Rafatbakhsh and Ahmadi
Asian. J. Second. Foreign. Lang. Educ. (2023) 8:41
https://doi.org/10.1186/s40862-023-00214-4

Asian-Pacific Journal of Second and Foreign Language Education

RESEARCH Open Access

Predicting the difficulty of EFL reading comprehension tests based on linguistic indices

Elaheh Rafatbakhsh[1*] and Alireza Ahmadi[1]

*Correspondence:
e.rafatbakhsh@shirazu.ac.ir

[1] Department of Foreign Languages and Linguistics, Faculty of Literature and Humanities, Shiraz University, Shiraz 7194684795, Iran

Abstract

Estimating the difficulty of reading tests is critical in second language education and assessment. This study was aimed at examining various text features that might influence the difficulty level of a high-stakes reading comprehension test and predict test takers' scores. To this end, the responses provided by 17,900 test takers on the reading comprehension subsection of a major high-stakes test, the Iranian National University Entrance Exam for the Master's Program were examined. Overall, 63 reading passages in different versions of the test from 2017 to 2019 were studied with a focus on 16 indices that might help explain the reading difficulty and test takers' scores. The results showed that the content word overlap index and the Flesch-Kincaid Reading Ease formula had significant correlations with the observed difficulty and could therefore be considered better predictors of test difficulty compared to other variables. The findings suggest the use of various indices to estimate the reading difficulty before administering tests to ensure the equivalency and validity of tests.

Keywords: Reading comprehension, Corpus linguistics, Reading difficulty, Readability, Computational linguistics

Introduction

Reading is a multifaceted cognitive ability that entails numerous subskills and processes, beginning with visual processes like decoding text and progressing to more complex levels involving syntax, semantics, and discourse analysis, and the ultimate stage of meaning-making through the reader's overall knowledge. In second language reading, this process can be even more intricate due to significant experiential, institutional, and sociocultural differences (Grabe & Stoller, 2002; Nassaji, 2011). For instance, word processing, as a critical element in fluent reading (Kintsch & Van Dijk, 1978; Perfetti, 2007), has been proved to be slower and less automatic in L2 speakers (Gollan et al., 2008; Izura & Ellis, 2004) and have greater variations compared to L1 word processing as a result of the linguistic differences and the amount of language exposure in each individual (Cop et al., 2015).

Selecting reading passages with appropriate levels of difficulty for teaching or assessment purposes has always been challenging for educators. Both quantitative and qualitative methods have been used to assess the difficulty of reading comprehension tests. Through qualitative methods such as think-aloud protocols and content analysis,

researchers have tried to explore linguistic features of texts, cognitive processes of test takers, and their strategies (Anderson et al., 1991; Bachman et al., 1988). Scholars also employed quantitative methods to examine difficulty level in reading comprehension by investigating the features of text, items, and the interaction between the two. Text features that influence item difficulty in reading comprehension include topic, vocabulary, syntax, number of words and sentences, rhetorical pattern, sentence length, and negation among others (Rupp et al., 2001).

Different quantitative methods for measuring text complexity and difficulty have been presented for native speakers and L2 language learners (e.g., Graesser et al., 2011; Xia et al., 2019). Formulas that can assess the readability and difficulty level of the texts have been assisting educators in selecting reading passages from among the exponentially vast number of available materials (Hiebert, 2002). Today in the fields of education, applications and tools of computational linguistic analyses are used to monitor learning experiences and also to assess educational texts (Dowell et al., 2016). In these studies, multiple regression analysis is normally used to determine how much various variables and indices account for the difficulty of reading texts.

When it comes to designing reading comprehension tests, the purpose of assessment is of great importance with the goal of maintaining the reliability and validity of the test and consequently reducing errors and increasing the accountability of the results (Foorman, 2009). In high-stakes exams where the effects of the test results are more significant, it is crucial to be able to predict the factors affecting the difficulty level of the test and test takers' approximate scores. Therefore, in this study, we aimed to identify the factors which had the strongest predicting ability of the difficulty of a high-stakes reading comprehension test.

Literature review

Traditional computer index of text ease/difficulty

In traditional readability formulas, text readability is measured on the basis of sentence and word length. Flesch reading ease (Flesch, 1948) and Flesch-Kincaid grade level (Kincaid et al., 1975) are the main traditional approaches to scaling texts that provide a single metric of text ease/difficulty. Longer sentences and words with complex syntax tend to be more challenging for the working memory. Therefore, this easy-to-compute approach has been known to be a good estimation of the reading time of a passage (Graesser et al., 2011). Among the studies done on the validity of this formula, Brown (1998) concluded that this formula is not a robust predictor of L2 reading difficulty, while Greenfield (1999) in his study proved otherwise. Overall, although the simplicity, unidimensionality, and practicality of the traditional approaches are appealing, they do not address deeper levels of discourse (Connor et al., 2007; McNamara & Magliano, 2009; Rapp et al., 2007).

Multilevel frameworks

Flesch-Kincaid Grade Level or Reading Ease is accepted by the educational community; however, deeper level analysis is surely needed to assess various levels of language-discourse. Over the years, researchers and scholars have identified and formed multiple levels of comprehension and developed frameworks accordingly (e.g., Graesser et al., 1997; Kintsch, 1998; McNamara and Magliano, 2009; Pickering and Garrod, 2004; Snow,

2002). Inspecting a large body of literature on reading comprehension, Graesser et al. (2011, p. 224) identified five recurrent levels proposed in the frameworks: "(1) words, (2) syntax, (3) the explicit textbase, (4) the situation model (sometimes called the mental model), and (5) the discourse genre and rhetorical structure (the type of discourse and its composition)." Knowledge of vocabulary and familiarity with word structure have a significant effect on the amount of time spent on reading a text and comprehending it (Perfetti, 2007; Rayner et al., 2001). Graesser and McNamara (2011) then divided the analysis of word characteristics into various levels such as analysis of parts of speech, word frequency, psychological ratings, and semantic content. Syntax is also among the factors affecting the difficulty level of a reading passage. By assigning parts of speech to words, grouping them into phrases, and assigning tree structures to sentences, syntax can analyze a sentence (Jurafsky & Martin, 2008). After wording and syntax, the textbase focuses on the meaning (Kintsch, 1998). Analysis of co-reference, lexical diversity, and latent semantic are among methods used to analyze a text. For instance, in the textbase, using co-reference, we connect propositions, clauses, and sentences that refer to the same person or thing (McNamara & Kintsch, 1996). The presence or lack of such cohesion (referential cohesion or referential cohesion gap) can affect the amount of time spent on reading and the level of difficulty in comprehending a text (O'Brien et al., 1998). Another level is the situation model, the deepest level of mental representation that surpasses the explicit textual meaning and the textbase. It refers to the mental representation that the comprehenders produce to describe the text at local and global levels (Tapiero, 2007).

The dimensions that Zwaan and Radvansky (1998) considered for the situation model include causation, intentionality or goals, time, space, and protagonists. Reading time and difficulty of a text increases when a break occurs in one or more of these dimensions. Finally, genre defines the category of text and decides whether it is narration, exposition, persuasion, or description, or their related subcategories (Biber, 1991). Different genres involve different levels of difficulty for their comprehension, for example, informational texts are more difficult to comprehend and recall than fiction (Graesser & McNamara, 2011).

Automated text analysis

With recent advancements in technology, text analysis has become automated which has helped to reduce the challenges of text selection and made it more practical for educators. Automated test analysis has been made possible as a result of the synthesis of the advances in disciplines and approaches including corpus linguistics, computational linguistics, psycholinguistics, discourse processing, and information retrieval (Graesser et al., 2004). Besides the importance of automated analyses of language in providing texts at the appropriate level for their students' learning, language assessment and high-stakes assessment, in particular, can highly benefit from this technology. Several indices that contribute to the ease/difficulty of a reading text can be measured using the latest automated text analysis tools, mostly available freely for researchers, teachers, and test developers.

The software for text readability measures dates back to 1963 when Danielson and Bryan developed a computer program for the readability formula and the

Farr-Jenkins-Paterson measure (Danielson & Bryan, 1963). Later, word-processing applications like Microsoft Word™ were created with the possibility of calculating measures such as Flesch-Kincaid. Today, tools such as Coh-Metrix are created that can measure text difficulty by focusing on different aspects of language and discourse. Moreover, natural language processing (NLP) tools such as TAALES and TAALED are currently available and can provide various measures of lexical diversity, syntactic complexity, text cohesion, grammar, and also sentiment.

Despite the importance of this topic in L2 reading, not enough attention has been paid to empirical studies in L2 contexts concerning the relationship between readability indices/formulas and the difficulty level of reading passages. A few studies have used these automated text analysis tools to examine the relationships between different text characteristics and reading comprehension scores (e.g., Crossley et al., 2008; Graesser et al., 2011; Hamada, 2015; Kim et al., 2018; Paribakht and Webb, 2016; Rupp et al., 2001). For instance, Crossley et al. (2007) examined three variables of the number of words per sentence, CELEX frequency, and argument overlap in 32 cloze reading passages and found that all three factors had significant correlations with the test takers' scores and also yielded a prediction of reading difficulty. In another study, Crossley et al. (2011) compared the readability formulas of the Coh-Metrix L2 Reading Index to the traditional readability formulas to identify which formula best categorizes the text levels. The results revealed that the Coh-Metrix L2 Reading Index was considerably more effective.

In another study, Nelson et al. (2012) assessed the effects of 7 text difficulty metrics on predicting text difficulty of both narrative and informational passages in five sets of texts. For narrative texts, the metrics with broader ranges of linguistics indices had a more significant relationship with the text difficulty. However, the metrics including the variables of sentence length and word difficulty had higher correlations for informational texts.

Hamada (2015) examined the lexical, syntactic, and meaning construction indices in the Japanese Eiken English graded test. Overall, the results indicated that surface-level linguistic variables such as lexical and syntactic indices better predicted the difficulty of reading comprehension than the higher-level linguistic variables including meaning construction indices.

Also, Choi and Moon (2020) studied the relationships between 26 text features and the test difficulty of high-stakes English as a Foreign Language (EFL) reading and listening tests. Moderate to high correlation was found with the observed difficulty of the test sections. Vocabulary features such as type and token as well as variation features, and syntactic features including the mean of clauses per sentence and readability features showed a significant correlation with the difficulty level. However, the correlation between the pragmatic features and the difficulty level was not significant.

The corpora chosen for the previous studies included simplified news texts (Crossley et al., 2011), Bormuth graded cloze passages (Chall & Dale, 1995; Crossley et al., 2007), standardized EFL tests such as TOEIC (Choi & Moon, 2020), and graded corpora such as the TASA corpus or Japanese Eiken English graded tests (Graesser et al., 2011; Hamada, 2015) among others. In this study, however, we tried to investigate a rather different corpus, the reading comprehension subsection of a high-stakes test which is used for university admission purposes. Despite the importance of high-stakes exams, there is still a paucity of research on the linguistic features that affect the difficulty of such tests,

thereby having great consequences on test takers' lives. As such, the main purpose of the study was to examine which linguistic features are related to the reading comprehension test difficulty of this high-stakes test. After a thorough review of the relevant literature, a total number of 14 factors and two readability formulas were included in the study. The following research question was accordingly put forward:

How well do textual measures, Coh-Metrix L2 reading indices, lexical diversity measures, and readability formulae predict reading test difficulty in INUEE?

Method

Data

Corpus

The reading comprehension subsection of a major high-stakes large-scale test, the Iranian National University Entrance Exam (INUEE) for the master's program, was selected for this study. This exam is held annually for the master's program admission purposes and it contains multiple-choice items on various subject matters including English language proficiency specifically designed for each field of study. Students with higher scores can enter different universities and pursue their master's degree. The reason for the selection of this test is that it is one of the most influencing and important exams in the context of Iran as its results directly affect individuals' life prospects both socially and financially.

The English language proficiency section includes items on vocabulary, grammar, and reading comprehension. For this research, we chose our exams based on the availability of data. The Iranian National Organization for Educational Testing granted access to the data of seven fields of studies namely physics, mechanics, Persian literature, English language related fields, agricultural management, food hygiene and quality control, and urban planning. The reading comprehension subsection in each test version includes three reading passages. We examined the reading passages in three years, 2017–2019, therefore, the overall number of reading passages examined was 63. Each reading passage is followed by 5 multiple choice items on getting the main idea, finding the supporting details, and dealing with vocabulary. The mean number of words in the passages was 275.43, with a minimum number of 141 and a maximum of 480 words (SD: 71.04).

Test takers

The Iranian National Organization for Educational Testing provided us with the information of a total of 17,900 test takers for the mentioned seven fields in three years. The data included 6500 responses for the year 2017, 5900 for 2018, and 5500 for 2019. The participants were female and male non-native speakers of English who ranged in age from 22 to 62. While there were no specific data about their proficiency level, their overall scores on the English language proficiency exam (ranging from 0.00 to 51 out of 60 for English majors and 0.00 to 27 out of 30 for non-English majors) could show that they belonged to different proficiency levels. A lot of test takers choose not to answer some part of the exam and prefer to leave the parts that they are not good at empty since this is a timed exam with a negative scoring system. Therefore, from among the available data of the test takers, we eliminated the ones whose overall English language

proficiency scores were lower than one standard deviation below the mean. The total number of the data of the test takers was then reduced to 10,386.

Variable selection

The variables for this study were selected from the previous research done on the readability and ease/difficulty of reading texts. Graesser et al. (2011) used the principal component analysis for 53 Coh-Metrix measures of text characteristics. They identified five major factors that systematically accounted for the variability and difficulty among texts. We included these five factors, namely, narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep (causal) cohesion in this study.

Crossley et al. (2008) examined and validated three indices of content word overlap, semantic similarity, and CELEX frequency scores which are accountable for reading difficulty based on many psycholinguistic studies. These three indices are the components of the Coh-Metrix L2 Reading Index formula.

We also took into account the three dimensions of lexical diversity in the current study. We selected the factors validated by Kyle et al. (2021) including volume, abundance, and variety. For each of the volume and abundance dimensions, one index was selected. Four indices of variety, relatively independent of text length, including MATTR, an instantiation of D, and two versions of MTLD were selected.

Finally, we measured the traditional readability formula of Flesch-Kincaid Reading Ease and also the Coh-Metrix L2 Reading Index to see if there is any relationship between these factors and the test difficulty and to further compare the two readability formulas. Overall, a total of 16 corpus features were measured using two corpus analysis tools (Table 1).

Data analysis

To measure the indices of narrativity, syntactic simplicity, word concreteness, referential cohesion, deep (causal) cohesion, CELEX frequency scores, semantic similarity, content word overlap, Flesch-Kincaid Reading Ease, and the Coh-Metrix L2 Reading Index for each reading passage, Coh-Metrix 3.0 web tool (Graesser et al., 2004; McNamara et al., 2002) was employed. For the indices of lexical diversity in the reading passages TAALES 2.0 tool (Kyle et al., 2018; Kyle & Crossley, 2015) was used.

When there is a sound theoretical or conceptual rationale, the relationship between a continuous dependent variable and a number of predictors or variables could be checked through multiple regression (Pallant, 2020; Stevens, 2009; Tabachnick & Fidell, 2019). As discussed before, the variables for this study were all selected from the previous research conducted on the readability and difficulty of reading, and therefore seemed suitable for regression analysis. In multiple regression, a minimum of 10 data cases (with conservative models using 15 to 20) for each predictor is considered accurate (Crossley et al., 2008). In this study, multiple regressions using SPSS 26 were conducted separately for the different sets of predictors and the observed test difficulty values. Observed test difficulty is the mean difficulty of the items in each test. Item difficulty value was calculated by dividing the number of test takers who got the item wrong by the total number of test takers. Before using multiple regression, assumptions for using this statistical technique including normality, linearity, homoscedasticity, and independence of residuals (Pallant,

Table 1 Variables used in this study

Indices/variables	Definition
1. Narrativity	(Genre and rhetorical structure level) How much a text tells a story or presents characters, actions and procedures
2. Syntactic simplicity	(Syntax level) A subsection of syntax level that assesses a text on the basis of the number of words, their simplicity, and sentence syntactic structure
3. Word concreteness	(Word level) The level of meaningfulness of the content words and evoking of mental images
4. Referential cohesion	(Text base level) The degree of the connectedness of content words and ideas as the text unfolds
5. Deep cohesion	(Situation model) The extent to which clauses and sentences in a text are connected to causal and intentional or goal-oriented connectives
6. Content word overlap	(Word level) The measure of the content word overlap between two adjacent sentences
7. Semantic similarity	(Textbase level) The uniformity of parallel syntactic constructions at the phrase level and also parts of speech
8. CELEX frequency scores	(Word level) The frequency of the words in the CELEX (Baayen et al., 1996) from the early 1991 version of the COBUILD corpus
9. Volume	(Word level) The number of words in a text
10. Abundance	(Word level) The total number of different types (lemmas) in a text
11. HD-D	(Word level) The "probability that a word in a text would be included in a random sample from that text" (Kyle et al., 2021, p.7).
12. MATTR	(Word level) MATTR (Moving average type-token ratio) is the average of "type-token ratios across multiple, overlapping, equal sections in the text" (Kyle et al., 2021, p.8).
13. MTLD-original	(Word level) MTLD-original (The measure of textual lexical diversity) represents the mean number of words to reach a point of type-token ratio stabilization
14. MTLD-w	(Word level) MLTD-w "moving average is a variant of MTLD that uses a moving average approach" (Kyle et al., 2021, p.8)
15. Flesch-Kincaid Reading Ease	A traditional readability formula that measures text readability according to sentence and word length
16. Coh-Metrix L2 Reading Index	This formula consists of three variables of a word overlap index, a word frequency index, and an index of syntactic similarity to examine the readability of a text

2020; Stevens, 2009; Tabachnick & Fidell, 2019) were checked to make sure the use of this technique was justified.

Results

As stated before, to find out how the selected independent variables could collectively predict the difficulty of the reading comprehension tests in the entrance exam for the EFL test takers, the observed difficulty and the independent variables were explored using multiple regression. First, a multiple regression analysis was estimated for the five indices of narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep (causal) cohesion. Table 2 shows the descriptive statistics for the dependent and independent variables and Table 3 presents the results of the regression analysis.

As can be seen from Table 2, the observed difficulty of the test seemed to be high with the mean value of 0.77, which means that the test was seemingly difficult for this population. The difference between the minimum and maximum test difficulty (0.50 & 0.92) showed a wide range of difficulty among the reading passages. Word concreteness and deep cohesion had an overall positive mean (0.31 & 0.29 respectively),

Table 2 Summary of the descriptive statistics

Variables	Min.	Max.	Mean	SD	N
Observed difficulty	0.50	0.92	0.77	0.11	63
Narrativity	− 2.57	1.47	− 1.12	0.72	63
Syntactic simplicity	− 2.44	1.00	− 0.66	0.83	63
Word concreteness	− 1.79	2.37	0.31	1.02	63
Referential cohesion	− 2.33	2.59	− 0.09	0.96	63
Deep cohesion	− 1.63	4.37	0.29	1.10	63

Table 3 Regression analysis for predicting variables of EFL reading difficulty

Variables	β	R	t	p	R	R^2	ΔF	Sig.
	–	–	22.472	0.000	0.267	0.071	0.873	0.505
Narrativity	− 0.013	− 0.036	− 0.101	0.920				
Syntactic simplicity	− 0.159	− 0.110	− 1.102	0.275				
Word concreteness	0.085	0.074	0.654	0.516				
Referential cohesion	− 0.240	− 0.190	− 1.789	0.079				
Deep cohesion	− 0.028	− 0.049	− 0.199	0.843				

*β: Standardized coefficients

Table 4 Summary of the descriptive statistics

Variables	Min.	Max.	Mean	SD	N
Content word overlap	0.03	0.33	0.11	0.05	63
Sentence syntax similarity	0.04	0.14	0.08	0.02	63
CELEX frequency scores	2.58	3.15	2.87	0.13	63

however, narrativity, syntactic simplicity, and referential cohesion showed a negative average (− 1.12, − 0.66, & − 0.09 respectively).

According to the results of the multiple regression analysis, the combination of the five factors together produces a multiple correlation of 0.267 and a corresponding adjusted R2 of 0.071 this means that all five variables account for 7.1% of the variance in the observed difficulty of the 63 passages. Therefore, this model can only predict 7.1% of the difficulty of the reading passages. The Table also indicates that none of the variables has a significant contribution to the observed difficulty ($p > 0.05$).

Next, the indices of content word overlap, semantic similarity, and CELEX frequency scores were studied concerning the observed difficulty of the reading passages. Table 4 shows the results of the descriptive statistics of the three variables. The mean value for content word overlaps between the adjacent sentences is 0.11 while the syntactic similarity indicated the mean value of 0.08.

Regression analysis was then conducted with the three features entered in the model; the results appear in Table 5. All three variables together account for only 12% of the observed difficulty of the tests ($F = 2.672$, $p \leq .05$). Concerning the correlation between each variable and the observed difficulty, only the content word overlap

Table 5 Regression analysis for predicting variables of EFL reading difficulty

Variables	β	*r*	*t*	*p*	R	R^2	ΔF	Sig.
(constant)	–	–	2.67	0.010	0.346	0.120	2.672	0.06
Content word overlap	−0.333	−0.344	−2.61	0.011				
Sentence syntax similarity	−0.040	−0.127	−0.31	0.756				
CELEX frequency scores	−0.003	−0.024	−0.02	0.981				

index is found to be significant ($p < .05$). This indicates that syntax similarity and CELEX frequency did not have a significant relationship with the observed difficulty.

The indices of lexical diversity were the next corpus features selected to be analyzed in the study. The indices included volume, abundance, HD-D, MATTR, MTLD-original, and MTLD-w (Table 6). According to the table, the number of words in each passage manifested a large difference in the size of the passages, ranging from 141 words to 480 (Mean = 275.43). The number of lemmas in the text (abundance) seemed to vary greatly in the reading passages.

The indices were then entered in the regression analysis and the results appear in Table 7. The findings indicated that the six variables accounted for 8.2% of the variance in the 63 reading passages which is not significant ($F = 0.831$, $p > .05$). The statistics of each of the variables also show no significant correlations with the observed difficulty of the tests.

Finally, we compared the two readability formulas of the Coh-Metrix L2 Reading Index and the traditional Flesch-Kincaid Reading Ease. The descriptive statistics are presented in Table 8. There are great differences between the minimum and maximum values of the formulas in the reading passages.

Table 6 Summary of the descriptive statistics

Variables	Min.	Max.	Mean	SD	N
Volume	141	480	275.43	71.05	63
Abundance	91	466	177.22	75.58	63
HD-D	0.64	0.85	0.77	0.04	63
MATTR	0.72	0.88	0.80	0.03	63
MTLD-original	35.52	129.39	70.45	20.44	63
MTLD-w	37.60	133.01	72.32	22.16	63

Table 7 Regression analysis of the lexical diversity indices predicting EFL reading difficulty

Variables	B	*r*	*t*	*p*	R	R^2	ΔF	Sig.
(Constant)	–	–	0.199	0.843	0.286	0.082	0.831	0.551
Volume	0.169	0.188	0.981	0.331				
Abundance	−0.015	0.104	−0.092	0.927				
HD-D	0.170	0.234	0.553	0.582				
MATTR	0.068	0.238	0.185	0.854				
MTLD-original	0.138	0.220	0.254	0.800				
MTLD-w	−0.162	0.215	−0.309	0.758				

Table 8 Summary of the descriptive statistics

Variables	Min.	Max.	Mean	SD	N
Flesch-Kincaid Reading Ease	1.01	70.36	35.44	17.28	63
The Coh-Metrix L2 Reading Index	−0.11	21.91	0.11	0.05	63

Table 9 Regression analysis of the readability measures predicted scores

Variables	β	*r*	*t*	*p*	R	R^2	ΔF	Sig.
Flesch-Kincaid Reading Ease	−0.328	−0.319	−2.522	0.011	0.320	0.103	3.43	0.039
The Coh-Metrix L2 Reading Index	0.025	−0.086	0.196	0.846				

We then performed a multiple regression for the two formulas and the observed difficulty. The results are demonstrated in Table 9.

It is clear from the results that the two formulas explain for the overall 10.3% of the variance in the passages ($R^2 = 0.103$). Comparing the two reading formulas, we found that Flesch Reading Ease, with the beta value of −0.328 (regardless of the negative sign), makes a stronger prediction of the reading difficulty than the Coh-Metrix L2 Reading Index ($\beta = 0.025$). Furthermore, the contribution made by the Flesch Reading Ease ($r = .319$, $p < .05$) unlike the Coh-Metrix L2 Reading Index ($r = 0.086$, $p > 0.05$) is significant.

Discussion

Multiple indices and two readability formulas were studied in 63 reading passages in a national university entrance exam and their relationships with the test difficulty were investigated. First, we addressed the five factors of narrativity, syntactic simplicity, word concreteness, referential cohesion, deep (causal) cohesion (Tables 2 and 3). The low mean value of narrativity (−1.12) in the reading passages was in line with the contents being tested in an academic context since informational texts include less degree of narrativity (Biber, 1991; Graesser et al., 2011). Also, the syntactic simplicity (in this case −0.66) in informational texts is lower to compensate for the challenging subject matters. According to the results of the multiple regression analysis, the five factors can only predict 7.1% of the difficulty of the reading passages. The results are not in line with the findings of Graesser et al. (2011). The authors found the five levels responsible for 67.3% of the variance in 37,520 texts. In their study, they found significant correlations between each variable and the grade leveled texts. Contrary to their study, we found no significant correlation between these factors and the observed difficulty levels. However, the studies are not very similar in the construct. They used the TASA corpus which included leveled texts with an associated Degrees of Reading Power (DRP) score of text difficulty while in the current study we examined the observed difficulty of the reading passages of a high-stakes university entrance exam answered by L2 test takers with various levels of proficiency. Therefore, the findings should be compared with caution.

We then explored the three Coh-Metrix indices of content word overlap, semantic similarity, and CELEX frequency scores. Together, they were responsible for 12% of the observed difficulty, with only the content word overlap variable being significant. The

Coh-Metrix index content word overlap is one of the influencing factors that can help readers with meaning constructions. The more the vocabulary overlap between two adjacent sentences, the easier the comprehension is (Douglas, 1981; Rashotte, 1983). However, these findings differ from the results of the study by Crossley et al. (2008). In their study, the combination of these three indices accounts for 86% of the variance in the cloze test scores of the Japanese students while in the current study this percentage is much lower (12%). Moreover, in their study, all the three variables significantly correlated with the observed difficulty of the tests. The main difference between the two studies is the type of reading comprehension test. Crossley et al. (2008) used Bormuth (1971) corpus of 32 passages which were validated in previous similar studies and then they collected scores according to fifth-word deletion cloze tests.

We also wanted to find out whether lexical diversity indices were predictors of the test difficulty, therefore, we ran multiple regression analysis for the variables of volume and abundance and four indices of variety including HD-D, MATTR, MTLD-original, and MTLD-w. Together, they showed to account for 8.2% of the variance but none of them indicated to have a significant correlation with the observed scores.

Finally, the two readability formulas of Flesch-Kincaid Reading Ease and Coh-Metrix L2 Reading Index were measured and compared. The results were in line with the findings of Greenfield (1999), as the Flesch-Kincaid Reading Ease formula was found to be a strong predictor of the test-takers' scores. Also, Nelson et al. (2012) confirmed that the traditional components of readability formulas, including sentence length and word difficulty, are more relevant for text difficulty assessment for reading comprehension tests with informational texts compared to the narrative genre. However, our findings are in contrast with the results of the study conducted by Crossley et al. (2011). They compared the traditional formulas of Flesch-Kincaid Grade Level and Flesch Reading Ease to the Coh-Metrix L2 Reading Index to find out which formula best classifies the level of simplified L2 reading texts. The Coh-Metrix L2 Reading Index showed significantly higher discriminating power between different levels of the texts. The main difference between their study and the current one is that the texts examined in their research were non-academic in nature.

Overall, among the 16 variables investigated in this study, the scores predicted by the factor of content word overlap and the Flesch-Kincaid Reading Ease formula had significant correlations with the observed test difficulty of the reading passages. We ran multiple regression for each of these two indices and found that content word overlap accounted for 11.8% and Flesch-Kincaid Reading Ease for 10.2% of the variance in the observed difficulty.

One reason for the outperformance of Flesch-Kincaid Reading Ease to the Coh-Metrix L2 Reading Index might be the fact that traditional readability formulas work perfectly for strictly academic genres and informational texts (Greenfield, 1999; Nelson et al., 2012). The reading passages selected for the university entrance exam were most often extracts of the academic and informational passages related to each field of study.

Another explanation can be the type of exam. The entrance exam is a timed multiple-choice test in which test takers have to answer multiple sections including the English proficiency section. Therefore, speed is an important factor in such tests which can extensively affect their performance. Moreover, as this test has a negative scoring

system, test takers usually tend to skip the parts they are not good at or the parts that require more time to answer. The Flesch-Kincaid Reading Ease is computed using the length of words and sentences, and in fact, it is "a robust predictor of the amount of time it takes to read a passage" (Graesser et al., 2011, p. 224). As long sentences impose more challenges on working memory, and normally they are syntactically more complex, reading comprehension gets more difficult especially with time limitations. This can also be a justification for the significance of the content word overlap index on the observed difficulty of the tests. The significance of the variable of content word overlap is that it corresponds to meaning construction in reading comprehension. It is an important factor that can highly impact text comprehension as well as reading speed (Douglas, 1981; Rashotte & Torgesen, 1985). Lack of enough time inhibits test takers from reading each passage more than once and the content word overlap in adjacent sentences can greatly assist them in comprehending a text and answering the related items. Also, informational texts include more unfamiliar words and require more knowledge (Graesser & McNamara, 2011). The content word overlap can help readers guess the meanings of unfamiliar words and therefore have a better comprehension.

Conclusion and implications

One of the first and the most important concerns of test developers is to develop tests with appropriate levels of difficulty and discrimination for the target test takers. The purpose of this study was, therefore, to examine if certain variables are associated with the difficulty level of reading tests and test takers' scores. From among the variables existing based on a priori assumptions from the previous studies, 16 indices were selected to be investigated on the reading comprehension section of the university entrance exams for the master's program. According to the results, the content word overlap and the traditional readability formula of Flesch-Kincaid Reading Ease were stronger and more significant predictors of the test takers' scores compared to the other variables. The rather different observations of the current study are indicative of how unique each context is and how different factors can contribute to the difficulty of tests.

The robustness of tests can partly be achieved by predicting item difficulty using both objective and subjective methods. Corpus analysis, using automated text analysis tools, can complement the objective methods relying on expert judgments. Assessing the difficulty of tests before administrating them is a crucial step which is now more possible and feasible with the recent advances in language assessment approaches. Studies such as the current one can identify various predicting factors that influence item difficulty in reading comprehension tests. As Bailin and Grafstein (2001, p. 292) proposed, "there is no single, simple measure of readability". Therefore, the procedures should differ depending on the characteristics of each test. For instance, in the case of our study, the stress of a high-stakes, time-limited one-off academic reading assessment which has important consequences, makes the test quite different from other types of tests.

The distribution of test takers' scores is highly influenced by the difficulty of the test. On that account, for such high-stakes university entrance exams maximizing consistency and maintaining a similar difficulty level in different test forms is critical (Kolen & Brennan, 2014). Assessment of the difficulty of reading comprehension can help with a more reliable and accurate screening of candidates seeking entrance to universities.

Test designers and materials developers should consider word overlap in texts as it is directly linked to meaning construction. It measures lexical connections which can help to build patterns of meaning (Crossley et al., 2008).

Another implication is that although new advanced formulas and indices have been proved to outperform the traditional readability formulas in predicting reading test difficulty in general, their selection should be made with caution. The utility of traditional formulas to predict difficulty for informational texts suggests that the type of text is an important criterion in selecting appropriate text difficulty indices and tools.

Limitations and further research

This study was not without limitations. The first limitation is related to the small sample size used for regression analysis. The present study included a sample of 63 reading passages. More dependable results could have been obtained if more passages had been included. Further studies should be conducted with larger corpora including a greater number of reading passages.

The difficulty level of a reading comprehension test may also be influenced by factors such as the type of items included, topic, genres, and communicative functions among others. Future studies should consider such factors for a more in-depth analysis.

Author contributions
All authors contributed to the study conception, material preparation, data collection and analysis. The manuscript was written and reviewed by both authors and they read and approved the final manuscript.

Funding
The authors received no financial support for the research, authorship, or publication of this article.

Availability of data and materials
All data generated or analyzed during this study are included in this published article.

Declarations

Ethics approval and consent to participate
Not Applicable.

Competing interests
The authors declare no competing interests.

Received: 14 October 2022 Accepted: 24 August 2023
Published online: 08 December 2023

References

Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8*(1), 41–66.
Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). The CELEX lexical database (cd-rom). *Linguistic Data Consortium*.
Bachman, L. F., Swathi Vanniaraian, K., & Lynch, B. (1988). Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries. *Language Testing, 5*(2), 128–159.
Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language & Communication, 21*(3), 285–301.
Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.
Bormuth, J. R. (1971). *Development of standards of readability: Toward a rational criterion of passage performance*. Bureau of Research.
Brown, J. D. (1998). An EFL readability index. *JALT, 20*(2), 7–36.
Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.
Choi, I. C., & Moon, Y. (2020). Predicting the difficulty of EFL tests based on corpus linguistic features and expert judgment. *Language Assessment Quarterly, 17*(1), 18–42.
Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007). Algorithm-guided individualized reading instruction. *Science, 315*(5811), 464–465.

Cop, U., Keuleers, E., Drieghe, D., & Duyck, W. (2015). Frequency effects in monolingual and bilingual natural reading. *Psychonomic Bulletin & Review, 22*(5), 1216–1234.
Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language, 23*(1), 84–101.
Crossley, S. A., Dufty, D. F., McCarthy, P. M., & McNamara, D. S. (2007). Toward a new readability: A mixed model approach. In *Proceedings of the annual meeting of the cognitive science society*.
Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 42*(3), 475–493.
Danielson, W. A., & Bryan, S. D. (1963). Computer automation of two readability formulas. *Journalism Quarterly, 40*(2), 201–206.
Douglas, D. (1981). An exploratory study of bilingual reading proficiency. In S. Hudelson (Ed.), *Learning to read in different languages. Linguistics and literacy series: 1. Papers in applied linguistics* (pp. 33–102). Center for Applied Linguistics.
Dowell, N. M., Graesser, A. C., & Cai, Z. (2016). Language and discourse analysis with Coh-Metrix: Applications from educational material to learning environments at scale. *Journal of Learning Analytics*, *3*(3), 72–95.
Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233.
Foorman, B. R. (2009). Text difficulty in reading assessment. In E. H. Hiebert (Ed.), *Reading more, reading better* (pp. 231–250). Guilford Press.
Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language*, *58*(3), 787–814.
Grabe, W., & Stoller, F. L. (2002). *Teaching and researching*. Allyn & Bacon.
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., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods Instruments & Computers*, *36*(2), 193–202.
Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223–234.
Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, *48*(1), 163–189.
Greenfield, G. R. (1999). *Classic readability formulas in an EFL context: Are they valid for Japanese speakers?* Temple University Press.
Hamada, A. (2015). Linguistic variables determining the difficulty of Eiken reading passages. *JLTA Journal*, *18*, 57–77.
Hiebert, E. H. (2002). Standards, assessment, and text difficulty. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (3rd ed., pp. 337–369). International Reading Association.
Izura, C., & Ellis, A. W. (2004). Age of acquisition effects in translation judgement tasks. *Journal of Memory and Language*, *50*(2), 165–181.
Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing (prentice hall series in artificial intelligence)*. Prentice Hall.
Kim, M., Crossley, S. A., & Skalicky, S. (2018). Effects of lexical features, textual properties, and individual differences on word processing times during second language reading comprehension. *Reading and Writing*, *31*(5), 1155–1180.
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*. Naval Technical Training Command Millington TN Research Branch.
Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge university press.
Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394.
Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.
Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046.
Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757–786.
Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, *18*(2), 154–170.
McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*(3), 247–288.
McNamara, D. S., & Magliano, J. P. (2009). Self-explanation and metacognition: The dynamics of reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of Metacognition in Education* (pp. 60–81). Routledge.
McNamara, D. S., Louwerse, M. M., & Graesser, A. C. (2002). *Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension*. Technical report, Institute for Intelligent Systems, University of Memphis, Memphis, TN.
Nassaji, H. (2011). Issues in second-language reading: Implications for acquisition and instruction. *Reading Research Quarterly*, *46*(2), 173–184.
Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. Council of Chief State School Officers.
O'Brien, E. J., Rizzella, M. L., Albrecht, J. E., & Halleran, J. G. (1998). Updating a situation model: A memory-based text processing view. *Journal of Experimental Psychology: Learning Memory and Cognition*, *24*(5), 1200–1210.
Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (7th ed.). Taylor & Francis Group.
Paribakht, T. S., & Webb, S. (2016). The relationship between academic vocabulary coverage and scores on a standardized English proficiency test. *Journal of English for Academic Purposes*, *21*, 121–132.
Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383.
Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190.
Rapp, D. N., Broek, P., McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading, 11*(4), 289–312.

Rashotte, C. A. (1983). *Repeated reading and reading fluency in learning disabled children*. The Florida State University.
Rashotte, C. A., & Torgesen, J. K. (1985). Repeated reading and reading fluency in learning disabled children. *Reading Research Quarterly, 20*(2), 180–188.
Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2*(2), 31–74.
Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing*, *1*(3–4), 185–216.
Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Rand Corporation.
Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). Routledge.
Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson Education Inc.
Tapiero, I. (2007). *Situation models and levels of coherence: Toward a definition of comprehension*. Taylor & Francis.
Xia, M., Kochmar, E., & Briscoe, T. (2019). Text readability assessment for second language learners. Preprint retrieved from https://arxiv.org/abs/1906.07580.
Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*(2), 162–185.

Publisher's Note