Point to Point

Problems in Automating Brown's EFL Readability Index

Brett Reynolds
Sakuragaoka Girls' Junior and Senior High School
Michael Geffon
Sakuragaoka Girls' Junior and Senior High School

Brown (1998) developed a readability index using the difficulty of 50 cloze passages as the dependent variable. Four independent variables were combined to calculate the index: the number of syllables per sentence, the average frequency of 30 targeted words elsewhere in the passage, the percentage of words with seven or more letters, and the percentage of function words. The index correlated strongly with passage difficulty (*r* =.74). In the process of writing a computer program to automate these calculations, we found a number of inconsistencies within Brown's article and between Brown's calculations and our own. We also found procedural problems that would limit the generalizability of the index. In a series of personal communications with Brown (December 2001 to March 2002) we were unable to reconcile all discrepancies. Based on our calculations, the best correlation with the difficulty of the cloze passages was found to be *r* =.64, suggesting that the value of an EFL readability index remains to be convincingly demonstrated.

Brown（1998）は５０個のクローズテスト難易度を従属変数にし、それに基づいて「外国語としてのリーダビリティー推定値」を考案した。推定値の関数には次の４つの独立変数を使う（1文あたりの音節数、単語の平均出現頻度、７文字以上の単語の占める割合、機能語の占める割合）。「英文難易度」と推定値との間には、強い相関があることが発表された（r ＝ .74）。しかしながら、我々が推定値を自動的に計るソフトを作成しているときBrownの発表した結果とソフトが計量した結果との間に矛盾が生じた。その上、推定値の一般化には限界があるという問題も見つかった。Brownとの個人的なやり取り（2001年12月〜2002年3月）をした上で、全ての問題を解決することはできなかった。我々の計算ではBrownの推定値と英文難易度との相関は r = .64 という結果が出たので、「外国語としてのリーダビリティー推定値」については未確認といえるだろう。

Brown (1998) reports the development of an EFL readability index. The index was based on data collected from 2,298 Japanese university students. Fifty passages ranging in length from 366 to 478 words (M = 412.1) were selected randomly from a public library. A cloze test was developed for each passage by deleting every 12th word, starting from the second sentence of the passage, until 30 words had been deleted. The remainder of the passage was left intact. Each student was randomly given one of these tests such that, on average, 46 students took each test. The mean score of each test was "normalized by converting them to z values (relative to each other) then to percentiles" (p. 16). The result was considered to be the passage difficulty. This was used as the dependent variable in developing the EFL readability index.

Brown then calculated for each passage "a large number of second language linguistic predictor variables" (p. 18). These included such variables as characters per word, syllables per word, syllables per sentence, and percentage of loan words to Japanese. Four of these were selected based on "correlation, factor, and regression analysis as being orthogonal and most important in predicting EFL difficulty" (p. 18). These are:

1. Syll/Sent
   The average number of syllables found in the sentences in each passage.

2. Pass Freq
   The average frequency with which the correct answers in the 30 blanks appeared elsewhere in the passage.

3. % Long Words
   The percentage[1] of words that contained seven or more letters in the passages.

4. % Func Words
   The percentage[2] of function words among the 30 deleted words in each passage. The remaining words were content words. Function words included articles, prepositions, conjunctions, and auxiliaries. Content words included nouns, pronouns, verbs, adjectives, and adverbs. (Brown 1998, p. 19)

Multiple regression was then used to produce a formula for the EFL readability index using these four variables. The resulting index was reported to correlate strongly with passage difficulty (r =.74). This was far better than any of the existing first language indices that Brown looked at, the closest second being Gunning-Fog (r =.55). Note that Brown is cautious about generalizing the results, citing the low reliability[3] for some cloze tests and the fact that only Japanese university students participated. He suggests that replications may result in different values.

Upon first reading, we thought that this index might be useful in helping teachers and students select level-appropriate passages. However, as Brown points out, "The counts that are necessary and the computations are not only laborious, but are also very prone to calculation errors if done by hand" (p. 28). Consequently, we undertook to write computer software to do this for us. Specific calculations performed by the software, which we call the English as a Foreign Language Readability Indexer (EFLRI), are described below. A general description of the software appears as Appendix A.

When we first ran EFLRI using Brown's variables and coefficients on a non related text of approximately the same length and difficulty as those used in the original study, we initially got a large negative value although the index is designed to produce values between 0 and 100 for such texts. This prompted us to look for errors in our calculations and in the software. However, after all our debugging efforts the output still did not match our expectations.

At this point we asked Brown to send us the original passages so that we could test the program on them. Upon receiving these, we analyzed passage 43 using EFLRI[4]. We chose this passage because it is used in a step-by-step example in Brown (1998, p. 28). We found that none of the variables produced by EFLRI matched Brown's, and that three of the four variables were considerably different.

Table 1. Values for Passage 43 as Reported in Brown (1998) and Calculated by EFLRI.

	Brown (1998)	EFLRI
Syllables/Sentence	76.63	43.77
Passage Frequency	0.41	0.29
% Long Words	19.22	18.59
%Function Words	23.33	33.33

Syllables per Sentence

Syllables per sentence refers to the average number of syllables found in the sentences in each passage. As noted in Table 1, Brown reports *syll/sent* =76.63 for passage 43 while EFLRI calculated 43.77. EFLRI calculates the number of syllables in the passage using a simple algorithm: Each consecutive sequence of vowels found in a word denotes a syllable, excluding terminal *e*. That total is then divided by the number of sentences in the passage to yield *syll/sent*. Our hand counts, done independently by two people, both resulted in *syll/sent* = 43.54 (566 syllables[5] and 13 sentences). This small variation between the computer and hand counts was expected as the heuristic used for counting syllables in the software is imperfect. However, the large difference between our counts and Brown's was unexpected. One explanation we considered was that the passages had become mislabeled but discussions with Brown ruled this out.

The descriptive statistics for *syll/sent* (see Brown, 1998, Table 5) show that the discrepancy is not limited to passage 43. Taken as a group, EFLRI's counts were substantially lower than those reported by Brown (see Table 2). Furthermore, EFLRI's maximum *syll/sent* was lower than the 76.63 reported by Brown for passage 43. Because Brown's original counts for each passage are unavailable, we have no way to reconcile this difference. Consequently, we decided to proceed with the values as calculated by EFLRI. These correlated moderately with the dependent variable (*r* =.50). This is lower than (but similar to) the correlation reported in Brown's (1998) Table 7 (MR =.55)[6].

Table 2. Descriptive Statistics for *Syllables per Sentence* as Reported in Brown (1998) and Calculated by EFLRI.

	Brown (1998)	EFLRI
Mean	36.95	29.45
Maximum	76.63	60.77
Minimum	15.57	10.45
SD	12.62	10.77

Passage Frequency

Passage frequency refers to the average frequency with which the correct answers in the 30 blanks appeared elsewhere in the passage. In calculating this variable, we noticed that total passage length would affect *Pass Freq*, an issue that Brown does not discuss. The fifty passages in Brown's study ranged in length from 366 to 478 words with an average of 412.1 words per passage. Thus, in developing the index, passage length was relatively constant. However, if this calculation were applied to a text of 5000 words, the results would be quite different[7]. To overcome this, we used a scrolling window approach. Only a window of 412 word tokens around each cloze word was searched in turn for other occurrences of that word. Where possible, the window was constructed so that the target cloze word lies at the middle, but was adjusted ahead or behind into the text as necessary to accommodate target cloze word positions closer to the start or end of the passage. Word tokens were normalized to lowercase, with hyphens removed, prior to a character-wise comparison. Other stemming or lemmatizing of the word tokens was not performed.

Another issue we discovered was inconsistent intervals between deleted words in Brown's cloze tests; in examining a number of the tests we found some intervals of 11 or 13 words although Brown states that exactly every 12th word was deleted. Although this will result in slight variations between EFLRI's values and Brown's, it seems unlikely to have a large effect on the predictive power of the index, particularly when it is applied to new passages. This difference also affects *% Func Words*.

Lastly, we found that in his Table 5, Brown (1998) gives the *Pass Freq* minimum as 5.66 (see Table 3). However, on page 28 the value for passage 43 is given as *Pass Freq* =.41. In Note 1, Brown (1998, p.31) reports that *Pass Freq* was transformed in all analyses using a standard log transformation. However, the $\log_{10}$ of Brown's reported minimum is .75. Thus, it is unclear how this *Pass Freq* value was arrived at. Personal communications with Brown failed to resolve this issue.

We calculated the raw *Pass Freq* both by hand and using EFLRI. These two values were identical. Next, we calculated *log Pass Freq*. For each of the 30 cloze words, the $\log_{10}$ of the raw frequency with all words included was calculated. The mean of these transformed frequency counts is reported as the *log Pass Freq* of the passage (see Table 3). Because this resulted in a stronger correlation between the variable and Difficulty (*r*

=-.19 vs. *r* =.02 for *Pass Freq*), we decided to use *log Pass Freq* in EFLRI. Note that the negative correlation is to be expected as the less frequent the words are, the more difficult the passage would probably be.

Table 3. Descriptive Statistics for *Pass Freq* as Reported in Brown (1998) and *Log Pass Freq* as Calculated by EFLRI.

	Brown (1998)	EFLRI
Mean	6.96	0.47
Maximum	8.82	0.65
Minimum	5.66	0.25
SD	0.59	0.09

Percentage of Long Words

Percentage of long words refers to the percentage of words that contained seven or more letters in the passages. As reported in Table 1, Brown's value for passage 43 was 19.22 while EFLRI outputs 18.59: the same value we arrived at through hand counts. These values are quite similar, with the difference likely due to minor counting errors. Word tokens were considered "long" if they contained 7 or more characters, excluding hyphens. Overall descriptive statistics for this variable were also quite similar between Brown's findings and EFLRI's output (See Table 4). We found that these values correlated moderately with the dependent variable (*r* =.48).

Table 4. Descriptive Statistics for *% Long Words* as Reported in Brown (1998) and Calculated by EFLRI.

	Brown (1998)	EFLRI
Mean	20.52	20.19
Maximum	34.33	34.27
Minimum	9.89	9.11
SD	5.94	6.11

Percentage of Function Words

Percentage of Function Words refers to the percentage of function words among the 30 deleted words in each passage. In calculating this variable, Brown used a fixed list of function words composed of articles, prepositions, conjunctions, and auxiliaries, as quoted above. Use of such a list can be problematic if it does not differentiate between parts of speech (POS), for example, *be* as auxiliary vs. *be* as main verb. Brown is unclear on whether such differentiation was performed, and we were unable to obtain the list of words that he used. Consequently, we could not determine precisely how his *% Func Words* was calculated. In the end, we decided to implement this variable using an existing software program, POS tagger[8], allowing us to categorize function words based on their actual usage rather than merely on identical spelling. The POS tagger uses the Penn Treebank tag set, of which we counted the following as being function words: CC (coordinating conjunction), DT (determiner), EX (existential *there*), IN (preposition or subordinating conjunction), MD (modal auxiliary), RP (particle), and TO (*to* as a preposition or infinitive marker). While our hand counts using this set were identical to EFLRI's output, this set does not overlap perfectly with the definition provided by Brown. For example, determiners include articles but also include words that Brown may have counted as content words (e.g., *many*).

Having used different sets of words we were surprised that the overall results were remarkably similar, especially given the differing values for passage 43 (see Table 1). We found that EFLRI's calculated *% Func Words* correlated weakly with the dependent variable (r =.24), although we had expected a negative correlation: the fewer function words there are, the more difficult the passage would likely be.

Table 5. Descriptive Statistics for *% Func Words* as Reported in Brown (1998) and Calculated by EFLRI.

	Brown (1998)	EFLRI
Mean	31.55	31.40
Maximum	50.00	50.00
Minimum	13.33	13.79
SD	8.17	8.85

The Index

Using multiple regression, Brown arrived at the following coefficients for the index.

$$
\begin{aligned}
\text{EFL Difficulty Estimate} = 38.7469 \quad & + (.7823 \text{ x Syll/Sent}) \\
& + (-126.1770 \text{ x Pass Freq}) \\
& + (1.2878 \text{ x \% Long Words}) \\
& + (.7596 \text{ x \% Func Words})
\end{aligned}
$$

Although we did our best to replicate the original study, we were unable to obtain similar results for three of the four variables. Consequently, the index as calculated using EFLRI-calculated values did not predict the passage difficulty as well as it did using Brown's values. While the original study reports a correlation of .74 between the calculated index scores and the passage difficulty, we found a lower correlation (*r* =.64)[9].Surprisingly, by recalculating[10] the regression coefficients based on our new values for the four variables, we were unable to increase this correlation. However we found that a three-variable solution, dropping syllables per sentence, performed equally well (see Tables 6, 7, & 8).

Table 6. Multiple Regression Summary Statistics

Statistic	
Multiple R	0.64
R^2	0.40
Adjusted R^2	0.37
Standard Error	22.40

Table 7. ANOVA Results

	Df	SS	MS	F
Regression	3	15,708.22	5,236.07	10.44*
Residual	46	23,081.54	501.77	
Total	49	38,789.75		

* $p < 0.0001$

Table 8. *t*-test Results

	Coefficient	SE	t	p
Intercept	45.28	17.94	2.52	0.02
Log Pass Freq	-151.69	41.86	-3.62	0.00
% Long Words	2.01	0.55	3.63	0.00
% Func Words	1.24	0.46	2.68	0.01

The formula for this solution is:

New EFL Difficulty Estimate = 45.28 + (2.01 x % Long Words)
+ (1.24 x % Func Words)
+ (-151.69 x Pass Freq)

The resulting correlation, while lower than the .74 reported by Brown, is still higher than any of the L1 readability indices (Brown, Table 4), but only slightly (lowest is Fry =.48; highest is Gunning-Fog =.55). However, it seems likely that some of this advantage for the new index results from doing the comparison using the very passages that it was based on. Whether this advantage transfers to other passages remains to be seen.

Summary

We set out only to automate the existing formula. In the course of reviewing the original paper and writing the EFLRI software, a number of issues surfaced, some of which we were able to overcome but many of which were irreconcilable. As we were unable to determine how the original data were arrived at in many cases, we were left with no choice but to recalculate them using our own procedures. Instead of the correlation of .74 that Brown reported, we were only able to obtain values that correlate at .64. Although we will not assert that Brown's data are in error, we do believe that they cannot be obtained using only the methods described in the published report. While we were not able to improve on the predictive strength of Brown's published formula, we did arrive at a more parsimonious solution.

Although the correlation that we obtained is below that originally published, it is still better than existing L1 indices. It seems unlikely that this advantage would translate to other passages; however, this remains an open question. Another question that is still unanswered is whether the index will work equally well for easier or more difficult passages than for those used in the study. It is also unclear how the length of the passage will affect the outcome, although we have attempted to control for this.

In sum, both the formula published in Brown (1998) and the simpler solution that we found seem likely to be at least as good as L1 readability indices for predicting the difficulty of the passages for Japanese university students. However, the value of an EFL readability index, as distinct from L1 readability indices, remains to be convincingly demonstrated.

Acknowledgements

We would like to acknowledge the cooperation of J. D. Brown in sending us his original passages and taking the time to answer our questions. We also thank Hamish Cunningham and the entire GATE community in the NLP Group at the University of Sheffield for all their good cheer and hard work in making the GATE software available.

Brett Reynolds was with Sakuragaoka Girls' High School during this project. He is currently with the Humber institute of Technology & Advanced Learning.

Michael Geffon has administered the CALL program at Sakuragaoka Girls' High School since 1998. He is currently a student in the Educational Technology Leadership Program at George Washington University.

Notes

1. The variable is percentage of long words. Thus, if there were 50 long words and the passage contained 400 tokens (including the 50 long words) the value for this variable would be 12.5 (not 0.125).

2. As with Note 1, the variable is percentage of long words.

3. It seems to us that test reliability is not the main issue here as it is the relative distribution of the tests themselves that is of interest not the relative distribution of students. Put another way, were the students reliable in testing the difficulty of the passages?

4. For all calculations we removed passage and section titles because these were not used in counting every 12th word in Brown's cloze tests.

5. We initially disagreed about two words, but resolved this by referring to a dictionary.

6. It appears that a printing error resulted in MR^2 being rendered as MR± throughout Brown (1998).

7. A longer passage would provide more opportunities for words to reappear, thus increasing the value of this variable. Note that the variable is simply a count, not a ratio.

8. The POS tagger is a modified version of the Brill tagger, the accuracy of which has been found to be around 95%, depending on the text and conditions (Hepple, 2000).

9. In preliminary attempts using other lexical variables (e.g., word frequency band and type / token ratio) we have obtained MR values higher than Brown's 0.74.

10. Statistical calculations were done using R, an open-source statistical programming language, and duplicated using Microsoft Excel, both running on Macintosh computers under OS X.

References

Brown, J. D. (1998). An EFL readability index. *JALT Journal, 20* (2), 7-36.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002, July). GATE: A framework and graphical development environment for robust NLP tools and applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics* [ACL'02]. Retrieved February 21, 2003, from http://www.aclweb.org/anthology/P02-1022

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Ursu, C., & Dimitrov, M. (2002). *Developing language processing components with GATE (a user guide) For GATE version 2.1 beta 1 (August 2002)*. Retrieved February 21, 2003, from http://gate.ac.uk/sale/tao/tao.pdf

Hepple, M. (2000, October). Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* [ACL-2000]. Retrieved February 21, 2003, from http://www.aclweb.org/anthology/P00-1036

Microsoft. (2001). Excel X for Mac (Service Release 1) [computer software]. Redmond, WA, U.S.

R (Version 1.6.2) [computer language]. Retrieved December 10, 2002, from ftp://ftp.u-aizu.ac.jp/pub/lang/R/CRAN

(Received March 8, 2003; Revised June 19, 2003)

Appendix A

Notes on The EFLRI Software

Some of the problems we had reproducing Brown's data likely stem from the software tools he employed: As text processing is a messy business, it is likely that the heuristics, like our own for counting syllables, were not exact. Brown lists three programs, but does not indicate which were used for which calculations. None seem to still be available.

In undertaking to write EFLRI, we quickly realized that the calculations were more complex than they first appeared: Accurately determining sentence boundaries in heavily abbreviated or quoted text, for example, is a non trivial task. A second issue was our desire to construct ELFRI so as to standardize our operationalization of the variables to simplify replication. Happily, an existing Natural Language Processing (NLP) framework we found to perform most of the mundane functions of parsing and tokenizing the passages, GATE, was also designed to facilitate just such "quantitative measurement of accuracy and repeatability of results for verification purposes" (Cunningham, Maynard, Bontcheva, Tablan, Ursu, & Dimitrov, 2002, p. 3), thereby resolving both of these issues.

The GATE architecture, written in Java, allows for custom component-based modules, known as Processing Resources (PRs), to be written and executed on a given document or corpus of documents. PRs are usually strung together in a "pipeline" to build applications (Cunningham, Maynard, Bontcheva, & Tablan, 2002). The EFLRI program itself is a four-stage pipeline: The first three processing steps are done with standard GATE PRs, which, respectively, tokenize the text, determine sentence boundaries, and perform POS tagging. We wrote the fourth PR to then calculate the independent variables, as described above.

Word tokens, as described throughout this paper, are a formal GATE construct, defined as "any set of contiguous upper or lowercase letters, including a hyphen (but no other forms of punctuation)." (Cunningham, Maynard, Bontcheva, & Tablan, 2002) However, our program for comparison of word token equality ignores case and hyphenation. Other distinct tokens in the source text not included in any of our counts are number, symbol, punctuation, and space tokens.

The Author Responds to Reynolds and Geffon

James Dean Brown
University of Hawai'i at Manoa

Insofar as I am able to understand their article at all, I think Reynolds and Geffon (hereafter, R and G) are trying to make two points, which are summarized in their abstract (p. 197, this volume):

1. "...we found a number of inconsistencies within the article and between Brown's calculations and our own."
2. "We also found procedural problems that would limit the generalizability of the index."

There are so many logic fallacies, and problems with definitions, quantification methods, explanations, writing, etc. in the piece by R and G that I was tempted to trust the readers of *JALT Journal* to see the many flaws. However, in the end, I could not restrain myself; I could not let such nattering go unanswered. I will focus my responses on their two basic "arguments" (inconsistencies and lack of generalizability).

Inconsistencies

To repeat for the sake of clarity, R and G maintain that "we found a number of inconsistencies within the article and between Brown's calculations and our own" (p. 197, R&G abstract, this volume).

Inconsistencies Within My Article

The primary inconsistencies they point to within my article are disparities in (a) the deletion patterns in my cloze tests, (b) the values reported for the *Pass Freq* variable, and (c) percentages and proportions reported in different places in my article.

Inconsistencies in deletion patterns

In their words: "...in examining a number of the tests we found some intervals of 11 or 13 words although Brown states that exactly every 12th word was deleted." (p. 201, R&G, this volume). First, I did not state that, "*exactly* every 12th word was deleted" (emphasis mine). However, I have gone back through the passages, and was surprised to indeed find some cases of 11th, 13th and even 14th word deletion. I would fire the graduate assistant who made a few mistakes in counting out *1500* items, but he finished his doctorate many years ago at Florida State University. I am disappointed to find that he did not do *exactly* what I asked him to do, but in any case, I assume these few minor discrepancies are errors in counting which average about every 12th word deletion. I further assume these counting errors are randomly distributed and therefore should have exactly zero effect on the results of the readability index.

Inconsistencies in the Pass Freq variable

The inconsistencies in values reported for the Pass Freq variable are clearly due to my reporting the raw frequencies in one place and the log transformations of those frequencies in another place, as I clearly explained in my article. When they tried one of many possible log transformations, they did not get the same results I had. They therefore felt obliged to use a transformation of their choosing in their computer program.

Inconsistencies in percentages and proportions

The inconsistencies in percentages and proportions, which they apparently found very disturbing, seem to me to be the silliest of the bunch: Anyone can easily understand the astounding difference between .125 as a proportion and 12.5%. Most people can make the shift in decimal points in their heads without even thinking about it. I often shift back and forth in my lectures in my testing classes, a process that Reynolds seemed to follow quite readily when he was a student in that class in the M.Ed. program at Temple University Japan.

Inconsistencies Between My Original Article and Their Attempts

The inability of R and G to recreate my results may be a consequence of one or more of the following:

1. Differences could be due to errors in my calculations. However, most of my counts were automated using well-established professional software and the algorithms written into those programs (*Scandinavian PC Systems*, 1988; *Que Software*, 1990; *PC-Style* by Button, 1986). In other words the counts done with the same software would be 100% the same, yet might not be the same as counts derived from the their homegrown EFLRI computer program.

2. Differences might have occurred because R and G were looking at a different "passage 43". What was labeled passage 43 in the text files may be different from the passage 43 in the analyses. That is my guess for the differences found in Table 1.

3. Differences in the multiple regression analyses could be due to variations in the definitions they were using for each of the variables as compared to the ones I used. Indeed, since their definitions are not very clear, this could be a major source of differences. For example, their "simple algorithm" for syllables per sentence is as follows: "Each consecutive sequence of vowels found in a word denotes a syllable, excluding terminal *e*" (p. 200, R&G, this volume). What is a "consecutive sequence of vowels"? And, how can such a definition account for closed syllables, syllabic consonants like *l* and *r*, and other aspects of the rather complex concept of *syllable* in English? I know for a fact that Reynolds took a course at Temple University Japan called "Sound Systems" and should therefore know much more about English syllables than their definition would indicate.

4. Differences could be due to variations in the algorithms used to calculate each of the variables. That is my guess for the differences found in Table 2.

5. The differences reported in Table 3 are clear. They report the average passage frequency for the Brown data and the log of passage frequency for their own calculations. It is no surprise that they are radically different. It appears they are also using a different log transformation from the one I was using.
6. Differences could also be due to accumulated rounding errors. That is my guess for the differences found in Tables 4 and 5.

Errors do happen. Some may have occurred in my research and some in their piece. It is important to recognize that errors can accumulate from two sources. In the case of R and G's piece, the differences could have resulted from errors in my calculations or errors in theirs. They seem to be implying that the errors are all in my calculations because they "have no way to reconcile" these differences. Since my calculations are based on proven software programs, I will stand by them. The bottom line is that, just because they were not able to replicate my results with their homegrown EFLRI computer program, does not mean my results were not accurate, nor does it diminish the value of my research.

Generalizability[1]

To quote again from their abstract for the sake of clarity, R and G maintain that, "We also found procedural problems that would limit the generalizability of the index" (p. 197, R&G abstract, this volume). Did R and G read my article all the way through? It appears that they got themselves so focused on trying, and failing, to get their EFLRI computer program to work that they failed to understand what my article was actually about. Helping them to get their computer program to work was not the purpose of my research. *My* point was as follows: "The primary point is not that this particular index is the magical answer to determining the readability of passages for use in ESL/EFL curricula and materials, but rather that such an index can be created, one that is more highly related to the performance of second language learners than are the first language readability indices" (Brown, 1998, p. 30).

Ironically, R and G argue: "Based on our calculations, the best correlation [sic: should read "multiple correlation"] with the difficulty of the cloze passages was found to be r = .64, suggesting that the value of an EFL readability index remains to be convincingly demonstrated." Their different result is not surprising given that they defined the variables in their own vague ways and given that they used fewer independent variables in the multiple regression than I did. More importantly, they made no effort to explain how they applied multiple regression. For example, how were the variables entered into the equation? I used a *stepwise* approach (as I explained in the title for Table 6 on p. 25), which is also known as the *forward-stepping* method (as I explained in the text at the bottom of p. 24). What did they use? Did they use the entry, forward-stepping, backward-stepping, forward and backward stepping, hierarchical, or one of the other available methods? My guess is they don't know. Yet the method of entering the independent variables crucially affects the results and should always be explained. Is it surprising that they got different results when they used fewer independent variables, defined those variables differently, and maybe used a different form of regression? No, not at all.

What is surprising is the degree to which their results support my contentions without them even understanding it: Even with their apparently inaccurate computer program, they managed to create a readability formula that has a multiple correlation of .64 with the EFL difficulty of the passages, which is far better than the correlations of any of the first language readability indices with EFL difficulty. Thus their results support my main contention that "an index can be created, one that is more highly related to the performance of second language learners than are the first language readability indices"

(Brown, 1998, p. 30), and their findings corroborate the *transferability* (see footnote 1 above) of that contention. It seems, like Hamlet's statement about Rosencranz and Guildenstern, that we have another R and G, who are *hoist upon their own petard?*

Conclusion

In the end, I'm not sure what their motivation was for writing their piece. It may be something as simple as petulance over the fact that I was too busy to get back in touch with them after our last meeting. I did meet with R and G, at which time I agreed to supply them with the further information they clearly needed. As I often do, I waited for an e-mail request from them to trigger that reaction in my busy schedule. I never received such an e-mail request and figured they had probably lost interest. In any case, along the way, I had read Greenfield's recent paper (discussed below) and had therefore concluded that their project was at best a quixotic waste of time. The next thing I knew, R and G e-mailed me what must have been a draft version of their piece (given its incoherence), and then I heard that a revised version was coming out in the *JALT Journal*.

Whatever their motivation, if I had problems replicating another person's research (as suggested by their title: "problems in automating Brown's EFL readability index"), I would tend to first figure out where *I* went wrong. I would be embarrassed to expose my own inabilities and the inadequacies of my own computer program as they have apparently done. Personally, I prefer to write articles that make positive contributions to the field, rather than joining the ranks of those who pick endlessly at the articles of other people who have made such positive contributions.

Recently, Dr. Jerry Greenfield *has* made a positive contribution to research on EFL readability. He did his doctoral dissertation (Greenfield, 1999) and other recent research comparing my approach, his Miyazaki approach, and the traditional readability estimates (Greenfield, unpublished ms.). If R and G had taken a bit more scholarly approach to their piece, they might have turned up Greenfield's interesting work. Greenfield was not only able to use my formula, but was also able to develop an extension of it. It seems odd that Greenfield was able to use my formula, but R and G were not. How can this apparent difference be reconciled? Could it be that Greenfield is a trained and experienced researcher?

Greenfield's findings indicate that my formula works better with his data than it did with my own and that it works well. However, more interestingly, in his study, the first language readability estimates correlated much better with student performances than they did in my study, a fact that he (probably correctly) attributes to differences in the passages selected for study (mine were randomly selected from the native speaker reading materials in an American public library, whereas his were based solely on educational materials). I think there were probably also critical differences in the student populations used in the two studies (mine were 2,298 students from 18 different universities, while his were 200 students from Miyazaki International College, at which content courses are taught in English) that could have made big differences in the distributions involved in the analyses, which in turn could have affected the magnitude of the various correlation coefficients. Greenfield concludes, quite rightly I think, that, because the first language readability estimates in his study are easier to use and yet correlate with student performance almost as well as my readability formula and his, the first language readability estimates will generally be more useful. (Note that, in Brown, Chen, and Wang (1984), I was part of a team that advocated the use of first language readability estimates almost two decades ago.)

Thus, while Greenfield's findings take issue with my results, indeed contradict them in some ways, I don't find his paper the least bit objectionable because it is positive, logical, responsible, and scholarly. That is how we make progress in research. I interpreted my results as I saw them, Greenfield interpreted his results the way he saw them, and gradually we all come to a consensus as a field. I can tell you for sure that I will be citing Greenfield's article for years to come because he made a valuable contribution to the field. I look forward to seeing his new article published soon.

Interestingly, if R and G had taken a more scholarly approach and read Brown, Chen, and Wang (1984) and Greenfield's work (instead of relying on a single article by a single author as they did), even they would surely have recognized that using one of the much simpler first language readability approaches would make more sense for their purposes. That realization would, in turn, have afforded them the opportunity to use approaches they *are* capable of applying—approaches already available in the *Microsoft Word*^{tm} software program (including both the Flesch Reading Ease index and the Flesch-Kincaid Grade Level index, both explained in Brown, 1998, p. 17). They might thereby have recognized the apparent irrelevance of their EFLRI computer program and could have saved themselves the considerable embarrassment of publishing their confusion to the world.

Note

1. After 25 years of doing quantitative research in ESL/EFL, I have come to the conclusion that "generalizability" in our research is a pipe dream. We are never able to randomly sample from a population, and therefore are never able to generalize from a sample to that population. Even in EFL studies, our "samples" are at best generalizable only to the EFL students in a particular school. In ESL studies, which typically involve students from many different nationalities and language backgrounds, there is no definable population to which results could conceivably be generalized. At best, we can strive for the concept of *transferability*, which can be defined as the degree to which the

results of one study can be transferred, or can be applied, to other settings (for more on the concept of *transferability*, see Brown, 2001, or Brown & Rodgers, 2002). Given these long-held beliefs, I would be the last person to claim generalizability for any of my studies.

James Dean Brown was educated at California State University Los Angeles (BA French), University of California Santa Barbara (BA English Literature), and University of California Los Angeles (MA TESL and PhD in Applied Linguistics). Currently Professor of Second Language Studies on the graduate faculty of the Department of Second Language Studies at the University of Hawaii at Manoa, his areas of specialization include language testing, curriculum design, program evaluation, and research methods.

References

Brown, J. D. (1998). An EFL readability index. *JALT Journal, 20*(2), 7-36.

Brown, J. D. (2001). *Using surveys in language programs.* Cambridge: Cambridge University Press.

Brown, J. D., & Rodgers, T. (2002). *Doing second language research.* Oxford: Oxford University Press.

Brown, J. D., Chen Yongpei , & Wang Yinglong. (1984). An evaluation of native-speaker self-access reading materials in an EFL setting. *RELC Journal, 15*(2), 75-84.

Button, J. (1986). *PC-Style: The program that has a way with words.* Grand Prairie, TX: Lone Star Software.

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

Greenfield, G. (unpublished ms.). Readability formulas for EFL. Unpublished ms. Miyazaki, Japan: Miyazaki International College.

Que Software. (1990). *Right writer: Intelligent grammar checker*. Carmel, IN.

Scandinavian PC Systems. (1988). *Readability program*. Rockville, MD.

(date received July 10, 2003; date revised July 11, 2003)