Articles

Readability Formulas For EFL

Jerry Greenfield
Miyazaki International College

EFL/ESL teachers use English readability formulas to match texts to their students' reading levels. However, the formulas' validity for EFL/ESL use has gone largely untested. Two studies have now addressed this issue, with divergent results. Brown (1998) found that classic formulas were not very accurate predictors of EFL difficulty, while Greenfield (1999) found that they predicted for EFL about as well as they did for native English readers. Both studies produced accurate EFL readability formulas. In the analysis presented here, the difference in the two studies' findings is attributed to Brown's random passage set. Brown's formula proves more accurate with the other study's passages than with his own, agreeing with observed EFL difficulty and predictions by classic formulas. This supports the finding that the classic formulas are valid for EFL use.

EFL/ESL教師は英語読解難易度推定式を使って学生の読解力水準に合った英文を選定する。しかし、これまでこの推定式のEFL/ESLにおける有効性については殆ど検証されてこなかった。この問題については、現在までに二つの考察がなされたが、それぞれの結果は異なっている。ブラウン(1998)の考察では、標準的推定式によるEFL難易度の推定精度はあまり高くないとしているが、グリーンフィールド(1999)は、EFL及び母語読者双方に対するその読解力推定精度はほぼ同等であるとしている。これら二つの考察は、それぞれ正確なEFL読解難易度推定式を導き出している。二つの知見の違いは、ブラウンの無作為に選択した文節セットに帰因すると考えられる。ブラウンの推定式の精度は、自らが選んだ文節セットに対してよりも、グリーンフィールドの考察で使われた文節に対しての方が精度が高く、またそれは実際のEFL難易度と標準的推定式から得られる推定結果とも一致している。そしてこのことは、標準的推定式のEFLに対する有効性を裏付けるものである。

EFL/ESL teachers along with other English teachers have long turned to readability formulas for aid in matching texts to students' reading levels. Until recently little attention has been paid to whether it is appropriate to apply these tools outside the native English contexts in which the formulas were originally developed. This question has now begun to be investigated, but with mixed results. A study by J. D. Brown (1998) found that several classic formulas were not very accurate in predicting EFL reading difficulty for Japanese university students. His conclusions cast serious doubt on the validity of the classic readability formulas for EFL use. Instead Brown proposed a new formula of his own that he found to be more accurate with his readers. Unfortunately, that formula is difficult to use, requiring long-word and passage-frequency word counts in addition to parsing into function and non-function words. Brown's results thus leave EFL teachers without an easy and reliable way to estimate a text's readability for their students.

At the time Brown's article appeared, another study (Greenfield, 1999) with another group of Japanese EFL readers was being completed. It found that the classic formulas discriminated text difficulty for those readers about as well as they do for native English readers. The study produced a formula scaled to those Japanese readers which is easy to use but offers only a marginal improvement in predictive accuracy over the traditional formulas, which themselves proved to be quite satisfactory in the EFL as well as native speaker contexts. While the results of this study are encouraging, they appear to disagree with Brown's on the question of the fundamental validity of applying the classic formulas in EFL contexts. Where does this leave us regarding this question and the viability of the proposed EFL formulas? It will be useful to compare the two studies more closely.

Readability Formulas

Very simply, readability formulas are multiple regression equations in which the dependent variable (the value we want to know) is the reading difficulty predicted of a text and the independent or predictor variables are two or more directly measurable characteristics of the text, such as the number of letters per word and the number of words per sentence. To use one of these formulas, you measure the independent variables in a piece of text, plug those values into the formula, do the math, and get a prediction of the text's difficulty expressed as a grade level, a cloze score, or a score on some set scale. Dozens of formulas

have been introduced, and some of the most popular formulas, such as the Flesch (1948) and Dale-Chall (Chall & Dale, 1948), have been around since the 1940s (for an overview of readability formulas and their histories see Chall, 1958, 1988; Klare, 1963, 1974-75, 1988). For a time during the 1980s, readability formulas came under attack because of their low face validity when viewed from the vantage point of psycholinguistic theories of reading (Bruce & Rubin, 1988; Rubin, 1985; Bruce, Rubin, & Starr, 1981; Smith, 1988). Nevertheless, the formulas have survived and are still widely used on account of their consistently high predictive validity (Chall & Dale, 1995; Fry, 1989). That is to say, they have been found empirically to do a good job of discriminating text difficulty even though it is not obvious why they should or how they could. In fact, the advent of computer word processing has made the formulas more accessible than ever. In older versions of Microsoft Word, it was possible to get a readability report on a Word document using the Flesch, Flesch-Kincaid, Coleman-Liau, and Bormuth formulas built into that application. With Microsoft Word 97/98, the number of formulas was cut back to include only the Flesch and Flesch-Kincaid, but they are there waiting to be used at the click of a mouse. The question is, are they valid to use for EFL/ESL?

Although there are many different variables that have been identified as playing a part in reading difficulty (Gray & Leary, 1935), factor analysis has narrowed these down to only a few which have high correlations with the others and so can be used to represent them. The predictor variables in classic readability formulas typically represent just two main text factors: vocabulary difficulty and grammatical difficulty. Depending on the formula, vocabulary difficulty may be represented as word familiarity, average word length in syllables, proportion of long words, average word length in either characters or syllables, or proportion of monosyllable words. Grammatical difficulty is typically measured by the average number of words or syllables per sentence, based on a strong association of sentence length with, for example, the incidence of compound-sentence and embedded-clause constructions, which are much harder to count. Proposed approaches to readability measurement using variables that are not so easily countable have not been widely adopted. Nor do they need to be, since research has found them not to deliver significantly better results than formulas with simpler variables (Bormuth, 1969, 1971; Chall & Dale, 1995).

Generally the predictive accuracy of the most commonly used formulas has been found to be very high, yielding correlations with inde-

pendent comprehension tests in the .8 or .9 range (Chall, 1958; Chall & Dale, 1995; Fry, 1989). Readability formulas assume that the readers for whom they predict difficulty and the texts to which they are applied are similar to the samples used to derive the formulas in the first place. Indeed, almost without exception the formulas have been validated by testing American native English readers. Using the formulas to predict difficulty for second language readers assumes that those readers are not significantly different from native readers in ways that affect how the measured text variables relate to reading difficulty. Surprisingly, this assumption has been left essentially untested.

The question might be pursued on theoretical grounds, if we had a detailed enough model of second language reading. However, the theoretical issues in this area are many and complex, and empirical studies to investigate them have not yet produced a clear and comprehensive account of second language reading and how it is similar to or different from first language reading. (For an overview of the state of the research up to 1990, see Bernhardt, 1991; see also Grabe, 1993; Paran, 1996. For a sampling of recent studies, see Carrell & Wise, 1998; Parry, 1996.) A new survey is overdue. In any case, arguing from a theoretical model can only make the case for or against whether native reader-based formulas *ought* to work with second language readers. Such arguments cannot establish in fact whether they *do* work.

Validating Formulas for EFL

Fortunately, it is not necessary to resolve the theoretical issues in order to determine whether the formulas historically based on L1 reading data are also valid for EFL/ESL readers. Readability formulas represent statistical correlations and *predict* difficulty rather than *explain* its causes. Formula validity depends simply on the accuracy of predictions. This can be determined for EFL/ESL readers empirically by testing them to see how closely their performance matches what formulas predict. The remarkable fact is that this was not done to settle the matter years ago.

Hamsik's Study

In fact, a small-scale study was done in 1984 by Hamsik, who investigated the ESL validity of the Flesch, Dale-Chall, Fry, and Lorge readability measures. Hamsik gave cloze tests on 18 academic passages to 40 Intensive English Center students at an American university. The students

are described as being from the Middle East, South America, and "the Far East." Hamsik found significant positive correlations of .775 to .819 between the rank orders of difficulty of the passages as indicated by the cloze scores and as predicted by each of the four readability measures. On the strength of this evidence, Hamsik concluded that "the four readability formulas and graphs...do measure readability [for] ESL students and that they can be used to select material appropriate to the reading level of ESL students" (p. iv).

Hamsik's small heterogeneous sample of ESL readers did not permit discriminating any effect of first language background. With this in mind, Hamsik included among her recommendations one that "future studies of this sort should take account of L1 background" (p. 55). She also suggested that it might be possible to develop a readability index for ESL students that would be more accurate than existing formulas.

Brown's Study

Further investigation of ESL/EFL formula validity was not forthcoming until Brown's 1998 article. In an earlier study of cloze item difficulty, Brown (1992) had administered cloze tests to nearly 2300 Japanese EFL university students. For the new study he reanalyzed the data for difficulty at the passage level and compared the observed mean cloze scores on the passages with scores predicted by six readability measures: the Flesch, Flesch-Kincaid, Fry Graph, Gunning, Fog Count, and Gunning-Fog. Brown found Pearson correlations ranging only from .48 to .55, leading him to conclude, "first language readability indices are not very highly related to the EFL difficulty" (p. 27).

To address this need, Brown developed a new formula using his observed EFL scores as the criterion, scaled to yield an EFL Difficulty Index ranging from 1 to 92. Multiple regression analysis found the best fit or most accurate prediction to be made using four text variables: syllables per sentence, passage frequency (how many times the deleted item appears elsewhere in the text), percentage of long words (seven or more letters), and the percentage of function words. The resulting formula, which he called the EFL Difficulty Estimate, had a multiple correlation of .74, which yielded an adjusted R-Square or coefficient of determination of .51.[1] Usually a coefficient in that range would not be considered particularly strong. Nonetheless, because his formula yielded a stronger correlation with the observed EFL scores than did the classic formulas in his tests, Brown speculated, "EFL/ESL readability might best be esti-

mated separately for students from different language backgrounds" (p. 30). In other words, Brown suggests, we need to replace the classic readability formulas with new formulas specific to different language groups. His formula was offered as one that might be used with Japanese EFL. Brown's formula is as follows:

Brown EFL Difficulty Estimate

EFL Difficulty = 38.7469 + (.7823 x Syllables per Sentence)
+ (-126.1770 x Passage Frequency)
+ (1.2878 x % Long Words)
+ (.7596 x % Function Words)

(R= .74, adjusted R^2 = .51, SE = 19.68, N = 50, p < .00001)

Again, the formula is scaled to predict passage difficulty from zero to a maximum difficulty of 100. Something is amiss, however, because when this formula was applied to Bormuth's standard passage set (Bormuth, 1971) assembled for that researcher's testing and used again to calibrate the New Dale-Chall readability formula, Brown's formula predicted difficulty scores ranging from -276 to +64.

The Miyazaki Study

The Miyazaki study (Greenfield, 1999) also involved Japanese university students and checked the Flesch Reading Ease and Flesch-Kincaid formulas along with the Coleman-Liau, New Dale-Chall, and Bormuth formulas. The EFL participants in this study were 200 Japanese students enrolled in a small liberal arts college in western Japan. Careful randomized testing procedures were followed, based on Bormuth (1971). Fifth-word deletion cloze tests were constructed on 31 of the 32 Bormuth academic passages. One passage was read by all participants as a control, and one was omitted for a balanced design. Pearson correlations between observed EFL mean cloze scores and scores predicted by the formulas are .691 for the New Dale-Chall formula, .765 for Coleman-Liau, .845 for Flesch Reading Ease, .847 for Flesch-Kincaid, and .861 for Bormuth. These results are shown in Table 1.

Bormuth left us his set of mean cloze scores for his 1971 passages, so it is possible to compare the observed EFL scores and Bormuth's native English reader criterion for the same passages.[2] That correlation is even stronger at .915. These correlations are generally consistent with

Table 1. Pearson Correlations Between Scores for Bormuth Passages Predicted by Miyazaki EFL Index, Original and Recalculated Brown Formulas, and Classic Formulas

	Observed EFL	Miyazaki EFL Index	Brown (Original)	Brown Recalculated
Flesch	-.845	.957	-.764	.902
Flesch-Kincaid	-.847	.980	.775	-.909
Coleman-Liau	-.765	.927	.736	-.850
Bormuth 1969	.861	.977	-.775	.913
Dale-Chall 1995	.691	.790	.654	-.820
Brown Recalculated	.907	.825	-.926	
Bormuth 1971	.915	.944	-.781	.890
Observed		.941	-.820	.895

All relationships significant at $p < .0001$, N = 31. The negative sign can be ignored as an artifact of contrasting scales.

inter-formula correlations and predictive accuracies reported in the L1 literature (e.g. Chall & Dale, 1995). These findings support the conclusion that the classic formulas are indeed fundamentally valid for a broad spectrum of English readers that includes non-native as well as native readers. In other words, the formulas work quite well to predict the relative EFL/ESL difficulty of English academic texts.

What remained was to check whether recalculating the classic formulas using the EFL scores would significantly improve accuracy with EFL readers. While some of the recalculations result in a small but statistically significant improvement, the gain did not seem sufficient to justify substituting them for the originals. Going a step further, a comprehensive check of all of the classic variables found that a regression of just two, letters per word and words per sentence, against the study's EFL criterion produced an EFL difficulty index that was as good as or slightly better than any of the classic formulas. The ANOVA and coefficients for that multiple regression are shown in Table 2.

The new formula, for convenience called the Miyazaki EFL Readability Index, turned out to be only marginally more accurate than the classic formulas. However, like Brown's, it has the practical advantage of

Table 2. ANOVA and Regression of Brown's Four Independent Variables Forced vs. Bormuth EFL Criterion

	DF	Sum of Squares	Mean Square	*F* Value	*p*-Value
Regression	4	4093.682	1023.420	29.990	< .0001
Residual	26	887.272	34.126		
Total	30	4980.954			
	Coefficient	Std. Error	Std. Coeff.	*t*-Value	*p*-Value
Intercept	33.232	9.649	33.232	3.444	.0020
Syllables per Sentence	-.249	.205	-.177	-1.216	.2351
Passage Frequency	12.834	3.045	.486	4.215	.0003
Long Words	-48.665	21.996	-.327	-2.212	.0359
Function Words	-65.650	26.595	.225	-2.468	.0205

N = 31

being scaled for EFL readers while, unlike Brown's, being simple to apply using easily found word counts. The Miyazaki formula is as follows:

Miyazaki EFL Readability Index

EFL Difficulty = 164.935 − (18.792 x Letters per Word)
− (1.916 x Words per Sentence)

(R= .862, adjusted R^2 = .723, SE = 10.558, N = 31, p < .0001)

Note that this formula delivers a reading ease score on a nominal 100-point scale, 100 being easiest.

Comparison of Brown's and the Miyazaki Study

On the face of it, Brown's study and the Miyazaki study seem to have come to contrary conclusions regarding the validity of classic readabil-

ity formulas for EFL use, leaving the matter still unresolved. However, a closer look reveals that the disagreement is less direct than it seems at first glance. In spite of obvious parallels in the research designs, there are important differences in what the two studies were actually testing.

Participants

Let us look first at the population samples. It seems reasonable to assume that the groups participating in the two studies were similar to each other except with respect to their size. Brown's group was comprised of 2,298 participants distributed across 18 universities in Japan, while the 200 participants in the Miyazaki study were from a single college. While ordinarily the larger sample is advantageous for statistical purposes, in this case there is no reason to suppose that the Miyazaki group was in fact importantly different from the group in Brown's study. The two population samples were each internally homogeneous and were similar to each other in first language, cultural background, and general educational level.

Brown tells us little about the English proficiency of his group beyond the fact that they were representative of Japanese university EFL students in this regard. A similar claim is made for the Miyazaki students. However, even if the two groups had differed in their English proficiency levels, there is no reason to suppose that the Miyazaki group was higher than Brown's rather than the other way around. In any case, it is the difference in accuracy rates obtained in the two studies—Brown's group scoring much lower overall than the Miyazaki group—that begs to be explained. More importantly it must be wondered why there is such a large difference in correlations with predictions by the classic formulas. On balance it seems very unlikely that a difference in the English proficiency of the two EFL population samples or any other differences in the two population samples could explain the large differences in the correlations between formula predictions and observed difficulty found in the two studies.

Passage Sets

The more likely source of the disagreement between the two studies lies in differences in the nature of their passage sets arising in turn from difference in the purposes of the studies. The Miyazaki study was looking at whether the classic formulas' ability to discriminate relative

difficulty is different for EFL and native English readers. To answer this question, it was desirable to compare EFL performance with native English reader performance while keeping the passage set constant. Using Bormuth's passages made it possible to compare observed EFL difficulty with his observed native English reader scores as well as with the New Dale-Chall formula predictions based on the same passages. Using the Bormuth passages with an EFL group focused squarely on whether the text variables *for those texts* relate to EFL difficulty the way they do for native English readers. The issue of the formulas' applicability to other kinds of texts was not addressed.

Brown, on the other hand, chose not to control for difference in passage type but was checking the universality of the classic formulas for predicting the difficulty of randomly selected texts. His method of selecting passages was quite different from that followed by Bormuth and other classic readability researchers. Typically a criterion passage set is deliberately chosen to exhibit a well-distributed range in the values of text variables to be regressed against test scores. Instead, Brown's passages were randomly selected to be "representative samples of the English language, at least the English language written in the books found in a U.S. public library" (Brown, 1998, p. 16). Brown verified their representativeness by comparing their lexical frequencies with frequencies published for the English corpus, finding them to correlate at .93. The ostensible advantage of such an approach is that it tests the EFL validity of the classic formulas not for only *academic* materials but rather for *any* English texts. To suppose that such a set is superior is to suppose that the academic materials traditionally used may not be fully representative of English texts at large. In effect, Brown concluded that *when applied to more generally representative texts than the kinds of academic texts on which they are based*, classic readability formulas do not work very well for EFL readers. Thus, Brown's finding does not directly contradict that of the Miyazaki study on the more specific question as to whether the formulas might, however, be valid for EFL readers when applied to academic texts.

Brown's conclusion is complicated by the fact that he was concerned with the issue of text representativeness as well with the question of whether the first language of the readers makes a difference in formula validity. As a result, it is not clear whether the inability of the formulas to predict the scores he observed was due to the fact that his readers were second language readers or to the fact that his texts were randomly selected, or some combination of both. Brown's suggestion that different formulas need to be developed for different EFL/ESL language

backgrounds implies that he believes that language background was the more important difference. However, since no one (to my knowledge) has ever provided a native reader difficulty criterion based on a random passage set, there is nothing against which to compare the EFL results to answer that question, and so Brown's data must be regarded as ultimately inconclusive on this point.

It might be argued that using a random passage set corrects for a shortcoming in the classic tradition of readability research, namely that that tradition has focused on a too narrow range of text types and, thus, has failed to achieve full generalizability. Such an argument is unconvincing on two counts. First, the principal use of readability formulas generally is in education-related contexts, where they are used not only by teachers but even more importantly by writers and publishers of educational materials. No claim is made that the various criterion passage sets used to derive the classic formulas are or should be representative of the English language in general. In fact, Bormuth deliberately selected his passages to represent the range of content, structure, style, and vocabulary specifically found in school texts. The decision to select criterion passages to represent educational materials in this way does not, on that account, limit the practical usefulness of formulas derived from them. On the contrary, there is little point in creating readability formulas for other kinds of texts. This is especially so for formulas to be used in EFL contexts.

Second, as Brown himself had already pointed out in his earlier article on natural cloze tests (Brown 1993), there is no assurance that a randomly selected set of texts will provide the variability in the text features needed to discriminate difficulty, regardless of how representative their combined lexical frequencies may be. In fact, Brown's data suggest that the passages did not work especially well for this purpose. The range of mean cloze scores for the passages was relatively flat and the passage scores overall very low (mean = 13.7%), with fully 40% of the mean passage scores falling below 10% accuracy. Brown attributes these low accuracy rates to the nature of cloze tests. However, in the Miyazaki study only 10% of the mean passage scores fell below 10% accuracy, with the range of mean raw scores much greater and the mean score for the set also considerably higher at 24.25%. The evidence suggests that Brown's randomly selected passages may have been too difficult overall to yield a robust difficulty variable to compare with formula predictions. To make the scores more suitable for regression analysis, Brown scaled them in such a way that the variability was statistically magnified. Although this

is a perfectly reasonable procedure, it may explain the out-of-range results found when his resulting formula is applied to a passage set more variable than his random set apparently was.

Brown's EFL Difficulty Estimate

This brings us back to Brown's formula. Although it correlates only weakly with the classic formulas, Brown's EFL Difficulty Estimate nonetheless is moderately strong in discriminating the readability of his passages for his representative sample of Japanese EFL students. But there apparently were problems with the passage set, and anyway it is unknown how that set relates to typical academic texts. We might therefore ask how the formula works for academic texts as represented, for example, by Bormuth's passages. This question was answered by applying the EFL Difficulty Estimate to those passages and comparing the results with predictions by the classic formulas, with Bormuth's native reader scores, and with the Miyazaki EFL scores on the passages. Those correlations are included in Table 1.

The results are a little surprising. The correlations between scores predicted for the Bormuth passages by classic formulas and scores predicted by Brown's formula range from .654 to .832, with Flesch at .766, Flesch-Kincaid at .778, and only the correlation with New Dale-Chall falling below .734. These are much higher than Brown found with his own passages. When Brown's estimates for the Bormuth passages were compared with the observed Miyazaki EFL mean cloze scores, the resulting correlation of .841 was comfortably consistent with correlations ranging from .691 to .861 found between the observed EFL difficulty and classic formula scores for those passages. In other words, Brown's formula worked about as well on the Bormuth passages read by the Miyazaki students as do the classic formulas and the Miyazaki formula.

It remained only to see whether Brown's formula could be improved further by recalculating its coefficients using the Miyazaki scores. This involved performing a new multiple regression using the four variables used in Brown's model. The ANOVA and coefficients for the four-variable regression are shown in Table 3.

Normally when performing a multiple regression, variables are entered or removed one at a time to discover which combination results in the strongest multiple correlation, or best fit, with the fewest variables. Since in this case all four variables are being forced into the equation, there is a possibility that one or more variables might be extra bag-

Table 3. ANOVA and Regression of Two Independent Variables vs. Miyazaki EFL Criterion (Miyazaki EFL Readability Index)

	DF	Sum of Squares	Mean Square	*F* Value	*p*-Value
Regression	2	8988.565	4494.282	32.874	< .0001
Residual	28	3121.061	111.466		
Total	30	12109.626			
	Coefficient	Std. Error	*t*-Value	*p*-Value	
Intercept	164.935	18.758	8.793	<.0001	
Letters per Word	-18.792	4.940	-3.803	.0007	
Words per Sentence	-1.916	.484	-3.793	<.0005	

*N = 31

gage, adding nothing to the strength of the regression. In fact, the three word-frequency variables were well within bounds, but the syllables-per-sentence variable failed the relatively liberal .1 probability limit to add/remove adopted by Brown in his (more proper) stepwise regression procedure. Removing this variable improved the probability of the remaining three variables and only very slightly reduced the adjusted R Squared from .794 to .791, which is very strong. The four-variable recalculated formula is as follows:

Recalculated Brown EFL Difficulty Estimate

Cloze = 33.232 + (−.249 x Syllables per Sentence)
+ (12.834 x Passage Frequency)
+ (−48.665 x % Long Words)
+ (−65.650 x % Function Words)

(R = .907, adjusted R^2 = .794, SE = 5.842, N = 50, p < .0001)

Note that EFL difficulty in the recalculated formula is given as a predicted cloze score and is not rescaled as in Brown's original formula.

While the three-variable version is slightly more accurate with the sentence-length variable dropped out, it has no variable ostensibly rep-

resenting a syntactic factor. The function-word variable, which might seem to be related to syntax, in fact is correlated in these passages only .174 with syllables per sentence, which is ordinarily taken to be a syntax variable and in this case is itself correlated .766 with observed difficulty. Since predictions by the two versions are nearly perfectly related at .994, it is redundant but otherwise does no harm to retain the four-variable recalculation as the more faithful to Brown's original model. Note that this recalculated formula has an adjusted R Squared of .794, much higher than the .51 achieved by Brown's original regression.

We may then compare predictions of the recalculated Brown formula with the observed EFL scores, Bormuth's 1971 scores, predictions of the original Brown formula, and predictions of each of the classic formulas. These correlations are also included in Table 1. In some of the corresponding correlations it is readily apparent that one correlation is stronger than its counterpart. At the same time, it is not obvious, particularly in pairs, which are closer in value, whether a given difference is significant. There is a test, the Williams *t*-test, that finds whether a difference between two related correlations is significant.[3] In this case it allows us to determine whether the Miyazaki formula, the original Brown formula, and the recalculated Brown formula are equally good, better, or worse as predictors of cloze performance in comparison with each other and with each of the classic formulas.

The results of the Williams *t*-test are shown in Table 4. The recalculated Brown formula has a small but statistically significant advantage over Brown's original formula, and both are superior to all of the classic indices except the Bormuth formula. The comparison shows the Miyazaki Index to correlate more strongly than Brown's formula, but the difference was not statistically significant.

The bottom line is that we have two new formulas developed from EFL data that appear to work very well in predicting the relative difficulty of academic texts. At the same time we have strong evidence that the new formulas have only a narrow, if any, advantage over the time-tested traditional formulas, especially the Flesch and Flesch-Kincaid, and Bormuth formulas. We may therefore use those formulas with some new confidence that they are valid for EFL. By extension, if they are valid for a first language group as different from English as Japanese, they are probably valid for other EFL and ESL contexts as well. Brown's own findings do not directly contradict this conclusion, if we understand that his passage set is not strictly comparable and may not have been appropriate for basing a measure of academic readability.

Table 4. Williams *t*-test: Recalculated Brown Formula versus Classic, Original Brown & Miyazaki EFL Formulas

Named Formula	Recalculated Brown vs Observed EFL	Named Formula vs Observed EFL	Named Formula vs Recalculated Brown	Williams *t*-value
Flesch	.907	-.845	.862	4.061
Flesch-Kincaid	.907	-.847	.864	4.005
Coleman-Liau	.907	-.765	.820	6.777
Bormuth 1969	.907	.861	.807	1.893
Dale-Chall 1995	.907	.691	.591	3.939
Original Brown	.907	.820	.915	8.573
Miyazaki EFL Readability Index	.907	.941	.808	2.462

Underlined values are significantly larger in that row's comparison at $p < .01$, with $t = 2.762$ needed for significance (2-tailed, $df = 28$). The negative sign can be ignored as an artifact of contrasting scales.

Applications

If the old formulas are already valid, is there any point in introducing a new formula that is not significantly or importantly more accurate? The answer is not as simple as it might seem. Predictive accuracy is only part of the story. Along with accuracy, we need to consider how easy any index is to apply. Bormuth took account of this when he constructed different formulas for hand scoring, computer calculation, and unrestricted research use. The Coleman-Liau formula (Coleman & Liau, 1975) is specifically intended for computer calculation. For application in the field, the ease of a simpler hand scoring formula compensates for a small loss in accuracy. Part of the attractiveness of the Flesch and Flesch-Kincaid (Kincaid, Fishburne, Rogers, & Chissom, 1975) formulas has been their ability to deliver accurate results with just two simple variables that are easy to count and calculate. Of course, this advantage disappears with computerized applications. However, it is not realistic to expect we would ever have an EFL formula built into Microsoft Word. Fortunately, the Miyazaki results indicate that the Flesch and Flesch-Kincaid formulas included in that application already can serve our needs.

Viewed against this availability, the complexity and difficult-to-count variables of Brown's formula make hand scoring with that formula not very attractive. The Miyazaki formula uses only two variables that not only are easy to count but are reported along with the readability report in Microsoft Word. Other word-processing applications with or without readability measurement also provide character, word, and sentence counts, making hand scoring any passage a simple affair. To make it even easier to use the Miyazaki index, the Miyazaki study provides a lookup table of scores (Greenfield, 2003) for a practical range of word- and sentence-length values, so no calculation is needed. Once you have these values, whether by counting or by using a word processor, it is straightforward to locate a score on the table that represents the EFL readability of the passage. The score is figured on a 100-point scale, with 100 being easiest and 50 representing a text of average difficulty for EFL students.

What these scores actually mean, however, is not so straightforward. EFL students achieved a mean cloze accuracy of about 27% on a text having a Miyazaki EFL Readability Index of 50. No one knows for sure what that means in terms of performance criteria for EFL reading. The conventional wisdom for native readers is that a score of 35% corresponds to a score of 50 on a well-constructed multiple-choice comprehension test, 45% corresponds to an MC score of 75, and 55% to an MC score of 90. Alternatively, some have said a 35% cloze accuracy rate is satisfactory for a text to be read with classroom support, 45% for homework, and 55% for extensive reading for pleasure. These suggested figures are open to question, however, even for native readers.

It might be thought that EFL readers have a lower tolerance for texts that are challenging than native readers do, but I suspect that the opposite is true. Applying a cloze accuracy criterion of 45% to texts used regularly in my own institution's content-based EFL classes would find many of them out of reach for most of the students who nonetheless do successfully read them. If this is so, is it because EFL students are prepared to work harder to comprehend a text than are native English students? Do EFL students have a higher tolerance for imperfect decoding and use other strategies to comprehend the text? Do EFL readers comprehend texts in a general way better than they are able to produce accurate cloze completions of individual items? In general, does the relationship of cloze accuracy to general comprehension tend to be different for EFL readers than for native readers?

That research has yet to be done. The problem of establishing performance criteria for EFL readability is still very much unresolved (Greenfield, 2001, 2003). Finding that the formulas do a good job of discriminating relative difficulty does not by itself address this question. Neither the Miyazaki study nor Brown's has taken this on as a research issue. Interpreting formula results to help make a determination about what reading materials are appropriate for a particular context still calls on the expertise of the successful teacher. We can now be more confident, however, that the information delivered by the formulas is in fact reliable and relevant to making such a judgment.

Jerry Greenfield is Professor of English at Miyazaki International College, where he also teaches courses in Information Technology and Aesthetics. In addition to his research interests in second language readability and reading assessment, he has presented and published on issues in Web page design and legibility.

Notes

1. This is the present author's calculation for comparison with the Miyazaki results; Brown reports the unadjusted R Square of .55. The adjusted coefficient takes account of the number of variables and provides a more precise estimate of validity.
2. Bormuth gives these scores in a probit metric to remove certain floor and ceiling effects in his data. This was accomplished by looking up percentage scores in a table showing the area under a normal distribution curve to find the corresponding deviation score. He then scaled the scores in such a way that cloze scores of .10, .30, and .50 took on probit values respectively of 372, 448, and 500 (Bormuth, 1971, p. 88). Without Bormuth's original raw scores, it is impossible to compare accuracies directly. However, this does not prevent using Bormuth's scaled criterion in testing correlations with other score sets for the passages.
3. Howell (1997). This test finds whether a difference between two related correlations is significant, or in this case whether a formula is equally good, better, or worse as a predictor of cloze performance in comparison with another formula. The formula for this test is as follows:

$$t = (r_{1.2} + r_{1.3}) \sqrt{\frac{(N-1) \bullet (1 + r_{2.3})}{2\left(\frac{N-1}{N-3}\right) \bullet |R| \bullet \left(\frac{(r_{1.3} + r_{2.3})^2}{4}\right) \bullet (1 - r_{2.3})^3}}$$

where $R = (1 - r^2_{1.2} - r^2_{1.3} - r^2_{2.3}) + (2r_{1.2} \bullet r_{1.3} \bullet r_{2.3})$

In this equation $r_{1,2}$ is the correlation between observed cloze mean scores and one formula's scores, $r_{1,3}$ is the correlation between observed cloze mean scores and another formula's scores, $r_{2,3}$ is the correlation between the two formulas' scores, and N is the number of passages. This ratio is distributed as t on N–3 degrees of freedom.

References

Bernhardt, E. B. (1991). *Reading development in a second language: Theoretical. empirical. and classroom perspectives*. Norwood, NJ: Ablex.

Bormuth, J. R. (1969). *Development of readability analyses* (Final Report, Project No. 7-0052, Contract No. 1, OEC-3-7-070052-0326). Washington, DC: U.S. Office of Education.

Bormuth, J. R. (1971). *Development of standards of readability: Toward a rational criterion of passage performance*. U. S. Department of Health, Education, & Welfare (ERIC Doc. No. ED O54 233).

Brown, J. D. (1992). What text characteristics predict human performance on cloze test items? In *Proceedings of the 3rd conference on second language research in Japan, Tokyo. November 16, 1991* (pp. 1-26). Niigata: Language Programs of the International University of Japan.

Brown, J. D. (1993). What are the characteristics of natural cloze tests? *Language Testing*, 10, 93-116.

Brown, J. D. (1998). An EFL readability index. *JALT Journal*, 29, 7-36.

Bruce, B., & Rubin, A. (1988). Readability formulas: Matching tool and task. In A. Davison & G. M. Green (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidered* (pp. 5-22). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bruce, B., Rubin, A., & Starr, K. (1981). *Why readability formulas fail*. Reading Education Report No. 28. Urbana, IL: University of Illinois Center for the Study of Reading. (ERIC Doc. No. ED 205 915)

Carrell, P. L., & Wise, T. E. (1998). The relationship between prior knowledge and topic interest in second language reading. *Studies in Second Language Acquisition*, 20, 285-305.

Chall, J. (1958). *Readability: An appraisal of research and application*. Columbus, OH: Ohio State University.

Chall, J. (1988). The beginning years. In B. L. Zakaluk & S. J. Samuels (Eds.), *Readability: Its past, present, and future* (pp. 2-13). Newark, DE: International Reading Association.

Chall, J., & Dale, E. (1948). A formula for predicting readability. *Educational Research Bulletin*, 27, 11-20.

Chall, J., & Dale, E. (1995). *Readability revisited: The New Dale-Chall Readability Formula*. Cambridge, MA: Brookline Books.

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60, 283-284.

Dale, E., & Chall, J. (1948). A formula for predicting readability. *Educational Research Bulletin*, 27, 11-20, 28, & 37-54.

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

Fry, E. (1989). Reading formulas: Maligned but valid. *Journal of Reading*, 32, 292-297.

Grabe, W. (1993). Current developments in second language reading research. In S. Silberstein (Ed.), *State of the art TESOL essays: Celebrating 25 years of the discipline* (pp. 205-236). Alexandria, VA: Teachers of English to Speakers of Other Languages.

Gray, W. S. and Leary, B. (1935). *What makes a book readable?* Chicago: University of Chicago Press.

Greenfield, G. (1999). *Classic readability formulas in an EFL context: Are they valid for Japanese speakers?* Ed.D. dissertation, Temple University. University Microfilms No. 99-38670.

Greenfield, G. (2001). Performance criteria for EFL/ESL reading difficulty. Paper presented at the TESOL 2001 Conference, St. Louis, Missouri, March 2, 2001 (Event #5271, audiotaped).

Greenfield, G. (2003). The Miyazaki EFL Readability Index. *Comparative Culture*, 9, 41-49. Miyazaki, Japan: Miyazaki International College.

Hamsik, M. J. (1984). *Reading. readability. and the ESL reader*. Unpublished doctoral dissertation, University of South Florida.

Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Belmont, CA: Duxbury Press.

Kincaid, J. P., Fishburne, R. P., Jr., 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*. Research Branch Report 8-75. Millington, TN: Naval Technical Training, U. S. Naval Air Station, Memphis.

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

Klare, G.R. (1974-75). Assessing readability. *Reading Research Quarterly*, 10, 62-102.

Klare, G. R. (1988). The formative years. In B. L. Zakaluk & S. J. Samuels (Eds.), *Readability: Its past, present, and future* (pp. 14-34). Newark, DE: International Reading Association.

Paran, A. (1996). Reading in EFL: Facts and fictions. *ELT Journal*, 50, 25-42.

Parry, K. (1996). Culture, literacy, and L2 reading. *TESOL Quarterly*, 30, 665-692.

Rubin, A. (1985). How useful are readability formulas? In J. Osborn, P. T. Wilson, & R. C. Anderson (Eds.), *Reading education: Foundations for a literate America* (pp. 61-77). Lexington, MA: Lexington Books (D. C. Heath).

Smith, F. (1988). *Understanding reading: A psycholinguistic analysis of reading and learning to read*. Hillsdale, NJ: Lawrence Erlbaum Associates.

(Received June 5, 2003; revised November 30, 2003)