). Investigating the validity of two widely used quantitative text tools. *Reading & Writing*, *31*, 813–833. doi:10.1007/s11145-017-9815-4

Davison, A., & Kantor, R. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, *17*, 187–209. doi:10.2307/747483

Duffy, T. M. (1985). Chapter 6 - Readability Formulas: What's the Use? In T. M. Duffy & R. Waller (Eds.), *Designing Usable Texts* (pp. 113–143). Toronto: Academic Press Inc. Retrieved from https://doi.org/10.1016/B978-0-12-223260-2.50011-6

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. doi:10.1037/h0057532

Furukawa, A. (2007). *Yomiyasusa level: A reading level for Japanese students.* Retrieved from https://www.seg.co.jp/sss/YL/What_is_YL.html

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223–234. doi:10.3102/0013189x11413260

Greenfield, G. R. (1999). *Classic readability formulas in an EFL context: Are they valid for Japanese speakers?* (Ed.D.). Temple University, PA, USA. Retrieved from https://search.proquest.com/docview/304536830/abstract/CB62966525BF4987PQ/1

Greenfield, J. (2004). Readability formulas for EFL. *JALT Journal*, *26*(1), 5–24. Retrieved from https://jalt-publications.org/files/pdf-article/jj-26.1-art1.pdf

Holster, T. A., Lake, J. W., & Pellowe, W. R. (2017). Measuring and predicting graded reader difficulty. *Reading in a Foreign Language*, *29*(2), 218. Retrieved from http://nflrc.hawaii.edu/rfl/October2017/articles/holster.pdf

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl131hsuehchao.pdf

Kincaid, J. P., Fishburne, J., 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. Retrieved from http://www.dtic.mil/docs/citations/ADA006655

Lange, B. (1982). Readability formulas: second looks, second thoughts. *The Reading Teacher*, *35*(7), 858–861. Retrieved from https://www.jstor.org/stable/20198110

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15–30. Retrieved from http://nflrc.hawaii.edu/rfl/April2010/articles/laufer.pdf

Lorge, I. (1939). Predicting reading difficulty of selections for children. *The Elementary English Review*, *16*(6), 229–233. Retrieved from http://www.jstor.org/stable/41383105

Maxwell, M. (1978). Readability: Have we gone too far? *Journal of Reading*, *21*(6), 525–530. Retrieved from http://www.jstor.org/stable/40010923

MReader. (2018). The Kyoto Scale. Retrieved June 6, 2018, from https://mreader.org/mreaderadmin/s/html/Kyoto_Scale.html

Nation, P. (2012). *The BNC/COCA word family lists*. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Information-on-the-BNC_COCA-word-family-lists.pdf

Nation, P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language*, *26*(2), 1–16. Retrieved from http://www.nflrc.hawaii.edu/rfl/October2014/articles/nation.pdf

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage, and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6–19). New York: Cambridge University Press.

Pinchbeck, G. G., McLean, S., Brown, D., & Kramer, B. (2016). Part 2, Revisiting the Word Family: What is an Appropriate Lexical Unit for Japanese EFL Learners? In R. Waring, L. Anthony, C. Browne, & T. Ishii (Eds.), *Vocabulary Learning and Instruction (Vocab@Tokyo: Current Trends in Vocabulary Studies)* (pp. 26–28). Tokyo, Japan. VLI: A Journal of Vocabulary and Research. Retrieved from http://vli-journal.org/vocabattokyo/vocabattokyo_handbook_2016.pdf

Rankin, E. F. (1965). The cloze procedure - a survey of research. In L. T. Thurstone (Ed.), *Yearbook, National Reading Conference* (pp. 50–93). Milwaukee: National Reading Conference.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43. doi:10.1111/j.1540-4781.2011.01146.x

Selzer, J. (1981). Readability is a four-letter word. *The Journal of Business Communication (1973)*, *18*(4), 23–34. doi:10.1177/002194368101800403

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251. doi:10.1037/0033-2909.87.2.245

Stenner, A. J., Smith, M., & Burdick, D. S. (1983). Toward a theory of construct definition. Journal of Educational Measurement, 20, 305–316. doi:10.1111/j.1745-3984.1983.tb00209.x