In Search of New Dimensions for Readability for Japanese Learners of English

Ken NORIZUKI

(平成15年11月 4 日受理)

Readability statistics have been used for many decades in Japan for teaching and research purposes. These statistics have generally been found reliable and valid for first-language learners of English, but only a limited number of studies have been conducted to test the reliability and validity of the measures for Japanese learners of English, and even fewer studies have been directed toward developing a new readability formula. The aim of the present study is twofold. First, conventional readability statistics and their parameters are compared in terms of the extent to which they enable predictions to be made about reading test performance. Second, combinations of various linguistic variables are tested to seek the possibility of developing a new measure. Six readability statistics, their constituent parameters, and four new variables were correlated with average scores on six reading passages administered to 119 university students. Multiple regression analyses were performed to ascertain some of the strongest combinations of predictors from among the linguistic variables examined. Three new readability equations are proposed to form the basis of future scrutiny. The overall results and the implications are discussed to pave the way for continued research.

1. Introduction

When a text is selected for teaching/learning and testing purposes, there are two major criteria that teachers or learners take into account. The first criterion is the ease or difficulty of the text. If the text appears to be too difficult (or too easy) for particular learners, it will be deemed unsuitable for use. The second criterion concerns the learners' areas of interest, or the extent to which the learners find the text takes up interesting topics they would like to know more about.

While both criteria are important, the first criterion that is examined with regard to "elements which might affect the ease with which readers can comprehend a text" (Harrison and Bakker, 1997: 122) should be given first priority in most selection cases. An incomprehensible text that carries potentially interesting subjects for students to read is just as bad as a comprehensible, boring text, where no sense of

self-engaged reading and learning takes place. Thus, a text needs to be at the right (linguistic and cognitive) level, and is even better if it arouses learners' general or specific interest. This leads to a discussion of how to assess text comprehensibility, often simply called readability, in a reliable and valid manner.

A large number of readability formulas or analytic procedures have been developed and extensively discussed in the last century. These can be categorized into three main types of measures. One type is entirely based on the lengths of lexical variables, such as counts of the number of words and either syllables or letters. Flesch Reading Ease, Flesch-Kincaid Grade Level, Fry Graph, and Gunning Fog formulas are some examples. Some of these measures can be estimated easily using functions of PC word processors (e.g., two Flesch measures, which can be obtained from the grammar and spelling checker provided with Microsoft Word).

Another type relies on lists of general or familiar words as well as, or more than, on the length factors. The most well-known of these are the Dale and Chall formula (1948) and the new Dale and Chall formula (1995). Before the introduction of the new Dale-Chall formula, Klare (1963) concluded that the Dale-Chall formula (1948) had a slightly higher predictive power than other existing formulas (quoted in Kiyokawa, 1993). Kiyokawa (2000) commends the use of the new Dale-Chall formula, if time allows, particularly for research purposes (see Kiyokawa, 1988; Hasegawa and Norizuki, 1999; Duppenthaler, 2000; Kinoshita and Ohtsu, 2002).

The third type of analysis takes a more global approach, but tends to be one without specific formulas. Zakaluk and Samuels (1988), as described by Kiyokawa (1993), conceptualized three dimensions of comprehensibility: text readability level; adjunct comprehension aids, and learners' reading comprehension level. Naganuma (2000) statistically analyzed the grammatical and functional features of words in the texts by using corpora-based statistical analysis tools. Bormuth's formula (1969), which appears to be a rather complex variety of the second type, was an end product of sophisticated linguistic analysis of word depth, letter redundancy, independent clause count, and parts of speech or form class, in addition to surface features counted in the earlier formulas (see Greenfield, 1999).

The first type of measure has generally been found to be reliable and valid for users of English as a first language, and is preferred in most pragmatic educational and research settings. A large number of EFL studies in Japan have employed formulas of this type to demonstrate a striking gap between the students' reading ability and the difficulty of texts used in textbooks or entrance examinations (e.g., Kimura and Visgatis, 1998; Matsuo, 1998; Brown and Yamashita, 1995; Shiozawa and Komaba, 1990; Shiozwa and Aizawa, 1989). However, it has not been ascertained whether the reliability and validity of the measure is true for

exposure-limited EFL settings.

The second type of measure is expected to be more useful when software programs become more widely available. The findings in Hasegawa and Norizuki's study (1999) suggested that Kiyokawa's formula (1988), a variant of the earlier Dale-Chall formula, which was developed for Japanese high school and university students, was a potentially more accurate measure of text difficulty than the Flesch formulas.

The third type of measure has been of limited practical utility so far. However, Harrison and Bakker (1997: 136) have shown that "correlation between lexical density scores and perceived readability can be much higher than is the case with more conventional readability scores" (based on Flesch Reading Ease, Flesch-Kincaid, and Gunning Fog formulas), and that "distributions of sentence length and 'packet' length taken together make for a better predictive model than those relying heavily on the first variable." A major problem remains as to how to count these variables easily and quickly under practical time constraints.

2. The present study

The aim of this study is twofold. The first is to compare the effectiveness of conventional readability measures and their parameters. The second is to explore the possibility of developing a new readability formula that will be a better predictor of readability for Japanese learners of English than conventional measures. It is hoped that the new formula can be applied quite easily using Word and Excel functions and eventually even more easily, with the aid of a relatively simple computer program developed for this purpose.

First, six conventional readability formulas were selected. Flesch Reading Ease and Flesch Kincaid scores were included as they can be estimated from Word. Coleman-Liau was chosen on the basis of ease of calculation. These three formulas represent the first type of analysis. The two Dale-Chall formulas representing the second type were calculated because they are often regarded as the most accurate model in current use. The Bormuth formula was added to the list of comparisons because its prototype originated in the third type of analysis. This formula is also important because one of the parameters involved is based on a vocabulary list employed by the older Dale-Chall model that can be aligned with the new model's list.

Second, candidates for new parameters to be included in the new measure were examined. Final candidates were Hokkaido University Word List Level 1, Levels 1 and 2, the type/token ratio, or the number of different lexical items in a text (types) divided by the total number of words (tokens), and the index of Guiraud, which is computed by dividing the number of types divided by the square root of the number of tokens (see Daller *et al.*, 2003).

The type/token ratio was included because a passage containing the same long or difficult words repeatedly seems to be less difficult than a similar passage containing many different long or difficult words (see Klare, 1976). A problem with this measure is, as a passage gets longer, the type/token ratio tends to get lower. Guiraud is a statistical adjustment to the type/token ratio. Research shows that this measure has been found to be consistently more stable than its unadjusted version regardless of the lengths of passages (Daller *et al*, 2003).

The inclusion of Level 1 lexicon as one parameter and a combined lexicon of Levels 1 and 2 as another was designed to cover a wider range of learners' reading ability weighted separately in two interdependent, but distinct parameters. The Hokkaido Word List was chosen because it is readily accessible from a website and words included in the list and its grading are intuitively more appealing to Japanese learners of English, as compared with word lists developed in English-speaking countries, such as ungraded Dale-Chall word lists (1949 and 1995) and the graded University Word List (Nation, 1990).

3. Method

3.1 Materials

Six passages were selected from institutional examination papers constructed in past years by the writer of this article in conjunction with colleagues. Each passage was followed by five four-option test items. The passages were used in high-stakes examinations in such a way that text and test difficulty was carefully controlled on a par with examinees' general level of English language proficiency.

The six passages were roughly divided into three levels: ones designed for intermediate, upper-intermediate, and advanced levels set against institutional standards even though these standards varied somewhat depending on the changing levels of examinees involved. Six readability scores and their parametric values were calculated. Based on the derived new Dale-Chall (1995) scores, the six passages were grouped into two tests. Two of these passages were included in both versions of the tests so as to equate the two tests that could potentially be of unequal test difficulty levels. The two tests were constructed to be of parallel overall text difficulty. Each test consisted of two common and two uncommon passages (or four subtests) and 20 test items.

3.2 Subjects

One of the two tests was assigned to four classes of university students. Two classes were from Shizuoka Sangyo University and the other two were from another university in Shizuoka Prefecture. The students were all first-year students, ranging

from intermediate to advanced levels, relative to language proficiency levels of students in the institution that the writer is affiliated with. The two tests were numbered and arranged in order: Version 1 (henceforth, V1) of the test was odd-numbered (e.g., 1, 3, 5, ... 99) and Version 2 (henceforth, V2) was even-numbered (e.g., 2, 4, 6, ... 100). Thus, nearly the same number of students was randomly given either V1 or V2 test in each class. Sixty students took the V1 test and fifty-nine students took the V2 test (n=119).

3.3 Analyses

Six readability scores and their parametric values were obtained using Microsoft Word 2002 and Microsoft Excel 2002. A set of average scores for six passage-subtests was correlated with six readability scores and their parametric values. In this study the average subtest score was considered a dependent variable and all other variables were independent variables because of their potential relationships with the reading text difficulty dependent variable (see Brown, 1989). Pearson product-moment correlation and multiple regression analyses were performed to explore these relationships using SPSS Version 11. The Rasch model PROX procedure was carried out to equate V1 and V2 tests, with the aid of TDAP Version 2 (Ohtomo and Nakamura, 2002).

3.4 Results

In table 1, six measures of conventional readability statistics and their parametric values are reported in the ascending order of Dale-Chall (1995) text 'easiness' scores (the higher the score, the easier the text). According to Dale and Chall's (1995) conversion criteria, Passage 1 is graded at 11-12 reading levels; Passage 2 at 9-10; Passage 3 at 7-8; Passage 4 at 5-6; Passage 5 at 5-6, and Passage 6 is at 5-6. Flesch Reading Ease scores rank the passages in exactly the same order. Bormuth (1969) text 'difficulty' scores put the passages in the same ranks in a descending order (the higher the score, the more difficult the text). Other readability scores and parameters are not ranked in a unilateral difficulty/easiness order. Even the scores for Flesch Reading Ease and Flesch-Kincaid Grade Level, which are based on the same types of variables, are not placed in an identical order. (Henceforth, acronyms listed in table 1 will be used to refer to the conventional statistics.)

Table 2 shows the statistics for extra variables included in this study. The indices of Guiraud are placed in the exact order of DC48 scores. It also appears to match fairly well with the DC95 and BM scores, except for the order of passages 4 and 5. The percentage scores for words not in Hokkaido University Vocabulary Lists show a marked difference from those based on two DC word lists (W95 and W48) in table 1.

Moreover, the ratios of words beyond Level 1 and beyond Level 2 place the passages in a completely different order. This suggests that it may be worthwhile considering the inclusion of two levels as separate variables in developing a new formula. (Henceforth, acronyms listed in table 2 will be used to refer to the four new variables.)

Table 1　Six readability measures and their parameters

	DC95	DC48	BM	FRE	FK	CL	S/W	L/W	W/S	W95	W48
P1	21.61	8.56	79.59	29.6	12.0	14.35	1.76	5.12	27.64	24.55	23.00
P2	29.57	7.72	78.55	42.7	12.0	14.82	1.69	5.20	20.67	21.24	25.54
P3	37.53	5.57	74.45	56.7	7.9	14.81	1.61	5.20	13.65	17.95	18.32
P4	40.11	4.83	73.08	57.1	10.0	11.27	1.54	4.60	19.08	11.29	8.06
P5	41.39	5.10	71.52	63.3	8.7	10.87	1.49	4.53	17.28	11.25	10.93
P6	43.50	4.76	68.35	64.3	7.2	11.49	1.54	4.63	12.05	12.83	12.08

DC95: Dale-Chall 1995; DC48: Dale-Chall 1948; BM: Bormuth 1969; FRE: Flesch Reading Ease; FK: Flesch-Kincaid Grade Level; CL: Coleman-Liau; S/W: the number of syllables divided by the number of words; L/W: the number of letters divided by the number of words; W/S: the number of words divided by the number of sentences; W95: the percentage of words not included in DC95 word list; W48: the percentage of words not included in DC48

*NB1: FRE and FK scores are rounded off to one decimal place in the output of the Microsoft Word readability analysis.

*NB2: Passages were analyzed as a whole rather than on the basis of 100-word samples because passages were generally short and sampling seems to hamper the natural process of analysis when a statistical program is used.

Table 2　Four new variables for readability analysis

	T/T	GR	HL1	HL2
P1	.5116	10.06	28.17	10.59
P2	.5000	9.64	30.91	10.75
P3	.4505	9.35	24.91	11.72
P4	.5524	8.70	14.92	4.03
P5	.5016	8.85	17.04	9.65
P6	.4868	7.92	18.87	9.43

T/T: type-token ratio (the number of different lexical items divided by the total number of words); GR: the index of Guiraud (types divided by the square root of the number of tokens); HL1: the percentage of words outside Hokkaido University Vocabulary List Level 1; HL2: the percentage of words outside Hokkaido University Vocabulary List Levels 1 and 2.

Based on DC95 scores as shown in table 1, the two tests were made to be of approximately equal text difficulty. The V1 test was made up of passages 1, 4, 5 and 6; The V2 test of passages 2, 3, 4 and 6. Each test had twenty items, five items for each passage, and 10 items for passages 4 and 6 served as a linking mechanism to equate the two versions of the tests.

The raw average scores suggest that the two tests or the two groups of the examinees were not at exactly the same level of difficulty or ability though the difference was not found to be statistically significant.

The raw scores were converted into logit scores using the Rasch model PROX procedure. The difference in mean difficulty for linking items between the V1 and V2 tests was used as a translation constant and was simply added to all the V2 item difficulty calibrations. This equating process allows all the thirty items on the two tests to be positioned on a common metric of difficulty. Recalibrated item difficulties were averaged across each reading passage, or subtest, and these means could be compared with readability or text difficulty scores (see Henning, 1987 for the PROX and the PROX-based linking procedures).

Logit item difficulties are expressed on a continuum scale running from minus infinity (low difficulty) to plus infinity (high difficulty). For ease of interpretation, these statistics were further recalibrated on the WIT scale, which is written as:

WIT = 9.1*logit item difficulties + 100

On the WIT scale, a passage that matches the group ability is positioned around 100 points. In the present analysis, passage 1 was found to be the most difficult (WIT=113.15) and passage 3 was the easiest (89.41) (see Ohtomo, 1996 for WIT and other similar scales).

Table 3 displays simple correlation coefficients between readability scores, associated parametric values, selected extra linguistic variables and pre- and post-equated subtest scores. Readability scores correlated in the range of the middle .70s to the high .80s with subtest scores, except for Coleman-Liau's statistics. Among the conventional parameters, the number of words per sentence (W/S), or the average sentence length, is most highly correlated with post-equated WIT scores. All other conventional parameters as well as some of the extra variables examined show moderately high (.60s) to high (.80s) correlations with the subtest scores.

Based on the results of simple correlation analysis, multiple regression analyses were run between subtest scores as the dependent variable and various combinations of conventional and new readability factors as independent variables. Although pre-equated percentage scores consistently showed higher correlations

Table 3 Correlation Matrix for All Variables

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. DC95	1																
2. DC48	-.981	1															
3. BM	-.940	.922	1														
4. FRE	.994	-.967	-.931	1													
5. FK	-.854	.851	.890	-.876	1												
6. CL	-.729	.747	.791	-.697	.478	1											
7. S/W	-.960	.953	.898	-.960	.760	.840	1										
8. L/W	-.729	.747	.791	-.697	.478	1.000	.840	1									
9. W/S	-.880	.827	.825	-.896	.912	.353	.736	.353	1								
10. W95	-.940	.948	.885	-.918	.688	.893	.982	.893	.665	1							
11. W48	-.835	.900	.833	-.799	.641	.915	.896	.915	.512	.946	1						
12. T/T	-.084	.015	.091	-.165	.478	-.467	-.062	-.467	.498	-.227	-.347	1					
13. GR	-.904	.877	.965	-.871	.778	.792	.838	.792	.775	.864	.806	-.056	1				
14. HL1	-.810	.875	.814	-.775	.606	.932	.888	.932	.467	.939	.998	-.383	.786	1			
15. HL2	-.373	.466	.310	-.284	-.010	.649	.442	.649	-.014	.598	.703	-.841	.433	.711	1		
16. PC	.866	-.931	-.807	.846	-.858	-.537	-.788	-.537	-.799	-.788	-.808	-.125	-.749	-.771	-.416	1	
17. WIT	**-.843**	**.894**	**.745**	**-.830**	**.848**	**.402**	**.734**	**.402**	**.845**	**.718**	**.702**	**.236**	**.688**	**.656**	**.323**	**-.983**	1

*PC: proportion of correct responses

*Bold figures represent correlation coefficients between the dependent variable (WITs) and independent variat

with individual linguistic variables than post-equated WIT scores, as shown in table 3, it was found that the latter always outperformed the former in the analyses of multiple correlations. This corroborates our earlier prediction that equating would play a substantive role in the present study.

Table 4 reports some of the highest multiple correlations between post-equated WIT scores and combinations of 4 or less variables. Among the highest was the combination of word length based on the number of letters (L/W), sentence length based on the number of words (W/S) and unfamiliar words based on the DC48 word list (W48), producing a multiple correlation (MR) of .9999, an MR^2 (multiple correlation squared) of .999 and an adjusted MR^2 of .997. This finding implies that, as far as the present sample of data is concerned, DC48 will assume a perfect predictive formula for subtest scores, by adding the word length factor to its parameters. Likewise, a highest possible level of multiple correlation was found for a combination of L/W + W/S + HL1 + HL2, and a slightly lower but still extremely high multiple correlation for a combination of L/W + HL1 + HL2 + GR.

Table 4 Multiple Regression Analyses (best fits)

DEPENDENT VARIABLE	INDEPENDENT VARIABLES	MR	MR^2	AM^2
1) WIT	= L/W + W/S + W48	.999	.999	.997
2) WIT	= L/W + W/S + HL1 + HL2	.999	.999	.992
3) WIT	= L/W + HL1 + HL2 + GR	.991	.981	.906

Each combination of these variables will yield the best possible linear prediction of subtest or passage scores. In the present pilot analysis, the following regression equations are tentatively suggested as the basis for further scrutiny.

$$RS1 = 214.544 - 33.694*L/W + 0.838*W/S + 2.038*W48 \quad (1)$$
$$RS2 = 205.749 - 36.225*L/W + 1.052*W/S + 2.143*HL1 + 0.354*HL2 \quad (2)$$
$$RS3 = 238.461 - 57.723*L/W + 3.038*HL1 - 0.590*HL2 + 8.831*GR \quad (3)$$

4. Discussion

The present study is in its pilot stage, and thus any conclusive statements should be deferred until an extended follow-up study is conducted. The discussion will now turn to the aim of the study outlined in section 2, implications of some findings and major operational problems that must be solved in formulating a readability estimation procedure of practical utility.

4.1 Conventional readability statistics and their parameters

Among the six readability statistics, DC48 was most highly correlated with passage scores. FK, DC95 and FRE had correlations of over .8 and BM of over .7. CL was the only index that failed to predict more than 50% of variation in the performance of reading subtests (R^2=.162). DC 48's superiority over DC95 holds even when their word list variables (W48 and W95) are compared in multiple regression analyses. This finding is rather counterintuitive as the older word list contains a large number of archaic or uncommon words that EFL learners rarely encounter in their reading (e.g., afar, gooseberry, nevermore, thee, washtub), and words of relatively recent invention are not included in the list (e.g., computer, motorcycle, TV). A major pitfall of DC95 may lie in the greater ambiguity of its operational guidelines, as compared to those of DC48. In each text analysis, quite a few words were not found in the list but had to be treated as a variety of a word type which was in the list. It seems that DC48 rules are more straightforward, more explicit and easier for users to read and follow when these words are identified.

In the present analysis, it appears that DC48 is the most valid readability measure, but FK and FRE as the least sophisticated type of readability statistics are fairly close to DC48, almost as good as DC95, and better than BM as the most sophisticated type of analysis. Moreover, the convenience of FK and FRE, which can be readily accessed from popular computer programs, may well be preferred in most practical settings.

Among the variables included in the six readability formulas, it was W/S that correlated most strongly with passage scores. In the subsequent multiple regression analyses, W/S was employed as one of the constituent parameters in equations 1 and 2. S/W and W95 produced moderately high correlations, but in the multiple regression analyses, L/W and W48 performed better, interacting with other variables. In fact, the L/W variable, showing a rather low simple correlation, appears in all the three equations and should continue to be posited as one of the most crucial components in a future discussion of readability.

4.2 Candidates for new parameters and formulas

HL1, containing 786 words as opposed to 3000 words in W95 and W48, had moderate correlations with passage scores. On the other hand, HL2, made up of Levels 1 and 2 combined lists of 2564 words, had very low correlations. It appears that HL2 failed to discriminate between difficult and easy passages, probably due to the fact that the combined lists cover almost all the essential words in the six passages, lacking well-tuned sensitivity to substantial lexical difficulty variations.

Nevertheless, HL2 appears to work well with other variables to form a strong set

of predictors as in equations 2 and 3. The effect of employing graded word lists may be partially substantiated. In future analyses, Level 2 could be subdivided into two groups on the basis of word familiarity, and only the more familiar subgroup of words along with HL1 words subsumed under the revised HL2.

The superiority of GR over T/T seems to be assured in the present study. GR had moderately high simple correlations and was included as one of the four independent variables in multiple regression equation 3.

In section 3, three new formulas were tentatively presented as the basis for on-going research. Even though it is too early to conclude whether one or more of these formulas will prove reliable, valid and useful in Japanese EFL settings, the following alternative directions can be inferred from the present line of research.

1) One or more of the conventional readability statistics may prove more reliable and valid than any alternative measures.
2) One or more of the conventional readability statistics may undergo minor revisions including the addition or deletion of the conventional parameters.
3) One or more new readability formulas may be developed on the basis of old and new parameters and be used along with conventional measures.
4) One or more new readability formulas may be developed on the basis of new linguistic variables and be used along with conventional measures.
5) One or more readability measures may be developed and supplant the conventional measures.

4.3 Current problems and future directions

One of the most arbitrary parts of readability estimation is to judge whether a word in question is in the word list, a variation of a word type in the list, one that is not identified in the list or one that is treated as a proper noun, *n*-suffix proper adjective (W48) or a complex or simple numeral (W95), etc.

Guidelines for DC48, for example, are more explicit than those for DC95, but involve some serious problems that cannot be ignored. They stipulate that *n*-suffix proper adjectives such as *American, Australian* are exceptionally treated as familiar, which means *Japanese, Chinese, British, German* are all considered unfamiliar. W48 does not include words of recent invention. The word list is likely to be biased against topics of very recent interest. Some *er*- or *est*-ending words such as *'dancer', 'longer', 'bravest',* are exemplified as familiar words for W95, because the base forms are in the list. This being so, what about *'tier', 'pager', 'digest', 'earnest'* (as opposed to orthographically similar, but semantically unrelated word forms like *'tie', 'page', 'dig', 'earn'* which are in the list)? No clear grammatical classifications are provided to help

users make a quick and reliable judgment.

One solution to some of the above problems is to provide all the possible forms of the words in question. This raises another question. Can it be said unequivocally that the word forms '*be*', '*am*', '*are*', '*were*', '*been*', '*being*' are familiar to all learners of English in Japan? The answer is in the negative. This has led to the inclusion of HL1 and HL2 in the present study.

HL1 and HL2 are not without problems, however. Some basic word forms are accidentally missing (e.g., *were*). In the present pilot analysis, these missing basic words were counted in appropriate levels after being cross-checked with other online word lists founded upon national instructional guidelines or authorized textbooks for Japanese junior and senior high schools. The development of reclassified graded word lists as well as the specification of word entries or associated guidelines merits attention in any future study.

The present study has restricted its attention to variables that can be counted with relative ease. It is hoped that one line of future research will also look at a wider range of variables including lexical density, packet length, as proposed by Harrison and Bakker (1998), as well as more global or implicit factors including text structure and topic. A question is how far the depths of readability can be explored without computer programs and how far computer programs can facilitate such complicated estimation procedures.

A list of pragmatic considerations for future research might be something like the following.

1. A similar study should be conducted on a larger scale involving far more subjects and passages linked through a series of equating procedures. This study is expected to pave the way for one of the directions outlined in the preceding section.

2. If one or more new formulas are construed as stable, readability estimates from these formulas as well as existing models will be computed for many new sets of passages and correlated with passage scores or difficulty self-ratings. If new formulas consistently produce higher correlations with criterion measures than existing models, they will be deemed as sound new formulas.

3. A set of readability estimation procedures will be programmed. The finalized program will not simply help users to get the score quickly but allow them to monitor the process of estimation done in a reliable and valid manner.

5. Conclusion

A variety of readability statistics have been used in EFL teaching and research contexts without questioning how reliable and valid these measures are. The present study compares six conventional readability statistics, their constituent parameters and four interdependent linguistic variables with average scores for six passage-subtests. Based on the results of simple correlation coefficients, multiple regression analyses were performed to see if there are any practical sets of predicting variables that may outperform the existing measures. Three readability equations are tentatively proposed as the baseline of further scrutiny. This line of study should be replicated over a larger number of subjects using a wider variety of reading passages in due course.

Acknowledgements

This study was conducted as part of a larger research project which is being carried out in collaboration with Professor Yoshinori Miyazaki, funded by a research grant from Shizuoka Sangyo University. The author would like to express his special gratitude to Professor Hideo Kiyokawa of Wayo Women's University for his invaluable advice. The author owes his thanks to Professors Adrienne Garden and Takashi Umemoto, as well as Alice Norizuki for administering the tests in their classes. The author also wishes to thank anonymous reviewers of this paper and students who participated in the study.

REFERENCES

Brown, J.D. 1989: Cloze item difficulty. *JALT Journal* 11: 46-67.

Brown, J.D. and Yamashita, S.O. 1995: English language entrance examinations at Japanese universities: What do we know about them? JALT Journal 17: 7-30.

Chall, J. and Dale, E. 1995: *Readability Revisited: The New Dale-Chall Readability Formula.* Brookline Books.

Dale, E. and Chall, J. 1948: A formula for predicting readability: instructions. *Educational Research Bulletin* 27: 37-54.

Daller, H., Van Hout, R., Treffers-Daller, J. 2003: Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics* 24, 197-222.

Duppenthaler, C. 2000: Readability measures of some English readers used in Japanese high schools.

Greenfield, G.R. 1999: Classic readability formulas in an EFL context: Are they valid for Japanese Speakers? Unpublished doctoral dissertation, Temple University, Japan.

Hasegawa, K. and Norizuki, K. (長谷川和則・法月健) 1999: 「CAI英語演習の教授・学習過程とその結果が示唆するもの（その2）—速読演習と穴埋め演習のデータ処理結果を通して—」[The process of CAI English language learning and teaching (Part 2)—an analysis of learners' rapid reading and gap-filling exercises—.] 『静岡産業大学国際情報学部研究紀要』1: 69-80.

Harrison, S. and Bakker, P. 1998: Two new readability predictors for the professional writer: pilot trials. *Journal of Research in Reading* 21, 121-138.

Henning, G. 1987: A Guide to Language Testing: Development, Evaluation, Research. Rowley, Mass: Newbury House Publisher.

Kimura, S. and Visgatis, B. 1996: High school English textbooks and college entrance examinations: a comparison of reading passage difficulty. *JALT Journal* 18: 81-96.

Kiyokawa, H. (清川英男) 1988: 「高校・大学生用リーダビリティ公式の開発」[Development of a readability formula for high school and university students.] 『和洋女子大学英文学会誌』21: 43-63.

—— 1993: 「リーダビリティ研究のためのいくつかの課題」[Some issues associated with readability studies.] 『和洋女子大学英文学会誌』27: 53-67.

—— 2000: 「リーダビリティ」[Readability.] 高梨庸雄・卯城祐司『英語リーディング事典』[*Dictionary of Reading in English.*]

Klare, G. R.A. 1976: Second look at the validity of readability formulas, *Journal of Reading Behavior* 8: 129-152.

Matsuo, H. (松尾秀樹) 1998: 「高校教科書と大学入試問題—リーダビリティから見た比較—」[High school textbooks and university entrance examinations: comparative analysis of readability.] 『英語教育研究』41: 43-63.

Nation, I.S.P. 1990: *Teaching and Learning Vocabulary.* New York: Heinle and Heinle.

Naganuma, K. (長沼君主) 2001: 「コーパスに基づいたリーダビリティの測定とその教育的応用」[Corpus-based analysis of readability.] *On JALT2000—Towards the New Millennium.* 129-135.

Ohtomo, K. (大友賢二) 1996: 『項目応答理論入門』[*Introduction to Item Response Theory.*] 大修館書店.

Ohtomo, K. and Nakamura, Y. (大友賢二・中村洋一) 2002: 『テストで言語能力は測れるか～言語テストデータ分析入門～』[*Can Tests Measure Language Proficiency?: Introduction to Analysis of Language Test Data.*] 大修館書店.

Shiozawa, T. and Aizawa, K. (塩澤利雄・相澤一美) 1989: 「中学校英語教科書のリーダビリティ」[Readability analysis of junior high school textbooks.] 『現代英語教育』6月号 42-44.

Shiozawa, T. and Komaba, T. (塩澤利雄・駒場利男) 1990: 「英語ⅠBの教科書について—リーダビリティを中心に」[On English IB textbooks: focus on readability.] 『現代英語教育』2月号: 13-15.