Corpus-Based Study on Readability of Integrated English Textbooks

HU Tunan[a],*

[a] Lecturer, School of Foreign Languages, Zhejiang University of Finance and Economics, Hangzhou, China.
* Corresponding author.

Abstract

Readability has been used extensively as a quantitative indicator to evaluate the difficulty of reading textbooks. Based on a corpus consisting of 96 texts from two versions of integrated English textbooks published by the top-tier publishing house in China in 2013 and 2020 respectively, this paper employs three readability formulas (i.e. FRE, FKGL and LR) to examine their readability trends and differences in readability. The results show that: 1) the readability of both book sets is moderate, corresponding to the reading level of 8th-9th grade students in the US, 2) no significant differences are found in three readability indices between two book sets, 3) some of LR subindices, such as deep cohesion, show an opposite tendency to the change of overall readability. It is claimed that: 1) both book sets are difficult for the students in the corresponding grades to read and have a scientific hierarchy of readability; 2) the upgrade does not mean that reading difficulty should increase limitlessly; 3) some subindices can be tapped into purposefully to moderate the overall read difficulty. This study tries to provide a quantitative approach to evaluate English textbooks in terms of readability.

Key words: Corpus-based; Readability assessment; Integrated English textbooks

1. INTRODUCTION

Integrated English, also known as extensive reading, is one of the fundamental courses that prepares Chinese English-major undergraduates for future advanced courses. According to the *English Teaching Syllabus for English Majors* issued by the Ministry of Education, the purpose of this course is to gradually improve students' reading ability through comprehensively training the foundational skills in language. Thus, as the primary source of reading materials, textbooks used in this course play a crucial role in the actual teaching process. To some extent, the quality of textbooks, particularly the hierarchy of readability in the same book or book set, potentially influences the success or failure of teaching.

Readability refers to the ease of reading or understanding certain reading or written materials (Dale & Chall, 1948; Klare, 1963; McLaughlin, 1969; Richards et al., 2006). It has been proved that readability is important for knowledge development (Hartley et al., 2002). Assessing the extent to which vocabulary and sentence structure relate to the understanding of materials, readability is helpful for teachers to select proper reading materials for learners (Sheehan et al., 2014). So far, readability formula based on specific indices, such as word length and sentence length, is the prevalent method for evaluating the readability of texts. Since the first readability formula was developed by Lively and Pressey (1923), scholars have endeavored to devise more precise formulas. As a result, a growing number of readability formulas have been proposed. Some of them, such as Flesch Reading Ease Readability Formula (Flesch, 1948), Automated Readability (Smith & Senter, 1967), Smog (Mc Laughlin, 1969) and Flesch Kincaid Formula

(Kincaid et al., 1975), has been more widely adopted for decades. However, an increasing number of scholars have been questioning their validity (Benjamin, 2012; Crossley et al., 2019; Graesser et al., 2011; Lu et al., 2014; Jin et al., 2021). Because in these formulas, word length, sentence length, and percentage of hard words serve as essential subindices for measuring readability. Although the aforementioned indices are directly related to text difficulty, they only represent part of the superficial features of text, that is, part of text difficulty. Recent evidence from cognitive science suggests that the reading comprehension process involves multidimensional levels of linguistic complexity ranging from lexical and syntactic one to cohesion and sentiment (Kintsch & van Dijk, 1978; Just & Carpenter, 1980; Graesser & McNamara, 2011). As a result, to overcome the limitation of traditional readability formulas, recent studies suggested the necessity of considering multiple subindices reflecting different dimensions of linguistic complexity in calculating readability (Stevens et al., 2015). Coh-Metrix L2 Readability Formula is one of many results, which can describe a text in accordance with its syntax, vocabulary, morphology, semantics, etc. (Crossley et al., 2008). Data from several studies show that the accuracy of Coh-Metrix L2 Readability Formula is higher than that of the previous ones (Crossley et al., 2011).

In recent years, the high level of economic growth and development in technology is presenting new challenges for college English education. The most prominent are the changes in the available pathways to obtain fresh reading materials which have become more and more readily available online. The role of this challenge is two-fold. It has, on the one hand, weakened the function of textbooks to provide reading materials in English learning, and on the other, necessitated upgrading the textbooks continuously. As a result, selecting and evaluating textbooks becomes a major area of interest within the field of teaching practice and research. However, compared with qualitative studies in this field, far too little attention has been paid to quantitative analysis (Zhao & Zheng, 2006; Yang & Chen, 2013). Among the few quantitative studies, it is the major focus to assess readability of college English textbooks for non-English-major students with traditional unidimensional formulas (Gu & Guan, 2003; Deng, 2013). Since readability of English major textbooks has not been explored fully yet, the present study is to investigate the readability trend in this unexplored field based on two versions of the same set of integrated English textbooks. More specifically, we aim to solve the following three research questions: (1) Are there any significant differences in three readability indices among the different volumes within the same book set? (2) Are there any significant differences in three readability indices between two book sets? (3) Are there any significant differences in eight subindices of Coh-Metrix L2 readability between two book sets?

2. RESEARCH DESIGN

2.1 Data Collection

The reading materials were taken from two versions of the same set of integrated English textbooks (hereafter referred to as V1 and V2 individually). V2 is the revised version of V1, to be specific. Both were four-volume books published by the premier publishing house in China in 2013 and 2020 respectively. In each book set, volumes were numbered following their order. Volume numbers from 1 to 4 corresponded to grades from low to high, respectively. Typically, the former two volumes were used in the first and second semesters of the freshman year, the latter two in that of the sophomore year. Each volume had fourteen units, each holding two texts. According to the guidelines for the use of the textbook, we excluded texts whose types were self-learning materials after class. In other words, only the first text in each unit was included for the readability assessment. This resulted in a corpus of 96 texts (about 95,785 words in total). The number of words per volume is listed in Table 1.

Table 1
Descriptive Statistics of the Integrated English Textbook Corpus

Version-Volume	V1-1	V2-1	V1-2	V2-2	V1-3	V2-3	V1-4	V2-4
Number	10,379	11,666	11,936	12,236	12,064	12,554	12,042	12,908

Consequently, the resulting corpus does not necessarily fulfill the criteria of representativeness and balance generally recommended in corpus linguistics.

As can be seen from the table above, the number of total words in each volume increased steadily with the volume number. From this, it was observed that in terms of the number of total words, both versions of textbooks were in line with the development law of students' reading ability which should keep increasing along grade years.

Both could meet the needs of students in different grades to improve their reading ability.

2.2 Analysis Tools

In order to assess the readability characteristics of different versions of the same set of integrated English textbooks, we planned to carry out a quantitative analysis from two perspectives: the unidimensional one and the multidimensional one. Thus, in the experiment,

we employed the automated tool Coh-Metrix version 3.0, developed by McNamara team at the University of Memphis (available at http://141.225.61.35/cohmetrix2017).

For the first and second research question, Coh-Metrix automatically calculated the readability scores of both book sets with three formulas: Flesch Reading Ease Readability (hereafter, FRE), Flesch-Kincaid Grade Level (hereafter, FKGL) and Coh-Metrix L2 Readability (hereafter, LR). Below, we will introduce these formulas briefly one by one. The result of FRE ranges from 0 to 100, with a lower score indicating more difficult reading. And the output of FKGL is a number from 1 to 18+. The higher the number, the harder it is to read the text. Both of them are unidimensional formulas which measure readability based on the length of sentences and words within the text, and have their own corresponding difficult level list. In contrast, LR formula is a multidimensional one without the corresponding difficult level list, taking cohesive devices between sentences into consideration. The result of LR can be interpreted in the same manner as that of FRE: the lower the LR score, the lower the readability, and the more difficult for readers to understand the material.

For the third research question, it computed scores on eight subindices of LR individually. To be exact, they are narrativity, word concreteness, syntactic simplicity, referential cohesion, deep cohesion, verb cohesion, connectivity and temporality (McNamara et al., 2014). In the following, we would like to introduce these eight subindices briefly one by one. Narrativity refers to the index that indicates the extent to which a text tells a story. Typically, a narrative text tends to present a story with some basic elements, such as who, what, where and when. So, the higher score of narrativity indicates that the text is more closely affiliated with oral and everyday conversation. Word concreteness is closely related to the word class of a text which can be classified into two categories: the concrete one encoding the concepts evoking mental images easily and the abstract one encoding the concepts difficult to represent visually. The more concrete word a text contains, the more easily it can be process and understand, and vice versa. Syntactic simplicity indicates the extent to which the sentences in a text contain fewer words and use simpler, familiar syntactic structures, which are less difficult to read and process. Referential cohesion is an index on the lexical level, indicating the degree to which there are overlapping words, to be exact nouns in a text. A text with higher score of referential cohesion contains more words that overlaps across expressions and the whole text, which form an explicit line connecting the text for its readers. In contrast, deep cohesion is a syntactic feature reflecting the degree to which a text contains logical connectives when there are logical relationships within the text. If a text contains many logical relationships, such as causality, but does not contain those corresponding connectives, its readers have to spend more efforts on inferring the relationships by themselves. Briefly, the text is more difficult to understand. Similar to referential cohesion, verb cohesion is an index that can be observed on the lexical level also. Specifically, verb cohesion implies the degree to which the same verbs repeat themselves in the entire text, which is relevant for texts intended for younger readers and for narrative texts (McNamara et al., 2012). Connectivity is another syntactic index showing the degree to which a text contains explicit adversative, additive and comparative connectives to express relations in the whole text. Thus, the score of connectivity can be interpreted in the same way with that of deep cohesion. To be exact, the absence of connectives of the above logical relationships in a text will make the material more difficult to read. The last one is temporality, also known as temporal cohesion. It refers to the consistency of tense and aspect in a text, which contributes to the reader's understanding of the events in the text.

2.3 Research Process

After collecting the reading materials and preparing the readability assessment, we began our analysis in three steps.

Step 1, three readability scores of the corpus were calculated to present the overall readability trend of V1 and V2, and the differences in three readability indices among different volumes within the same book set were calculated via One-way ANOVA.

Step 2, the differences in three readability indices between V1 and V2 as well as among different volumes within the same book set were calculated via Independent-Sample.

Step 3, the differences in eight subindices of LR between V1 and V2 were calculated via Independent-Sample T Test.

3. DATA ANALYSIS AND DISCUSSION

3.1 Overall Readability Trend of V1 and V2

This section focuses on presenting the overall readability trend of V1 and V2 from two perspectives: the unidimensional one (FRE, FKGL) and the multidimensional one (LR), and addressing the first research question. We calculated FRE, FKGL and LR scores of each text in our corpus as well as the average scores of each volume within both book sets. Due to space constraints, Table 2 only displays the descriptive statistics of the mean scores.

Table 2
The Mean Scores of FRE, FKGL and LR

	V1-1	V1-2	V1-3	V1-4	V2-1	V2-2	V2-3	V2-4
FRE	73.31	65.91	70.82	65.09	68.62	58.22	70.51	61.38
FKGL	7.23	8.58	7.43	8.83	8.12	9.88	7.59	9.57
LR	19.93	16.16	19.90	18.19	18.53	15.27	17.87	17.88

As mentioned above, although FRE and FKGL are both based on the same quantitative subindices, the manners of interpreting their results are in opposition to each other. Thus, from Table 2, it is clear that FRE and FKGL scores present a trend of negative correlation. More specifically, for the same text or volume, the higher the FRE score, the lower the FKGL one, and vice versa. Measured by FRE, the mean scores of each volume within V1 are 73.31, 65.91, 70.82 and 65.09, within V2 are 68.62, 63.22, 70.51 and 61.38, respectively. According to Flesch (1948), seven of these scores fall into the standard level (value ranging from 60-70), corresponding to the reading level of 8th-9th grade students in the US. And the remaining one (V2-2) falls into the fairly difficult level (value ranging from 50-60), corresponding to the reading level of 10th-12th grade students in the US. Similarly, calculated by FKGL, the mean scores of each volume within V1 are 7.23, 8.58, 7.43 and 8.83, within V2 are 8.12, 9.88, 7.59 and 9.57, respectively. In accordance with Kincaid et al. (1975), FKGL scores ranging from 7.0 to 7.9 correspond to the reading level of 7th grade students, and scores ranging from 8.0 to 9.0 correspond to that of 8th and 9th grade ones. In general, the FKGL difficulty level of both book sets is consistent with the FRE one. Assessed by LR, the mean scores of each volume within V1 are 19.93, 16.16, 19.90 and 18.19, within V2 are 18.53, 15.27, 17.87 and 17.88, respectively. Taken together, these results demonstrate that both book sets have moderate difficulty for reading, aligning with the curriculum orientation of integrated English which serves as a fundamental course in the initial phrase of English major.

According to the guide for faculty provided with the textbooks, different volumes are used accordingly in different semesters of different school years as described previously. Two of three readability indices show that in both book sets, the reading difficulty of the textbooks used in the second semester (vol. 2 and vol. 4) is ranked above that of the textbooks used in the first semester (vol. 1 and vol. 3), while the remaining one (LR) reflect that the reading difficulty of V2-3 and V2-4 are about equal. Ideally, it is expected that students' reading comprehension ability should keep increasing linearly along school years. Correspondingly, the reading difficulty of the textbook should change accordingly. Thus, in a sense, the results from the unidimensional formulas indicate that both book sets have a scientific readability hierarchy and can achieve the course objective that gradually enhances reading comprehension ability, whereas that of the multidimensional one shows that the textbooks of V2 used in the sophomore year cannot adequately meet the teaching demand in the that phrase. This implies that in terms of vol. 3 and vol. 4, V2 as the revised version of V1 may place greater emphasis on the influences of the unidimensional subindices on readability, ignoring that of the multidimensional ones beyond words and sentences.

Furthermore, the decreasing trend in readability between vol. 2 and vol. 3 within both book sets deserves our attention. Theoretically, based on the hypothesis on continuous increase of reading ability, vol. 3 used in the first semester of the sophomore year should be more difficult to read than vol. 2 used in the second semester of the freshman year. However, all three readability indices indicate that the mean difficulty of vol. 3 is much lower than that of vol. 2, even that of vol. 1 (the mean FRE score of V2-3 is higher than that of V2-1). Strikingly, this defect has already presented in V1 but remains unsolved in V2.

To answer the first research question, a one-way ANOVA was conducted to calculate the significance of the differences in three readability indices among different volumes within the same book set. The results are shown in Table 3.

Table 3
The Results of One-way ANOVA of Differences in Three Readability Indices among Different Volumes within the Same Book Set

	F [3, 44]	p
V1 (FRE)	1.243	0.306
V2 (FRE)	2.049	0.121
V1 (FKGL)	0.952	0.424
V2 (FKGL)	1.758	0.169
V1 (LR)	1.330	0.277
V2 (LR)	0.981	0.411

From Table 3, it can be seen that in the three types of readability, significant differences are absent among the scores of four volumes within V1(p = 0.306, 0.424, 0.277) and among that within V2 (p = 0.121, 0.169, 0.411) both. This finding appears to contradict previous one reported in Table 2. In accordance with Table 2, the results of three readability indices imply that the overall readability trends of both book sets are roughly in line with the law of the development of students' reading ability. In order

to achieve the primary objective of integrated English to improve reading comprehension gradually, the textbooks used in the second semester are purposefully designed to be more difficult to read than that used in the first one of the same school years, despite the fact that the hierarchy of readability between vol. 2 and vol. 3 within both book sets needs to be further optimized. However, the statistics reported in Table 3 showed the opposite result that the difference in readability between the textbooks used in the different semesters of the same school years was not significant. This does not mean that the reading difficulty of the volumes used in the second semesters is far from the expected goal. Our previous analysis revealed that the texts in both book sets are difficult to read, corresponding to the reading level of 8th-9th grade students in the US. Indeed, the reading materials on this difficult level can meet the basic need of integrated English course. It is worth noting that the reading difficulty is not "the higher the better". As a result, the quest for high reading difficulty should not be pursued blindly without understanding its relevance to the instructional objective and the instructional needs.

3.2 Differences in Three Readability Indices

The second research question is answered in this section. Intuitively, according to the previous analysis, larger or smaller differences exist in three readability indices between V1 and V2. It is necessary for us to examine whether the differences are significant statistically. An Independent-Sample T Test was conducted to examine the significance of the differences in three readability indices between V1 and V2. Table 4 reports the results of Independent-Sample T Test of differences in three readability indices between V1 and V2.

It can be seen from Table 4 that in terms of three readability indices, no statistically significant differences are detected between the scores of V1 and that of V2 ($t = 1.397, -1.295, 1.079, p = 0.166, 0.198, 0.283$). As mentioned previously, the new challenges in college English education require textbooks to raise difficulty. As a result, V2 should be more difficult to read than V1 theoretically. However, it does not mean that reading difficulty is allowed to increase linearly without limit. Because the reading difficulty of the integrated English textbooks is without absolute standard or value. In other words, it is entirely possible that the level of the reading difficulty of a textbook can be increased unlimitedly. If this case is true, the textbook will be reduced to a worthless collection of difficult reading materials, which fails to play its due role in English learning. As we have stated before, the blind pursuit of high reading difficulty is not desirable. The readability of textbooks should keep balance with students' reading comprehension ability and teachers' practical demands in the teaching process.

Table 4
The Results of Independent-sample T Test of Differences in Three Readability Indices between V1 and V2

	M	SD	t	p
V1 (FRE)	68.61	12.90	1.397	0.166
V2 (FRE)	64.68	14.59		
V1 (FKGL)	8.02	2.85	-1.295	0.198
V2 (FKGL)	8.79	2.97		
V1 (LR)	18.54	5.42	1.079	0.283
V2 (LR)	17.40	5.04		

3.3 Differences in Eight Subindices of LR

In this section, we focus on answering the third research question. As shown earlier, we found no significant differences in three readability indices between V1 and V2. Distinguished from the unidimensional formulas based on the length of sentences and words, such as FRE and FKGL, eight additional subindices are considered in LR which enables a more in-depth and integrated assessment of readability. To investigate further the similarities and differences between V1 and V2, an Independent-Sample T Test was used to compare the eight subindices of LR. Results are given in Table 5 below.

Table 5
The Results of Independent-sample T Test of Difference in Eight Subindices of LR between V1 and V2

	M	SD	t	p
V1 (Narrativity)	0.54	0.79	1.004	0.318
V2 (Narrativity)	0.37	0.87		
V1 (Syntactic simplicity)	0.00	0.67	0.444	0.658
V2 (Syntactic simplicity)	-0.06	0.59		
V1 (Word Concreteness)	0.03	0.72	0.167	0.868
V2 (Word Concreteness)	0.01	0.74		
V1 (Referential cohesion)	-0.60	0.63	0.616	0.539
V2 (Referential cohesion)	-0.68	0.62		
V1 (Deep cohesion)	0.29	0.57	-0.616	0.540
V2 (Deep cohesion)	0.36	0.53		
V1 (Verb cohesion)	0.25	0.68	0.128	0.898
V2 (Verb cohesion)	0.23	0.75		
V1 (Connectivity)	-2.38	0.98	1.708	0.091
V2 (Connectivity)	-2.71	0.92		
V1 (Temporality)	-0.24	0.72	-0.243	0.808
V2 (Temporality)	-0.21	0.76		

From Table 5, it can be seen that no statistically significant differences are found in the eight subindices of LR. Here the results confirm our previous analysis

that reading difficulty should be referenced to students' reading comprehension ability and teachers' practical demands in the teaching process. Moreover, as shown in Table 4, V2 is more difficult to reading than V1. With the decrease of the overall readability, seven subindices show the corresponding trend. To be exact, their scores decrease with the upgrade, implying the increase of reading difficulty. However, the remaining one, deep cohesion shows a trend opposite to the overall decline. The difference in the single index does not allow us to reach a firm conclusion about the pros and cons of the two book sets. Although V2 as the revised version should be difficult enough to respond to the new situations challenging the traditional English education, the overall change of readability does not mean that all the subindices have to change correspondingly. When necessary, it is reasonable that overall readability can be adjusted by controlling some subindices.

4. CONCLUSION

This research provides a readability assessment of two versions of the same set of integrated English textbooks. Different from the previous studies measuring readability through only the unidimensional formulas, the present study examined the readability of the textbooks from two perspectives: the unidimensional one (i.e. FRE and FKGL) and the multidimensional one (i.e. LR). Our findings first show that the texts of both book sets are observed to have lower readability and the hierarchy of readability between the textbooks contributes to improve student's reading ability gradually in the initial phrase. Our results also show that there are no significant differences in three readability indices as well as in eight subindices of LR between V1 and V2. Our analysis further reveals that readability is not a score to be interpreted as "the higher the better" and some subindices can be tapped into purposefully to moderate the overall reading difficulty.

As a preliminary study, there are still some shortcomings in the current study. First, not only need our corpus of the integrated English textbooks to be further expanded in size to include the reading materials from more versions of the integrated English textbooks published by different publishing house, but it also remains to be enhanced by greater depth, breadth, and complexity. For example, the self-learning materials after class which has been excluded by the current research can be incorporated into the future experiment as an independent variable to evaluate the readability of the entire textbooks. Moreover, in addition to comparing the latest version of a textbook with its previous version, future research may consider tracking the evolution of one set of the integrated English textbooks through evaluating the change of its readability diachronically from the first version to the latest one. Second, there is a large need to

develop an evaluation criterion for assessing whether the difficulty of specific reading materials fulfills students' needs at the corresponding stage. We will examine the relationship between *China Standards of English* issued by the Ministry of Education and State Language Commission both and the readability scores from the indices utilized in the current research. Third, we would also like to acknowledge that this study is constrained because of the shallow nature of readability formulas based on the superficial features of texts. Although LR is a so-called multidimensional readability formula, it still operates at the lexical and syntactic levels, ignoring the subjective feedback from readers. Finally, readability is only one of many factors that can influence the quality of textbooks, and it will be more useful to examine the interaction between readability and such factors as theme, genera, source, etc.

REFERENCES

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review, 24*(1), 63-88.

Biber, D. (1998). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Crossley, S. A., D. Allen & D. S. McNamara. (2011). Text readability and intuitive simplification: a comparison of readability formulas. *Reading in a Foreign Language*, *23*(1): 84-102.

Crossley, S. A., Salsbury, T., McCarthy, P. M., & McNamara, D. S. (2008), LSA as a measure of coherence in second language natural discourse. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 1906-1911). Washington, DC: Cognitive Science Society.

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading, 42*(3-4), 541-561.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, *27*(2), 37-54.

Deng, W. B. (2013). A comparative study on new and old editions of intensive reading textbooks of 21st century college English. *Journal of Changchun University*, *23*(1), 118-121.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221-233.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, *3*(2), 371-398.

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

Gu, X. D., & Guan, X. X. (2003). A sample study on readability of CET reading test and reading materials of college

English. *Journal of Xi'an International Studies University*, *11*(3), 39-42.

Hartley, J., Sotto, E., & Pennebaker, J. (2002). Style and substance in psychology: Are influential articles more readable than less influential ones? *Social Studies of Science, 32*(2), 321-334.

Jin, T., Duan, H. Q., Lu, X. F., Ni, J., & Guo, K. (2021). Do research articles with more readable abstracts receive higher online attention? Evidence from Science. *Scientometrics, 126*(8), 8471-8490.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329-354.

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*. Retrieved October 11, 2022, from http://www.dtic.mil/dtic/tr/fulltext/u2/a006655. pdf

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363-394.

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

Lively, B. A., & Pressey, S. L. (1923). A method for measuring the vocabulary burden of textbooks. *Educational Administration and Supervision, 9*(7), 389-398.

Lu, X., Gamson, D. A., & Eckert, S. A. (2014). Lexical difficulty and diversity in American elementary school reading textbooks: Changes over the past century. *International Journal of Corpus Linguistics, 19*(1), 94-117.

McLaughlin, G. H. (1969). SMOG grading: A new readability formula. *Journal of Reading*, *12*(8), 639-646.

McNamara, D. S., A. C. Graesser & M. M. Louwerse. (2012). Sources of text difficulty: across genres and grades. In J. P. Sabatini, E. Albro & T. O'Reilly. *Measuring up: Advances in how we assess reading ability* (pp. 89-116). Plymouth: Rowman & Littlefield Education.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Zhiqiang, C. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. New York: Cambridge University Press.

Richards, J. C., Platt, J., & Platt, H. (1992). *Longman dictionary of language teaching and applied linguistics*. London: Longman.

Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal*, *115*(2), 184-209

Smith, E., & Senter, R. (1967). *Automated readability index (AMRL-TR-66-22)*. Ohio: Aerospace Medical Re.

Stevens, R. J., Lu, X., Baker, D. P., Ray, M. N., Eckert, S. A., & Gamson, D. A. (2015). Assessing the cognitive demands of elementary school reading curricula: An analysis of reading text and comprehension tasks from 1910 to 2000. *American Educational Research Journal, 52*(3), 582-617.

Yang, G., & Chen, L. J. (2013). A study on use of reflections on skills under various interaction patterns during asynchronous online discussion. *Foreign Languages and Their Teaching, 269*(2), 16-19.

Zhao, Y., & Zheng, S. T. (2006). Theoretical Analysis on some foreign systems of English textbook evaluation. *Foreign Language Education*, *27*(3), 39-45.