Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 10, No. 1, 2023

Evaluation of Reading Difficulty Based on Factor Analysis and Linear Regression

Wenjing Ma

Anhui University of Finance and Economics, Bengbu, 233000, China

Abstract: This paper conducts research on how to measure the difficulty of English texts and how to extend it to other languages, such as Chinese. Firstly, text features that can represent reading difficulty are extracted and represented using parameters. Secondly, the extracted indicators are subjected to KMO and Bartlett's tests, and the results show that these indicators are suitable for factor analysis. Then, a comprehensive evaluation model of language difficulty based on factor analysis is constructed, and the difficulty coefficient composite score F of 16 English reading texts is calculated. Additionally, a linear regression equation is constructed to separately measure the correlation between Type-Token Ratio (TTR) and the text difficulty coefficient F. Finally, the conclusions obtained from the model are compared with the officially published difficulty coefficients of English reading in the postgraduate entrance examination, validating the rationality of the results. Furthermore, the above constructed multi-indicator comprehensive evaluation model is extended to other languages such as Chinese, confirming the rationality and universality of the model, effectively solving the problem of "measuring text reading difficulty.".

Keywords: Measuring text difficulty, Factor analysis, Linear regression analysis, Comprehensive evaluation model.

1. Introduction

The concept of graded reading was proposed by educators in the early 20th century. As people's emphasis on reading has increased, there has been growing attention towards graded reading. Automated grading technology for reading has also made certain advancements[1]. For English learners, choosing correct and appropriate English reading materials can help them to improve their reading ability in an orderly manner. Lexile[2], AR[3], DRA, GRL, etc. are common graded reading systems, and Flesch Reading Ease (FRE)[4] is also an effective method for scoring text readability. The difficulty of the text can be reflected according to the language feature index, and readers, parents, teachers, and librarians can know whether the text is suitable for reading as long as they judge whether the indicators are consistent with the reader's ability.

Based on the corpus-based foreign language reading assessment text difficulty, Zhan Xianjun[1] evaluated the average length and word level distribution of the 2012 postgraduate English reading text based on the TOEFL reading text. But it mainly provides theoretical enlightenment, without model conclusion demonstration. And the data is only taken from part A of the English reading text in one test paper (2012 postgraduate English test questions), and the sample is under-represented. Lv Jiantao[6] assesses English reading ability based on the Rasch model, putting reading ability and material difficulty on the same scale so that they can be compared with each other. The Rasch formula calculates "reading ability-material difficulty = possibility of reading comprehension". However, the application is tested by students, and the teacher draws conclusions after analyzing the characteristics and scores of the students for subsequent prediction. Not adapted to the situation where there are many readers in the society. Cheng Yong and Xu Dekuan[6]used a computer to study the automatic grading of Chinese text reading difficulty based on the fusion of multilingual features and deep features, and used a neural network pre-training model based on BERT to extract the depth features of sentences in the text, and build on this basis An end-to-end neural network is used to fuse language features and depth features to achieve automatic classification. However, the statistical sample is only for the study of Chinese, and English and other language systems are not considered. All in all, the existing literature has more or less deficiencies in formulas, models, and theories. The indicators used are mostly limited to the text itself, lacking readers' own influence factors, insufficient experience and data support, which need to be improved.

2. Basic Model

2.1. Terms, Definitions and Symbols

2.1.1. Explanation of Terms

(1) Text readability

Readability refers specifically to the product of a certain writing style, referring to the complexity of text and sentence structure in a piece of content. The task of readability analysis is, given a text, by analyzing the text, give the difficulty value of the text or judge which level of reader the text is suitable for. The assumption is that complex sentences are more difficult to decompose and read than simple ones. Generally described as the ability to read easily.

(2) Types/Tokens ratio

Types: How many word forms are there in the text. Tokens: Refers to the total number of words in the text. Types/tokens ratio refers to the proportion of types and tokens, that is, the variability of words. The role is to explain the change and richness of the vocabulary.

(3) Number of new words

We constructed a basic vocabulary based on the Chinese Students' Spoken and Written English Corpus (SWECCL) and Brown Corpus , and compared all the vocabulary of the selected text with the basic vocabulary. The number of words beyond the range of the basic thesaurus is the number of new words.

(4) Subject difficulty factor

According to the English text subject difficulty coefficient

correspondence table released by the Library Association of the United Kingdom, we have selected the subject matter of the text as a consideration factor and matched the selected English text to obtain the subject difficulty of each sample text coefficient.

(5) Proportion of low-frequency words

The proportion of low-frequency words refers to the proportion of low-frequency words among all words used in a text. In Chinese text analysis, we use Peking University's Modern Chinese Corpus as the basic vocabulary. In this paper, we set low-frequency words as words that appear less than 1,000 times in the basic vocabulary.

2.1.2. Symbol Description

Table 1. Symbol Description

Serial number	Symbol	Symbol Description
1	x_1	Sentences
2	x_2	Mean (in words)
3	x_3	Subject matter difficulty factor
4	x_4	Unfamiliar word
5	y	Type/Token ratio (TTR)
6	x_1'	Average word frequency
7	x_2'	Sentence length variation
8	x_3'	Word richness
9	x_4'	Proportion of low-frequency words
10	y'	type/token ratio (TTR)
11	2016(1)	The first text of the 2016 postgraduate entrance examination English text reading
12	2017text1	2016 College Entrance Examination Chinese National Paper One Essay Text

2.2. Basic Assumptions

(1) Assuming that the influence of readers' reading attitude, reading interest, reading experience and knowledge reserves on the readability of the text is ignored;

(2) It is assumed to ignore the influence of the form of reading material on the readability of the text (paper, electronic, etc.);

Assuming that the influence of the reading environment on the readability of the text is ignored (weather, location, etc.);

(3) Assuming that the influence of social and cultural background differences on the readability of the text is assumed to be ignored.

3. The Foundation of Model

The main objective of this paper is to establish a model for assessing the difficulty of reading English texts. According to experience and related language analysis papers, The reading difficulty of English text has a certain correlation with the following five characteristic indicators "Sentences", "Mean (in words)", "Subject matter difficulty factor", "Unfamiliar word", and "Type/Token ratio(TTR)"[8]. To facilitate subsequent analysis, we number the first four variable indicators as x_1, x_2, x_3, x_4, and the last indicator as y.

3.1. Data Preparation

3.1.1. Quantification Methods of Sample Index

We filtered through the Internet to obtain a certain representative English text: the four English readings of the 2015-2018 postgraduate entrance examination English：16 samples in total. The most important thing in these articles is the statistics of specific quantitative parameter values of the indicators. With the help of the language statistics tools Wordsmith tools and AntConc, the parameter values of the four indicators x_1, x_2, x_4, y of the difficulty of reading English text can be obtained respectively. Corresponding to the index type symbolic symbol ratio, the smaller the y , the greater the richness of the article, and the higher the reading difficulty level of the text. In particular, relative to the objectivity of other indicators, the indicator x_3 has a certain degree of subjectivity. In order to obtain quantitative indicators, we conducted the following statistical analysis on the difficulty coefficients of different texts by consulting the corresponding table of the difficulty coefficients of English texts published by the Library Association of the United Kingdom, see Table 2.

Table 2. Difficulty factor

Number	Topics	Difficulty factor
1	natural science	0.388
2	medicine	0.381
3	jurisprudence	0.380
4	economics	0.378
5	Political science	0.375
6	ecology	0.374
7	Literary review	0.369
8	sociology	0.367
9	pedagogy	0.367
10	psychology	0.362
11	Communication	0.361

3.1.2. Get sample index parameter values

By analyzing the subject matter of the English text reading for each postgraduate entrance examination, and corresponding to the subject difficulty coefficient table, we have obtained the quantitative values of all selected indicators, see Table 3.

Table 3. Subject difficulty coefficient table

Text file	Sentencesx_1	Mean(in words)x_2	Subject matter difficulty factorx_3
2015(1)	20	21.2	37.50
2015(2)	21	21.1	38.00
2015(3)	18	24.17	36.70
2015(4)	21	21.33	36.70
2016(1)	20	21.45	36.10
2016(2)	32	14.06	37.40
2016(3)	18	23.72	36.70
2016(4)	25	17.44	36.70
2017(1)	24	18.46	36.10
2017(2)	21	20.05	36.70
2017(3)	20	21.2	37.80
2017(4)	20	19.95	37.50
2018(1)	29	15.59	36.10
2018(2)	20	20.4	36.70
2018(3)	23	18.65	36.70
2018(4)	20	20.35	37.80
Text file	Unfamiliar wordx_4		Type/Token ratio (TTR)/%y
2015(1)	17		58.73
2015(2)	20		54.85
2015(3)	20		54.25
2015(4)	20		52.90
2016(1)	15		57.11
2016(2)	18		60.89
2016(3)	21		56.44
2016(4)	18		52.75
2017(1)	16		57.34
2017(2)	19		57.72
2017(3)	18		53.54
2017(4)	23		56.14
2018(1)	15		61.50
2018(2)	19		54.90
2018(3)	18		54.78
2018(4)	18		60.69

3.2. Model Establishment and Verification

3.2.1. KMO and Bartlett test

Since the prerequisite of factor analysis is whether there is a strong correlation between the observed variables, we use the KMO and Bartlett tests to examine the correlation among the selected variables. The test results are shown in the following Table 4.

Table 4. KMO and Bartlett test

KMO and Bartlett test		
KMO sampling appropriateness number		.622
Bartlett sphericity test	Approximate chi-square	39.917
	Degree of freedom	6
	Significance	.000

It can be seen from the table that the KMO value is greater than 0.6, and the significance of the Bartlett sphericity test approaches 0, indicating that the data is suitable for factor analysis.

3.2.2. Evaluation Model of Language Difficulty

The establishment and solution of the model can be divided into five main steps as follows:

Standardize the raw data

There are 4 index variables for factor analysis, namelyx_1, x_2, x_3, x_4, A total of 16 evaluation objects. The value of the j index of the i evaluation object is b_{ij}, $i = 1,2, \dots, 16$; j=1,..., 4. Convert each index b_{ij} value into standardized index $\tilde{b}_{ij}$:

$$\tilde{b}_{ij} = \frac{b_{ij} - \bar{m}_j}{n_j}, i = 1,2, \dots, 16; j=1,\dots,4 \quad (1)$$

$\bar{m}_j, n_j$ is the sample mean and sample standard deviation of the j indicator:

$$\bar{m}_j = \frac{1}{16}\sum_{i=1}^{16} b_{ij} \tag{2}$$

$$n_j = \sqrt{\frac{1}{16-1}\sum_{i=1}^{16}\left(b_{ij} - \bar{m}_j\right)^2} \tag{3}$$

Correspondingly, $\tilde{x}_j$ is the standardized indicator variable:

$$\tilde{x}_j = \frac{x_j - \bar{m}_j}{n_j}, j = 1, \ldots, 4 \tag{4}$$

Calculate the correlation coefficient matrix R

Correlation coefficient matrix $R = (r_{ij})_{4\times4}$, have

$$r_{ij} = \frac{\sum_{k=1}^{16} \tilde{b}_{ki} \cdot \tilde{b}_{kj}}{16-1}, i, j = 1, \ldots, 4 \tag{5}$$

Where $r_{ij} = 1, r_{ij} = r_{ji}$. r_{ij} is the correlation coefficient between the i indicator and the j indicator.

Calculate the elementary load matrix

Calculate the eigenvalues of correlation coefficient matrix R, $\lambda_1 \geq \cdots \geq \lambda_4 \geq 0$, and the corresponding feature vector $m_1, \ldots, m_4$, of which $m_j = [m_{1j}, \ldots, m_{4j}]^T$ Elementary load matrix:

$$\Lambda_1 = \left[\sqrt{\lambda_1 m_1}, \sqrt{\lambda_2 m_2}, \ldots, \sqrt{\lambda_4 m_4}\right] \tag{6}$$

Choose $p(p \leq 4)$ principal factor

According to the elementary load matrix, calculate the contribution rate of each common factor, and select p main factor. Rotate the extracted factor load matrix to get the matrix $\Lambda_2 = {\Lambda_1}^{(p)} T$, (of which ${\Lambda_1}^{(p)}$ is the first p columns of Λ_1, T is an orthogonal matrix), Constructive factor model.

$$\begin{cases} \tilde{x}_1 = \partial_{11}F_1 + \cdots + \partial_{1p}F_p \\ \vdots \\ \tilde{x}_4 = \partial_{41}F_1 + \cdots + \partial_{4p}F_p \end{cases} \tag{7}$$

Where

$$\Lambda_2 = \begin{bmatrix} \partial_{11} & \cdots & \partial_{1p} \\ \vdots & \ddots & \vdots \\ \partial_{41} & \cdots & \partial_{4p} \end{bmatrix} \tag{8}$$

We select two main factors, the first common factor F_1 is the total number of words in the text, and the second common factor F_2 is the number of new words. We use the Matlab program to calculate the factor contribution and contribution rate after rotation, see Table 5.

Table 5. Contribution rate

Factor	Contribution	Contribution rate	Cumulative contribution rate /%
1	2.357	58.4153	58.925
2	1.113	23.7688	86.747

The factor loading matrix see Table 6.

Table 6. The factor loading matrix

Index	Main factor 1	Main factor 2
Sentences	0.6656	0.2319
Mean (in words)	-0.3236	0.9345
Subject matter difficulty factor	0.9696	0.1537
Unfamiliar	-0.9213	0.0011

Based on the cumulative contribution rate, observing and analyzing the gravel map below, see Figure 1. We can see that the four factors we selected can roughly cover the information used by the variables.

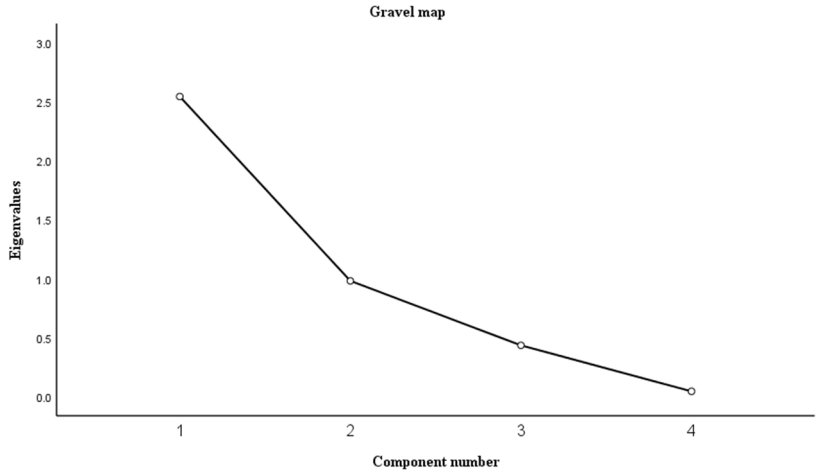

Figure 1. The gravel map

Calculate factor scores and conduct comprehensive evaluation

Use regression method to obtain single factor score function:

$$\hat{F}_j = \beta_{j1}\tilde{x}_1 + \beta_{j2}\tilde{x}_2 + \cdots + \beta_{j4}\tilde{x}_4, j = 1,2 \tag{9}$$

Record the estimated value of the ith sample point to tjth

factor F_j score:

$$\hat{F}_{ij} = \beta_{j1}\tilde{b}_{i1} + \beta_{j2}\tilde{b}_{i2} + \cdots + \beta_{j4}\tilde{b}_{i4}, i = 1,2,\cdots,16; j = 1,2 \quad (10)$$

Then there is

$$\begin{bmatrix} \beta_{11} & \beta_{21} \\ \vdots & \vdots \\ \beta_{14} & \beta_{24} \end{bmatrix} = R^{-1}\Lambda_2 \quad (11)$$

And

$$\hat{F} = \left(\hat{F}_{ij}\right)_{16\times 2} = X_0 R^{-1}\Lambda_2 \quad (12)$$

Where $X_0 = \left(\tilde{a}_{ij}\right)_{16\times 7}$ is the standardized data matrix of the original data; R is the correlation coefficient matrix; Λ_2 is the load matrix obtained in the previous step.

Calculate the score function of each factor:

$$F_1 = 0.3375\tilde{x}_1 + 0.1158\tilde{x}_2 + 0.4424\tilde{x}_3 - 0.3811\tilde{x}_4 \quad (13)$$

$$F_2 = 01635\tilde{x}_1 + 0.9859\tilde{x}_2 + 0.0509\tilde{x}_3 + 0.1013\tilde{x}_4 \quad (14)$$

Using the comprehensive factor scoring formula:

$$F = \frac{58.42F_1 + 23.77F_2}{86.747} \quad (15)$$

Calculate the comprehensive scores of the difficulty of 16 samples of English texts, see Table 7.

Table 7. The comprehensive scores

Rank	1	2	3	4	5	6	7	8
F_1	-0.5735	0.2371	-0.8102	0.0772	-0.6830	2.4088	-0.9687	0.6293
F_2	0.1895	1.5565	1.7247	-0.1726	-0.7559	0.5151	0.4635	-2.1179
F	-0.3321	0.6546	-0.0081	-0.0018	-0.7061	1.8096	-0.5155	-0.2400
Text file	2017(2)	2015(3)	2016(2)	2016(1)	2018(3)	2015(1)	2018(2)	2016(4)

Rank	9	10	11	12	13	14	15	16
F_1	0.6247	-0.2922	-0.6192	-0.7974	1.9452	-0.7176	0.2437	-0.7042
F_2	-0.7236	-0.1655	-1.0224	0.4866	0.9078	-0.0750	-1.1155	0.3046
F	0.1981	-0.2521	-0.7468	-0.3911	1.6170	-0.5143	-0.1864	-0.3850
Text file	2015(4)	2017(1)	2018(4)	2017(4)	2015(2)	2018(1)	2016(3)	2017(3)

From the above table, we can directly and clearly observe the two main factor function values F_1、F_2 comprehensive factor score F and comprehensive ranking of each sample: the second English text of the 2017 postgraduate entrance examination is the most difficult to read, and the other texts are successively less difficult to read.

Then, the conclusions drawn by the model are compared with the official English reading difficulty coefficient of postgraduate entrance examination to verify the rationality and accuracy of the results, see Table 8.

Table 8. The official English reading difficulty coefficient of postgraduate entrance examination

Year	Passage	Degree of Difficulty
2015	Text1	Upper middle
	Text2	difficult
	Text3	middle
	Text4	easy
2016	Text1	easy
	Text2	difficult
	Text3	Upper middle
	Text4	middle
2017	Text1	Upper middle
	Text2	middle
	Text3	easy
	Text4	middle
2018	Text 1	difficult
	Text 2	easy
	Text 3	middle
	Text 4	Upper middle

Finally, through correlation analysis, it is concluded that the correlation coefficient between the degree of difficulty of English text $F$and the Type/Token ratio (TTR) is 0.6544, which shows that there is a moderate correlation between the two. Therefore, we calculate the regression equation F of and TTR for:

$$F = 2.3010\text{-}0.0623y \quad (16)$$

In the previous literature, only the qualitative relationship between y, which is the ratio of symbol-like symbols, was given. The formula derived from this question model constructed a quantitative functional relationship between the degree of difficulty of English text and the ratio of symbol-like symbols.

3.3. Model Expansion

The model established above has a certain generality and can be extended to other languages as needed. Only the relevant variable indicators need to be modified during promotion. Under normal circumstances, the difficulty of reading Chinese text is related to the "average word frequency" and "sentence length variation". Variables such as "word richness", "proportion of low-frequency words", and the index of the"Type/Token ratio"[9]are also have a large degree of relevance with the difficulty of reading Chinese text. Set each variable and index as $x_1', x_2', x_3', x_4', y'$.

In order to obtain the quantitative relationship between the variables and indicators and the difficulty level, we select the modern text reading of the college entrance examination as a sample for analysis. The following Table 9 shows the specific statistical parameter values of each variable and indicator:

Table 9. The specific statistical parameter values of each variable and indicator

Text file	Average word frequency x_1'	Sentence length variation x_2'	Word richness x_3'
2017 text1	52.00	24.22	0.29
2017 text2	53.00	21.98	0.28
2017 text3	50.00	25.35	0.27
2018 text1	51.00	25.34	0.27
2018 text2	51.00	23.84	0.25
2018 text3	49.00	22.12	0.29
2019 text1	49.00	20.02	0.29
2019 text2	51.00	22.35	0.29
2019 text3	48.00	21.22	0.28

Text file	Proportion of low-frequency words x_4'	Type/Token ratio (TTR)/% y'
2017 text1	3.10	11.34
2017 text2	3.00	14.94
2017 text3	2.90	17.35
2018 text1	0.031	12.11
2018 text2	0.028	9.61
2018 text3	0.028	13.21
2019 text1	0.029	18.60
2019 text2	0.030	14.07
2019 text3	0.031	12.46

Note: Sentence length variation refers to the sentence length variation in the text; Word richness refers to the diversity of the semantic categories of the words in the text; Average word frequency The average word frequency value in the text; Proportion of low-frequency words The percentage of low-frequency words in the text

Below we perform quantitative calculations on the value of each variable. Based on the model established above, we can also obtain the quantitative linear relationship between the degree of difficulty F of English text and the Type/Token ratio (TTR):

$$F = 1.3381\text{-}0.0974y' \quad (17)$$

From this formula, the difficulty level of Chinese text reading can be directly and quantitatively calculated. Other languages can also use this model to quantitatively obtain the linear relationship between the text reading difficulty F and the TTR.

4. Conclusion

This article proposes a comprehensive language difficulty evaluation model based on multiple indicators, including "the number of sentences", "mean (in words)", "subject matter difficulty factor", "the number of unfamiliar words" and "type/token ratio (TTR)". The model predicts the difficulty coefficients of English reading comprehension questions in the official postgraduate entrance examinations from 2015 to 2018. The results obtained from the model are compared with the official English reading difficulty coefficient of postgraduate entrance examination, and it is found that the conclusions are generally consistent. This indicates that the model can effectively measure the difficulty level of English texts, helping readers choose texts suitable for their reading level. It can also be used to investigate whether the reading teaching goals determined in the syllabus are achieved and assist in adjusting the teaching guidance plan. In addition, the English text difficulty assessment model has good universality and can be extended to other languages, such as Chinese.

References

[1] S.M.Rao, H Zheng, S.J.Li: A Survey of Leveled Reading. Proceedings of the 20th Chinese National Conference on Computational Linguistics, 2021, p.689-702.

[2] Ar H R, Utomo P, Rajendra M Y: Classification of Lexile Level Reading Load Using the K-Means Clustering and Random Forest Method[J]. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control,2020.

[3] M C T, E S E, B K A, et al.: Effectiveness of accelerated reader on children's reading outcomes: A meta-analytic review.[J]. Dyslexia (Chichester, England),2022,29(1).

[4] Eleyan D, Othman A, Eleyan A. Enhancing Software Comments Readability Using Flesch Reading Ease Score. Information. 2020; 11(9):430.

[5] X.J.Zhan: Study on the validity verification of the validity of the reading text in the foreign language test - The perspective based on the corpus. Foreign language test and teaching, 2015 (03): 23-29.

[6] J.T.Lu, M.Z.Yang, D.C.Zhang, et al.: Testing and Evaluation of English Reading Ability based on Rasch Model. Science and Technology Vision, 2019 (29): 76-77-75.

[7] Y Cheng, D.G.Xu, J Dong: Automatic grading study of Chinese the difficulty of reading texts based on the integration of multilingual and deep features, journal Chinese Information, 2020, 34 (04): 101-110.

[8] Y Tao: Improve the corpus and improve the quality of Chinese cultural translation. China Journal of Social Sciences, 2017-09-26 (003).

[9] Y Cheng, D.G.Xu, J Dong: Analysis of key factors and easy-to-read formulas for text reading difficulty grading based on the language textbook corpus, 2020 (01): 132-143.