Routledge
Taylor & Francis Group

Language Assessment Quarterly

ISSN: 1543-4303 (Print) 1543-4311 (Online) Journal homepage: www.tandfonline.com/journals/hlaq20

# Predicting the Difficulty of EFL Tests Based on Corpus Linguistic Features and Expert Judgment

Inn-Chull Choi & Youngsun Moon

LANGUAGE ASSESSMENT QUARTERLY
2020, VOL. 17, NO. 1, 18–42
https://doi.org/10.1080/15434303.2019.1674315

# Predicting the Difficulty of EFL Tests Based on Corpus Linguistic Features and Expert Judgment

Inn-Chull Choi and Youngsun Moon

Korea University, Seoul, Korea (the Republic of)

**ABSTRACT**

This study examines the relationships among various major factors that may affect the difficulty level of language tests in an attempt to enhance the robustness of item difficulty estimation, which constitutes a crucial factor ensuring the equivalency of high-stakes tests. The observed difficulties of the reading and listening sections of two EFL tests were compared using corpus linguistic features and expert judgments, i.e., native and nonnative speakers' perceived difficulty of the test items. The research findings are as follows: Some corpus features and the predicted difficulties demonstrated a moderate to high correlation with the test sections' observed difficulty. The native and nonnative speakers' predicted difficulties significantly explained the observed difficulty of the test sections, where the nonnative speakers' predicted difficulty explained a similar variance. When entered separately, the corpus features showed a stronger explanatory power than the predicted difficulties. The corpus features and predicted difficulty together accounted for the largest variance, which was more than half of the variance of the test sections. The current study suggests that corpus features and expert judgment capture different aspects of item difficulty and future research in this area needs to consider how these two can be combined for robust item difficulty estimation.

## Introduction

In the process of developing tests to perform various functions (diagnostic assessment, formative evaluation, summative evaluation, etc.), every effort is made for the test to meet the degree of difficulty and discriminability appropriate for the characteristics of the target test-taker group, especially proficiency level. It is important to either conduct a simulated test or measure the difficulty of the test before it is actually administered so as to tailor it to the appropriate level of difficulty and discriminability. However, under ordinary circumstances in EFL environments, it may be impossible to administer a trial test. For instance, in developing the Korean version of the Scholastic Aptitude Test (referred to as College Scholastic Ability Test, or CSAT for short), a process of predicting the difficulty of test items in advance cannot be conducted due to complicated security issues. In order to develop valid and reliable high-stakes tests, it is deemed imperative that test developers devise a reliable and valid method of predicting the difficulty of a test prior to the initial implementation.

In particular, when developing a high-stakes test of criterion-referenced (C-R) nature, the process of predicting item difficulty is essential to ensure the robustness of difficulty, which serves as a crucial factor for the equivalency of C-R tests. This equivalency is a prerequisite for the complicated standard-setting process, in which a cut-off score is to be established. Methods of establishing a cut-off score include the Angoff method (Angoff, 1971), the Bookmark method (Lewis, Mitzel, &

**CONTACT** Inn-Chull Choi ✉ icchoi@korea.ac.kr Department of English Language Education, Korea University, 1 Anam-dong, Sungbook-gu, Seoul 136-701, Korea (the Republic of)

Green, 1996), and the use of score intervals such as 90, 80, 70, and so on as the cut-off score. The Angoff and Bookmark methods both require expert judgment of item difficulty, whereas a cut-off score based on fixed scores may be set according to a specific set of goals or based on the level of achievement. Although the Angoff and Bookmark methods of setting score criteria are ideal, for tests where test equating is virtually impossible, selecting a fixed score as a criterion may be the most feasible option available.

Whichever of the aforementioned methods may be employed for the standard setting process, maintaining a consistent difficulty is imperative, as the item difficulty may greatly influence the distribution of test-taker scores. This is all the more true for tests with cut-off scores set as inflexible fixed scores, whereas the cut-off scores of the Angoff and Bookmark methods are adjusted based on expert judgments of item difficulty, which may reflect the item difficulty in setting the cut-off scores. That is, in order to maintain a consistent scoring method for a test, maintaining a robust difficulty level is a prerequisite, especially for fixed cut-off scores.

This robustness, however, is not only critical for C-R tests, but is also essential even for non C-R high-stakes standardized tests, whose results go through sophisticated processes of test equating among different test sets administered on a regular basis. In order to maximize consistency in interpreting scores of high-stakes tests, it is imperative that the test forms used for these different test administrations maintain equivalent or comparable difficulty in terms of test input (Dorans, Moses, & Eignor, 2010; Kolen & Brennan, 2014). In other words, even non C-R high-stakes tests are expected to maintain robust difficulty to maximize overall reliability and validity. This goal of achieving robustness of difficulty can be facilitated by predicting item difficulty based on methods including an objective method involving corpus analysis and a subjective method that relies on expert judgments. Nonetheless, not much research has been carried out regarding this issue, although some have indicated various factors that influence item difficulty (Choi, 1994; Freedle & Kostin, 1993a; Rupp, Garcia, & Jamieson, 2001).

Therefore, this study attempts to explore plausible ways of identifying significant factors to maximize the robustness of the item difficulty estimation by utilizing measurements of content difficulty through corpus analysis in coordination with expert judgment. Various corpus features related to vocabulary, syntax, and pragmatics as well as expert judgments of both native and nonnative speakers will be used to investigate the extent to which these variables are related to the observed difficulty and identify the variables that can predict the difficulty of a test in a fairly accurate and reliable manner.

## Literature review

### *Corpus analysis*

#### *Corpus analysis method*

Noteworthy research conducted to explore ways to predict difficulty includes a study focusing on the well-defined constructs of reading (Freedle & Kostin, 1993a) and other studies utilizing sophisticated statistical analyses (Perkins, Gupta, & Tammana, 1995; Rupp et al., 2001). Another frequently adapted method by previous studies is using corpus analysis tools to investigate the test's linguistic characteristics.

Difficulty prediction based on objective features is possible with the use of corpus analysis. For each test item, features related to vocabulary, syntax, discourse, and pragmatics that may affect the difficulty of the test can be used as objective criteria for difficulty prediction. In doing so, various tools such as Coh-metrix, LCA, L2SCA, and AntWordProfiler can be used. Coh-metrix yields over 100 features covering vocabulary, syntactic, discourse, and phonological aspects (Graesser, McNamara, Louwerse, & Cai, 2004). In addition, the Lexical Complexity Analyzer (LCA) and L2 Syntactic Complexity Analyzer (L2SCA) are corpus tools for analyzing lexical and syntactic features of passages, respectively (Ai & Lu, 2010; Lu, 2010). Finally, AntWordProfiler (AWP) is an analytical tool that allows a vocabulary list to be set as a criterion for comparing the words in a selected list and the words within the passage to be analyzed, so that the composition or percentage of the vocabulary of the list within the passage can be obtained (Anthony, 2014). For example, AWP can provide the percentage of words from the Academic Word List contained in a passage to determine its academic characteristics.

On the other hand, some studies taken a critical view of corpus analysis (Crossley, Skalicky, Dascalu, McNamara, & Kyle, 2017; Sheehan, Kostin, Futagi, & Flor, 2010). Sheehan et al. (2010) stated that there are a few limitations to automated passage analysis tools, such as inadequate construct coverage, an overly narrow criterion for features, and an inadequate consideration of genre influences. In response, Sheehan et al. (2010) has attempted to develop a more accurate text analysis system to address the limitations of insufficient corpus analysis tools. Also, Crossley, Greenfield, and McNamara (2008) proposed a new text analysis method based on vocabulary, syntax, and content overlap features. On the other hand, a study that evaluated six text analysis tools and examined their explanatory power for the difficulty level of texts showed that all analytical tools were significantly correlated with the grade level of the passage and the students' actual test results (Nelson, Perfetti, Liben, & Liben, 2012). That study demonstrated that, although there are limitations to corpus analysis tools, as mentioned earlier, they can be useful in determining the characteristics of a text. Studies also showed that corpus analysis results can be used to predict item difficulty (Freedle & Kostin, 1993b; Nelson et al., 2012), indicating the possibility of utilizing the results of objective evaluation of passages obtained from corpus analysis to predict the difficulty of tests.

### Corpus features

When corpus analysis is conducted, numerous features that show a specific characteristic of the analyzed text are provided, and it is up to the researcher to choose the features relevant and important to research. The features utilized in the body of studies of difficulty prediction can be divided into vocabulary, syntax, discourse, pragmatic, and item factors. First, the corpus features related to vocabulary are among the most influential features in predicting item difficulty (Crossley et al., 2008). They include features that indicate the length of the texts such as the number of types, the number of tokens, the number of content words, and the number of sophisticated vocabulary items, and the features related to vocabulary difficulty include vocabulary frequency, vocabulary variation, lexical sophistication, and lexical density. Since students' comprehension ability decreases as the number of unknown words in the reading and listening process increases (Nation, 2013), vocabulary-related features are expected to have a direct impact on the difficulty of a test item.

Nonetheless, even if the meaning of the vocabulary is known, a passage can be comprehended only when the learner goes through the process of reading or listening to a sentence in accordance with relevant grammatical knowledge and understanding the meaning of the word within the whole sentence (Grabe, 2009). This shows that not only vocabulary but also syntactic features are likely to play an important role in research on predicted difficulty, as they also play an imperative role in determining the success and failure of comprehension. Syntactic features include syntactic complexity, readability values, the number of sentences/clauses/t-units, and the average length of sentences/clauses/t-units.

Corpus features related to discourse properties of the passage include narrativity, referential cohesion, deep cohesion, verb cohesion, connectivity, and temporality. These features are based on several levels of language and discourse and provide insight into text difficulties besides lexical and syntactic characteristics. In addition, features related to pragmatics may include the degree of concreteness or abstractness of the passage, its academic/non-academic character, and the academic/practical character of a passage, which are also important properties of the passage that can influence the difficulty of the item. Lastly, there are features specifically relevant to test items such as length or position of choices, item type (factual, inferential, main idea, flow, etc.), and variables for item stems (Freedle & Kostin, 1993a).

### *Corpus analysis for predicting difficulty*

There have been a number of studies that have examined the correlational relationships of corpus analysis results of test items and item difficulty as well as their explanatory power for item difficulty (Freedle & Kostin, 1993a, 1993b; Hamada, 2015; Loukina, Yoon, Sakano, Wei, & Sheehan, 2016; Rupp et al., 2001). For example, Choi (1994) investigated the correlation between Test Method Facets (Bachman, 1990; Test Task Characteristics: Bachman & Palmer, 1996) and item difficulty in order to examine the content and construct validity of an EFL test. The analysis of the length, content, topic, structure, and readability index of passages revealed many features having a significant correlation with the difficulty of the item. However, in order to determine the complex and in-depth relationship between corpus analysis results and item difficulty, it is necessary to conduct more research using statistical methods other than correlation. In addition, there have been studies designed to predict the difficulty level of the items (Beinborn, Zesch, & Gurevych, 2014; Hoshino & Nakagawa, 2010), but these results are also limited in that these studies mostly involved relatively few questions and participants, and thus call for further study.

Studies investigating item difficulty and its predictors have mostly concentrated on reading and listening comprehension items. According to the results of analyzing test items in reading comprehension and listening comprehension areas, syntactic complexity and vocabulary difficulty factors showed a significant relationship with item difficulty for both sections (Blau, 1990; Freedle & Kostin, 1996; Rupp et al., 2001). Freedle and Kostin (1996) found that the factors necessary for overall language comprehension were significantly predictive of the difficulty of both reading and listening test sections. Another study by Rupp et al. (2001) reported that when analyzing the factors that predict the difficulty of reading and listening areas using regression analysis, item difficulty showed a tendency to increase as the item's sentence length, vocabulary count, and vocabulary variation increased, and was also influenced by lexical density and item type.

However, various differences do exist in reading and listening comprehension tests, since the nature of reading and listening is clearly different and the linguistic elements that each section aims to assess are accordingly discrepant. Thus, a number of previous studies focused on a specific section of a test when examining the relationship between the results of the corpus analysis and the degree of difficulty. In order to predict the difficulty level of items within the reading comprehension section, previous studies have examined the influence of features on the test items, considering the characteristics of the reading comprehension section (Freedle & Kostin, 1991, 1993a, 1993b; Hamada, 2015). Some of the features that have been found to influence reading items' difficulty levels are the level of the vocabulary, the length of the passage, and propositional contents. Propositional contents include the type of information (concreteness/abstractness) of the passage, the distribution of information (compact vs. diffuse; number of content words/total number of words), and the degree of contextualization (context-embedded vs. context-reduced) (Bachman, Davidson, Ryan, & Choi, 1995). In a study by Bachman et al. (1995) that investigated the comparability of two tests, elements of propositional contents were found to be influential variables.

Hamada (2015) conducted a study on the reading comprehension items of the Japanese Eiken English test using two features each of vocabulary, syntactic structure, and semantic structure from Coh-metrix. All six variables significantly predicted the item difficulty, and among the three categories, vocabulary features had the strongest explanatory power. Another study by Freedle and Kostin (1993b) analyzed reading comprehension items of the TOEFL and used features divided into 12 categories, including negations, directives, vocabulary, sentence length, paragraph length, and concreteness. The results showed that the combination of features showed a maximum explanatory power of 58% for the item difficulty level.

On the other hand, there has been relatively less research on predicting the difficulty of the listening comprehension section compared to reading, and the features that showed significant explanatory power for the reading items were commonly used for analysis in the listening comprehension area as well (Rupp et al., 2001). Freedle and Kostin's (1996) study of the listening comprehension section focusing on short conversation items additionally included features specific to the characteristics of listening comprehension, such as emphasis, fillers, and diversity of topics as well as the features that they used for their study of the reading comprehension section (Freedle & Kostin, 1993a, 1993b). As a result of the correlation analysis, the features that showed a significant correlation with the listening items were also those found significant in their former study for the reading comprehension section, but the difference for listening items was that the test-taker's memory played a larger role. Nevertheless, as there is a significantly high correlation between reading and listening comprehension sections (Hale, Rock, & Jirele, 1989) and as research results indicate that the features related to the item difficulty of the two sections are common to both, using the same features to predict the difficulty level of items of reading and listening comprehension would not likely pose a problem (Freedle & Kostin, 1996).

### Expert judgment

During the test development process, the test developers' judgment of item difficulty plays a significant role. In particular, in the case of developing a test for which a pilot test cannot be conducted, the actual difficulty level of the test cannot be obtained; therefore, the adjustment of the test's difficulty level tends to depend on the experts' judgment of the difficulty. Despite this importance and influence of expert judgment in ensuring an appropriate level of test difficulty, not much research has been conducted regarding difficulty prediction and the accuracy of expert judgment, especially in written tests. Only a few studies have addressed this issue, focusing on oral language proficiency (Elder, Iwashita, & McNamara, 2002; Sydorenko, 2011) and essay tests (Hamp-Lyons & Mathias, 1994) in the EFL context. Thus, it is deemed necessary to investigate the extent to which the expert judgment can contribute to explaining the difficulty of standardized EFL reading and listening tests.

One study that compared experts' subjective difficulty predictions and text analysis results, in addition to predicting item difficulty with corpus features, used the order of the items in the test as a criterion for difficulty predictions (Loukina et al., 2016). In other words, the order of the items was used as the predicted difficulty, assuming that the order of the items in the test is determined by the difficulty of the item from easy to difficult. This judgment is likely to be based on the test developers' previous experience and their understanding of item difficulty factors. However, as mentioned in Loukina et al. (2016), there are many doubts regarding the difficulty prediction used in this study because the item sequence within a test may not entirely be based on difficulty prediction and it is impossible to control for the exact method or process of difficulty prediction. Despite these clear limitations, the lack of alternative measurements appropriate for difficulty prediction led this study to use item sequence instead of actual expert judgments.

When comparing the predicted difficulty obtained through this method and the complexity features of the passage, various corpus features predicted the difficulty of many items as well as the predicted difficulty, and in some cases, more accurately. In the case of a number of specific items, the text analysis features showed a stronger explanatory power than the difficulty predicted by experts on the basis of item sequence. As for the text analysis features, lexical properties of the test were the highest ranked feature of item difficulty. However, the results of this study should be considered with caution, as there is a critical limitation in the method of measuring experts' predicted difficulty. Therefore, further studies using the predicted difficulty of experts obtained with an accurate and reasonable method are needed.

Another aspect to explore of predicted difficulty is the comparison of difficulty predictions made by native speakers and non-native speakers for English tests administered within the EFL environment. A difference in item difficulty judgment between the two groups is likely to be observed, as native speakers will be able to review the authenticity of the language used in the test, while non-native speakers will be better experienced and informed about the level of EFL students and their expected responses to the test questions. To be specific, native speakers will assess test item difficulty based on a general perception of words and phrases in their view and context when predicting item difficulty, while non-native speakers will have relatively more knowledge of what L2 students may find difficult.

This may lead to differences in the elements they consider during their process of a rather holistic judgment of items. As documented in holistic judgment processes, evaluators tend to consider different elements and place different weights on each element (Eckes, 2008). Since there is no way of knowing the exact process of judgment even with rigorous training and detailed rubrics, it is necessary to investigate whether there is a significant difference in the difficulty prediction between native and non-native speakers. If there is a difference, steps should be taken to determine which group shows a more accurate predictability level for the test items. However, there have been almost no previous studies that have analyzed and directly compared the predicted difficulties of native and non-native speakers.

Therefore, in this study, the explanatory power of expert judgment by both native and nonnative speakers and the results of corpus analysis as a criterion for predicting the content difficulty of the Test of English for International Communication (TOEIC) and the New Test of English Proficiency developed by Seoul National Univ. (New TEPS) were analyzed and compared. Thus, an attempt to find a desirable method for predicting item difficulty through the accuracy and reliability of corpus analysis and expert (NS, NNS) judgment will be made by examining the following research questions of this study.

(1) Is there a noteworthy relationship between item difficulty, predicted difficulties (NS, NNS), and corpus features?
(2) Do the predicted difficulties of NS and NNS account for item difficulty? If so, which group shows higher explanatory power?
(3) What is the extent to which the corpus features differ from the NS- and NNS-predicted difficulty in accounting for item difficulty?
(4) To what extent does combining the corpus features with the predicted difficulties contribute to explanatory power than when applied separately?

## Research method

### Research data

The texts to be analyzed in this study are the two major standardized EFL tests (i.e., the TOEIC and the New TEPS) that have exerted the most significant washback impact on post-secondary English education in South Korea. The TOEIC simulation test has been conducted by leading educational institutions, and the New TEPS first pilot test has been administered by Seoul National University's TEPS Center. Only the reading and listening sections of the TOEIC and New TEPS were used in this study. For the analysis of the TOEIC test (whose official test was not available to this study), the mock TOEIC test was used, as it can be regarded as being almost in conformity with the composition and contents of the actual TOEIC test. The contents of the TOEIC are quite different from those of the New TEPS. For instance, 70 questions within the reading section of the TOEIC are divided into two parts: Part 6 with 16 fill-in-the-blank questions within a paragraph question, and Part 7 with 54 questions based on non-academic passages followed by more than two items per passage or multiple passages. As for the listening section, containing 100 questions, the parts are divided as follows: Part 1: picture description (6 questions); Part 2: infer a response to a short dialogue (25 questions); Part 3: answer questions on a long dialogue (39 questions); and Part 4: answer questions on a long non-academic monologue (30 questions).

The New TEPS 1st pilot test that was used in this study is the former version of the New TEPS implemented in Korea from May 2018. The New TEPS 1st pilot test consists of a test system based on the needs analysis of the examinees and test users that is somewhat similar to that of the original TEPS, but where the number of questions is reduced and the test time slightly shortened. Some minor differences from the pilot test and the newly implemented New TEPS test do exist, such as the sequence of parts within the Reading section. The composition of the 35 questions of the reading part is as following: Part 1: fill-in-the-blank within a passage (10 questions); Part 2: answer a question on an academic passage (13 questions); Part 3: answer questions on a non-academic passage (10 questions); and Part 4: choose the sentence that does not fit (2 questions). On the other hand; the 40 questions of the New TEPS Listening consist of the following Parts: Part 1: infer a response to a short dialogue (10 questions); Part 2: infer a response to a long dialogue (10 questions); Part 3: answer a question on a long dialogue (10 questions); Part 4: answer a question on a short monologue (6 questions); and Part 5: answer two questions on a long academic monologue (4 questions).

As for the participants, students from six universities within South Korea took the New TEPS pilot test and mock TOEIC test for a study designed to explore the extent to which the two tests are comparable in terms of test content and construct. The participants can be considered to represent a wide range of proficiency levels, as there were both high- and low-ranked universities within Korea among the six universities. The number of participants in the New TEPS reading and listening section was 251 and 315, respectively, while the number of participants in the TOEIC reading and listening was 124 and 169, respectively.

### Corpus analysis method

To obtain the corpus features, the original texts of New TEPS and TOEIC test reading and listening sections were converted into a text file (hereafter, the two tests will be referred to in alphabetical order). A corpus analysis was conducted considering the nature of the test texts after converting the instructions, the number of choices, and blanks. Text conversion was done collectively in accordance with the following principle:

- Delete the choice number (a) (b) (c) (d)
- Unify the blanks as ‘_____’ or fill them with the correct answers,
- and then delete the blanks
- Batch conversion of unrecognized symbols
- Delete the speaker in the script of the file ‘M:’, ‘W:’
- Delete ‘Q:’ indicating a question
- List non-level words in accordance with the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA)
- Delete all Direction parts
- Delete the symbols < and >, which cause errors in Coh-metrix

The reading and listening comprehension passages of the short test type designed to contain fewer than two exchanges were not included in the corpus analysis so as to ensure the robustness of the corpus analysis results. A corpus analysis was performed for the New TEPS listening area on the long conversations of Part 3, short monologues of Part 4, and long monologues of Part 5, from #21 to 40. For the listening comprehension area of the TOEIC, the long conversations of Part 3 and long monologues of art 4 were analyzed. In the case of reading comprehension, a corpus analysis of all items included in the New TEPS and TOEIC was conducted. The composition of the test parts included in this study is given in Table 1 below.

A total of five corpus analysis tools were used for the above text files and a total of 26 corpus features (see Table 2) were used in this study. Among the 53 features collected from the various corpus analysis tools, 26 features were chosen on the basis of their importance (Freedle & Kostin, 1993a, 1993b, 1996; Hamada, 2015) and relevance to this study, as well as their correlational relationship with the test difficulties (see Appendix).

**Table 1.** Test composition of the analyzed text.

| Test | Part | Item Type | Items | Passage |
|---|---|---|---|---|
| *New TEPS* | | | | |
| Reading | Part 1 (fill-in-the-blank) | OPOI | 10 | - |
| | Part 2 (multiple choice: academic) | OPOI | 13 | - |
| | Part 3 (multiple choice: non-academic) | OPMI | 10 | 5 |
| | Part 4 (sentence deletion) | OPOI | 2 | - |
| Listening | Part 3 (long dialogue) | OPOI | 10 | - |
| | Part 4 (short monologue) | OPOI | 6 | - |
| | Part 5 (long monologue) | OPMI | 4 | 2 |
| *TOEIC* | | | | |
| Reading | Part 6 (fill-in-the-blank) | OPMI | 16 | 4 |
| | Part 7 (multiple choice: non-academic) | OPMI | 54 | 15 |
| Listening | Part 3 (long dialogue) | OPMI | 39 | 13 |
| | Part 4 (long non-academic monologue) | OPMI | 30 | 10 |

* OPOI: One-Passage One-Item, OPMI: One-Passage Multiple-Items

**Table 2.** 26 corpus features used in the current study.

| Variable | Meaning |
|---|---|
| *Vocabulary features* | |
| 1. Token | The number of tokens in the passage |
| 2. Lexical Token | The number of lexical words in the passage (a word was considered a lexical word if it was classified as a noun, verb, but not modal or auxiliary verb, adjective, or lexical adverb) |
| 3. Type | The number of types in the passage |
| 4. Lexical Type | The number of different lexical words in the passage |
| 5. Syllable/Word | The average number of syllables in one word, indicating the length of words |
| 6. Type-Token-Ratio (TTR) | The ratio of the number of types to the number of tokens or words in a text (Templin, 1957) |
| 7. Standardized TTR | A standardized TTR is used to complement the TTR since it is sensitive to passage length |
| 8. Verb Variation | The ratio of the number of verb types to the total tokens or number of verbs (Harley & King, 1989) |
| 9. Lexical Density | The ratio of the number of lexical words to the number of words |
| 10. Lexical Sophistication | The ratio of the number of sophisticated word types, defined as words beyond the most frequent 2,000 words, to the total number of word types in a text (Laufer, 1994) |
| *Syntactic features* | |
| 11. Sentence | The number of sentences in the text |
| 12. Clause | The number of clauses in the text |
| 13. T-unit | The number of t-units in the text |
| 14. Complex Nominal | The number of complex nominals in the text |
| 15. Syntactic Complexity | The inverse of the syntactic simplicity index, which is based on the degree to which the sentences contain shorter lengths and simpler syntactic structures |
| 16. Voice | The number of incidences of agentless passive voice forms |
| 17. Mean Length of Sentence | The average number of words per sentence (MLS) |
| 18. Mean Length of t-unit | The average number of t-units per sentence (MLT) |
| 19. Clause/Sentence | The average number of words in one sentence, indicating the length and complexity of the sentence |
| 20. Ratio of Complex t-units | The ratio of complex t-units to simple t-units |
| 21. F minus | The inverse of the readability index, Flesch-Kincaid reading ease, indicating the complexity of the passage |
| 22. FOG | The readability index of a passage, indicating the years of education required to understand the passage |
| *Pragmatic features* | |
| 23. Negation | The proportion of negations (*no, not, nothing, never*, and so on) in the passage |
| 24. Subjunctive | The proportion of modal auxiliary verbs used to express counterfactual states or unreality in the passage |
| 25. Academic/Nonacademic | The coverage rate of the passage based on the Academic Word List, indicating the academic character of the passage |
| 26. Academic/Practical | The topic of the passage, either classified as academic, such as passages based academic topics such as history, social science, and science, or as practical, such as advertisements, e-mails, news articles, or invoices |

### Coh-metrix

Coh-metrix is a corpus tool that measures the consistency and coherence of the passages (Graesser et al., 2004). Among the various features, only the salient features were used for this study.

### Readability

Readability is an analytical tool that measures the readability of the passage. The features used in this study are Kincaid-Flesch reading ease and the FOG index showing the English education grade required to read the passage.

### AntWordProfiler

AntWordProfiler is a program that analyzes the word structure and complexity of passages, and the file that serves as the criterion of analysis can be uploaded directly (Anthony, 2014). In this study, it was used to identify the academic/non-academic nature of the test passages.

### Lexical complexity analysis (LCA)

LCA is a program that analyzes the lexical complexity of a text using various measurement methods (Ai & Lu, 2010; Lu, 2012). In this study, many features related to vocabulary (number of words, vocabulary variation, vocabulary characteristic, etc.) were obtained using LCA.

### L2 syntactical complexity analysis (L2SCA)

If the LCA is a tool for lexical level corpus analysis, L2SCA is a tool for the syntactic level (Lu, 2010). It analyzes various syntactic complexity features based on clauses, phrases, and t-units.

## Expert judgment

The predicted difficulties for the current study were provided by a total of 60 experts of English who determined the predicted difficulty on a scale of 1 point (very easy) to 5 points (very difficult) for each item of the New TEPS and TOEIC reading and listening sections. Among the 60 experts, 30 were educated native speakers and 30 were educated nonnative speakers, all of whom had taken an EFL assessment as graduate students majoring in TEFL. As all of them were involved in teaching EFL in the public or private sector, they had experience developing and reviewing EFL tests. Their understanding of difficulty required to develop good quality EFL tests through their professional TEFL training and experience is deemed to qualify them as appropriate subjects for making judgments on item difficulty.

The native and nonnative experts were not trained to evaluate the particular tests in question, nor were they given specific evaluation elements or a rubric. Considering the highly complicated nature of perceiving item difficulty due to a variety of factors, they were asked to judge the overall difficulty of each item in a holistic manner based on their previous experience and instinct. In the case of the listening tests, they were not given any transcript to read, but were asked to listen to the aural input. The average predicted difficulty score of the New TEPS and TOEIC for reading and listening comprehension of each item was used in the study distinguished by native speakers (NS) and nonnative speakers (NNS). The nature of the one-passage multi-item method made it virtually impossible to compare the passage-based corpus analysis and the item-based difficulty judgment with one-to-one correspondence. Therefore, the scores of the items in the form of a one-passage multi-item type were calculated as an average score to ensure one predicted difficulty score corresponding to each passage.

The interrater reliability of the predicted difficulty for each test was found to be reliable as the Cronbach's α ranged from 0.77 to 0.97. The reliability of the NS-predicted difficulty of the New TEPS reading and listening were 0.85 and 0.88, and 0.77 and 0.83 for the TOEIC reading and listening, respectively, while the reliability of the NNS-predicted difficulty was 0.94 and 0.93 for the New TEPS reading and listening, respectively, and 0.97 and 0.96 for the TOEIC reading and listening, respectively.

### Data analysis

The test items and passages of the two tests' reading and listening comprehension sections were first converted into text files in order to conduct corpus analyses addressing imperative and meaningful features relevant to the aim of this research. After the corpus analysis was completed, 60 experts were asked to submit their predicted difficulty for the items of New TEPS and TOEIC reading and listening comprehension sections. As the corpus analysis was conducted for each passage, the average of the NS- and NNS-predicted difficulty for each item in a one-passage multiple-item type was averaged once more. Once the data of the corpus analysis and predicted difficulty were obtained, inferential statistical analyses, including correlation analysis (Spearman's rho for the predicted difficulty, given the data of ordinal scale, and Pearson correlation for the corpus features), *t*-test, and regression analysis using SPSS 22, were conducted.

## Results and discussion

### Comparing the NS- and NNS-predicted difficulty

The descriptive statistics of the observed item difficulty and predicted difficulty for both NSs and NNSs are given in Tables 3 and 4 below. The observed difficulty is the proportion of test-takers who got the item wrong, and the predicted difficulty is a score based on a scale of 1 (very easy) to 5 (very difficult). A total of 30 reading and 18 listening passages of the New TEPS and 19 reading and 23 listening passages of the TOEIC were analyzed. To begin with, the observed difficulties of the New TEPS reading and listening section were 0.51 and 0.50 (see Table 3), respectively, and those of the TOEIC were 0.50 and 0.53 (see Table 4), respectively. All four test sections showed a similar average difficulty from 0.50 to 0.53. On the other hand, the predicted difficulty for each section by NSs (1.77 to 2.34) and NNSs (2.58 to 3.53) showed a wider range among the tests. The predicted difficulty of the TOEIC listening section was the lowest of the four sections for both NSs (1.77) and NNSs (2.58), whereas the New TEPS reading (NS = 2.34, NNS = 3.32) and listening (NS = 2.27, NNS = 3.53) showed a higher predicted difficulty by NSs and NNSs than the TOEIC. Overall, NSs tended to rate relatively lower predicted difficulty levels for test items than NNSs, as all four sections were found to have a lower predicted difficulty for NSs.

**Table 3.** Descriptive statistics of new TEPS reading and listening sections.

| Variables | *N* | Mean | *SD* | Min. | Max. |
|---|---|---|---|---|---|
| *New TEPS Reading* | | | | | |
| Observed Difficulty | 30 | 0.51 | 0.15 | 0.19 | 0.77 |
| NS-Predicted Difficulty | 30 | 2.34 | 0.71 | 1.23 | 3.78 |
| NNS-Predicted Difficulty | 30 | 3.32 | 0.77 | 1.77 | 5.00 |
| *New TEPS Listening* | | | | | |
| Observed Difficulty | 18 | 0.50 | 0.14 | 0.19 | 0.73 |
| NS-Predicted Difficulty | 18 | 2.27 | 0.74 | 1.00 | 3.37 |
| NNS-Predicted Difficulty | 18 | 3.53 | 0.82 | 2.27 | 4.67 |

**Table 4.** Descriptive statistics of TOEIC reading and listening sections.

| Variables | *N* | Mean | *SD* | Min. | Max. |
|---|---|---|---|---|---|
| *TOEIC Reading* | | | | | |
| Observed Difficulty | 19 | 0.50 | 0.06 | 0.40 | 0.68 |
| NS-Predicted Difficulty | 19 | 2.06 | 0.54 | 1.30 | 3.03 |
| NNS-Predicted Difficulty | 19 | 3.10 | 0.59 | 2.08 | 4.17 |
| *TOEIC Listening* | | | | | |
| Observed Difficulty | 23 | 0.53 | 0.10 | 0.36 | 0.70 |
| NS-Predicted Difficulty | 23 | 1.77 | 0.45 | 1.07 | 2.59 |
| NNS-Predicted Difficulty | 23 | 2.58 | 0.50 | 1.88 | 3.44 |

### Relations of observed difficulty and NS, NNS-predicted difficulty with corpus features

Next, in order to investigate the potential relationships between the observed difficulty, the NS- and NNS-predicted difficulty variables, and various corpus features, correlational analysis was conducted. For the correlation between the difficulty variables (observed difficulty, NS- and NNS-predicted difficulty) and corpus features, Pearson's *r* was used, and for the correlation between the observed difficulty and NS- and NNS-predicted difficulties, Spearman's $r_s$ was used. Based on the Bonferroni correction, no variables were found to be significant due to its stringent characteristic. Thus, the correlation results will be discussed based on the correlation coefficients rather than its *p* value which will be presented in the following tables. The correlation results between corpus features were excluded and only the results relevant to this study, which are the correlations between difficulty variables and corpus features, will be discussed.

The correlation analysis results of the New TEPS reading section are given in Table 5. Based on the correlation coefficients, syllable/word (*r* = .483), F minus (*r* = .475), FOG (*r* = .488), and academic/nonacademic passage (*r* = .369) demonstrated the strongest correlation with the observed difficulty. Somewhat surprisingly, features related to passage length, such as the number of types, tokens, or sentences, did not show a strong correlation, but a weak to moderate correlation. This may be because the passage length of the New TEPS is quite consistent throughout the reading section. Overall, the length of words and the readability features were found to have a moderate to strong correlation with the observed difficulty for the New TEPS reading test.

On the other hand, the features that had the strongest correlation with NS- and NNS-predicted difficulties were different from those of the observed difficulty. NS-predicted difficulty had a moderate correlation with the syllable/word variable, representing the average word length, and the readability features of F minus (*r* = .397) and FOG (*r* = .390), while NNS-predicted difficulty was moderately correlated to verb variation (*r* = .359), syntactic complexity (*r* = .391), F minus (*r* = .359), FOG (*r* = .366), and negation (*r* = .371). The readability features may have also influenced the experts' judgment of item difficulty based on these results.

Next, the correlation analysis results of the New TEPS listening in Table 6 show corpus features that have a moderate to strong correlation with the difficulty variables. For the observed difficulty, features related to length including token (*r* = .547), lexical token (*r* = .513), type (*r* = .504), and lexical type (*r* = .522) demonstrated a strong correlation. In addition, verb variation (*r* = .451), clause (*r* = .404) and the ratio of complex t-units (*r* = .508) also had a relatively strong correlation with the observed difficulty than the other corpus features. Compared to the New TEPS reading test, the

**Table 5.** Correlation of observed and predicted difficulties with corpus features of new TEPS reading.

| Variables | Token | Lexical Token | Type | Lexical Type | Syllable/Word | TTR |
|---|---|---|---|---|---|---|
| Observed Difficulty | .234 | .267 | .245 | .289 | .483 | −.253 |
| NS PD | .204 | .225 | .220 | .271 | .328 | −.252 |
| NNS PD | .227 | .238 | .250 | .295 | .325 | −.262 |

| Variables | Std. TTR | Verb Variation | Lexical Density | Lexical Sophistication | Sentence | Clause |
|---|---|---|---|---|---|---|
| Observed Difficulty | −.262 | .343 | .139 | −.032 | .195 | .230 |
| NS PD | −.240 | .266 | .009 | .058 | .103 | .226 |
| NNS PD | −.245 | .359 | −.060 | .086 | .155 | .268 |

| Variables | T-unit | Complex Nominal | Syntactic Complexity | Voice | MLS | MLT | Clause/Sentence |
|---|---|---|---|---|---|---|---|
| Observed Difficulty | .179 | .205 | .285 | .135 | .010 | .019 | .112 |
| NS PD | .121 | .189 | .194 | .129 | .173 | .126 | .241 |
| NNS PD | .170 | .191 | .391 | .213 | .140 | .061 | .280 |

| Variables | Complex T-unit | F minus | FOG | Negation | Subjunctive | Non/academic | Academic/Practical |
|---|---|---|---|---|---|---|---|
| Observed Difficulty | .114 | .475 | .488 | .175 | .226 | .369 | −.069 |
| NS PD | .207 | .397 | .390 | .212 | .210 | .124 | −.002 |
| NNS PD | .220 | .359 | .366 | .371 | .281 | .148 | −.099 |

PD: Predicted Difficulty

**Table 6.** Correlations of observed and predicted difficulties with corpus features of new TEPS listening.

| Variables | Token | Lexical Token | Type | Lexical Type | Syllable/Word | TTR |
|---|---|---|---|---|---|---|
| Observed Difficulty | .547 | .513 | .504 | .522 | .234 | −.304 |
| NS PD | .585 | .716 | .588 | .663 | .524 | −.111 |
| NNS PD | .601 | .751 | .614 | .695 | .738 | −.061 |

| Variables | Std. TTR | Verb Variation | Lexical Density | Lexical Sophistication | Sentence | Clause |
|---|---|---|---|---|---|---|
| Observed Difficulty | −.197 | .451 | .135 | .383 | .177 | .404 |
| NS PD | −.003 | .437 | .582 | .547 | −.107 | .055 |
| NNS PD | .064 | .307 | .657 | .585 | −.191 | −.022 |

| Variables | T-unit | Complex Nominal | Syntactic Complexity | Voice | MLS | MLT | Clause/Sentence |
|---|---|---|---|---|---|---|---|
| Observed Difficulty | .223 | .237 | .187 | .268 | .350 | .236 | .223 |
| NS PD | −.039 | .589 | .032 | .472 | .724 | .575 | −.039 |
| NNS PD | −.137 | .722 | −.173 | .394 | .837 | .732 | −.137 |

| Variables | Complex T-unit | F minus | FOG | Negation | Subjunctive | Non/academic | Academic/Practical |
|---|---|---|---|---|---|---|---|
| Observed Difficulty | .508 | .254 | .258 | .306 | −.323 | .032 | −.203 |
| NS PD | .173 | .539 | .532 | .013 | −.055 | .615 | −.304 |
| NNS PD | .175 | .742 | .707 | −.058 | −.152 | .675 | −.494 |

difficulty of the listening test seemed to have a strong correlation with the passage length and syntactic features of the listening.

As for the predicted difficulties, several corpus features related to vocabulary, syntax, and pragmatics had a considerably high correlation with both NS- and NNS-predicted difficulties. In particular, all of the features indicating the length of the passage and readability had a strong correlation with the NS- and NNS-predicted difficulty. Also, several lexical and syntactic features and the non/academic feature that did not have a strong correlation with the observed difficulty demonstrated a strong correlation with the predicted difficulties. This indicates the possibility that the item difficulty judgment may have been based on various lexical and syntactic features. Another noteworthy point is that the experts have judged lengthier and academic items as more difficult than other items.

The correlation analysis results for the reading section of the TOEIC are given in Table 7. There were only a few corpus features that showed a strong correlation with the observed difficulty, including the average number of syllables within a word ($r$ = .474), F minus ($r$ = .565), and subjunctive ($r$ = .507). Similar to the New TEPS reading, the observed difficulty of the TOEIC reading test was also not correlated to the passage length. Also, no syntactic features included in the correlation analysis demonstrated a moderate correlation with the item difficulty.

Meanwhile, the predicted difficulties of NSs and NNSs had relatively more features that had a strong correlation, which were mostly related to passage length and the variation of vocabulary within the passage. For instance, all four features related to length and the three features indicating vocabulary variation including TTR, standardized TTR, and verb variation were found to have a relatively strong correlation with both NS- and NNS-predicted difficulty. Unlike the actual observed difficulty, the experts may have taken the reading passage length into consideration in the item difficulty judgment process. Since all of the passages of the TOEIC reading section were of a practical rather than academic nature, the correlation result for the academic/practical feature was not given.

Lastly, the correlation analysis results for the TOEIC listening section are given in Table 8. For the observed difficulty, most of the features related to passage length showed a moderate to strong relationship, as did several other features, including the average number of syllables in a word ($r$ = .551), standardized TTR ($r$ = .422), verb variation ($r$ = .423), complex nominal ($r$ = .475), F minus ($r$ = .519), and academic/nonacademic passages ($r$ = .472). It seems like the difficulty of the TOEIC listening test was influenced by lexical and syntactic features including the passage length.

Meanwhile, for the predicted difficulties, there were relatively more features that showed a stronger correlation with the NS-predicted difficulty. Both NS- and NNS-predicted difficulty

Table 7. Correlations of observed and predicted difficulties with corpus features of TOEIC reading.

| Variables | Token | Lexical Token | Type | Lexical Type | Syllable/Word | TTR |
|---|---|---|---|---|---|---|
| Observed Difficulty | .239 | .271 | .228 | .254 | .474 | −.227 |
| NS PD | .747 | .744 | .754 | .733 | .402 | −.573 |
| NNS PD | .801 | .778 | .808 | .778 | .364 | −.641 |

| Variables | Std. TTR | Verb Variation | Lexical Density | Lexical Sophistication | Sentence | Clause |
|---|---|---|---|---|---|---|
| Observed Difficulty | −.129 | .168 | .299 | .208 | −.002 | .167 |
| NS PD | −.466 | .608 | .188 | .421 | −.201 | −.155 |
| NNS PD | −.537 | .639 | .110 | .413 | −.292 | −.264 |

| Variables | T-unit | Complex Nominal | Syntactic Complexity | Voice | MLS | MLT | Clause/Sentence |
|---|---|---|---|---|---|---|---|
| Observed Difficulty | .049 | −.059 | −.134 | −.189 | .117 | −.013 | .391 |
| NS PD | −.231 | −.277 | .217 | .193 | −.229 | −.256 | .078 |
| NNS PD | −.344 | −.276 | .322 | .153 | −.193 | −.162 | .053 |

| Variables | Complex T-unit | F minus | FOG | Negation | Subjunctive | Non/academic | Academic/Practical |
|---|---|---|---|---|---|---|---|
| Observed Difficulty | .278 | .565 | .364 | −.267 | −.507 | .112 | – |
| NS PD | −.037 | .381 | −.002 | −.433 | −.494 | .124 | – |
| NNS PD | −.016 | .328 | .029 | −.372 | −.404 | .121 | – |

Table 8. Correlations of observed and predicted difficulties with corpus features of TOEIC listening.

| Variables | Token | Lexical Token | Type | Lexical Type | Syllable/Word | TTR |
|---|---|---|---|---|---|---|
| Observed Difficulty | .356 | .434 | .462 | .498 | .551 | .212 |
| NS PD | .491 | .558 | .629 | .654 | .646 | .206 |
| NNS PD | .449 | .484 | .556 | .556 | .563 | .136 |

| Variables | Std. TTR | Verb Variation | Lexical Density | Lexical Sophistication | Sentence | Clause |
|---|---|---|---|---|---|---|
| Observed Difficulty | .422 | .423 | .391 | .161 | .119 | .078 |
| NS PD | .419 | .462 | .385 | .368 | .380 | .245 |
| NNS PD | .325 | .363 | .270 | .373 | .258 | .126 |

| Variables | T-unit | Complex Nominal | Syntactic Complexity | Voice | MLS | MLT | Clause/Sentence |
|---|---|---|---|---|---|---|---|
| Observed Difficulty | .009 | .475 | .009 | −.030 | .221 | .383 | .017 |
| NS PD | .360 | .433 | .006 | .065 | .139 | .173 | −.138 |
| NNS PD | .220 | .412 | .084 | .042 | .225 | .286 | −.161 |

| Variables | Complex T-unit | F minus | FOG | Negation | Subjunctive | Non/academic | Academic/Practical |
|---|---|---|---|---|---|---|---|
| Observed Difficulty | .212 | .519 | .287 | −.374 | .238 | .472 | – |
| NS PD | −.154 | .614 | .375 | −.271 | .118 | .318 | – |
| NNS PD | −.159 | .506 | .346 | −.290 | .109 | .398 | – |

demonstrated a strong relationship with all the vocabulary features indicating passage length, as well as with a few other features. However, other corpus features had a relatively weaker correlation which indicates that the judgment of experts does not necessarily reflect these features of the test. The correlation value was not given for the academic/practical index because, as mentioned for the reading section of the TOEIC, the listening section also consisted of only practical passages.

In order to examine the relation between the observed difficulty and predicted difficulty by NSs and NNSs, the correlation analysis results of the three variables are presented in Tables 9 and 10. The predicted difficulty by both NSs and NNSs of the New TEPS reading ($r_s$ = .695; $r_s$ = .773, respectively), listening ($r_s$ = .615; $r_s$ = .566, respectively) and TOEIC listening ($r_s$ = .604; $r_s$ = .594, respectively) showed moderate to high correlations. However, while the NS-predicted difficulty of the TOEIC reading had a strong correlation ($r_s$ = .544), NNS PD had a relatively moderate correlation with the observed difficulty ($r_s$ = .406). Overall, the predicted difficulties were strongly correlated with the observed difficulty for the New TEPS and TOEIC tests. In addition, the correlations between NS- and NNS-predicted difficulty were found to be strong for all four test sections. This does not imply that the predicted difficulties judged by NSs and NNSs were congruent

**Table 9.** Correlations of difficulty variables of new TEPS reading and listening.

| Variables | 1 | 2 | 3 |
|---|---|---|---|
| *New TEPS Reading* | | | |
| 1. Observed Difficulty | – | | |
| 2. NS-Predicted Difficulty | .695 | – | |
| 3. NNS-Predicted Difficulty | .773 | .860 | – |
| *New TEPS Listening* | | | |
| 1. Observed Difficulty | – | | |
| 2. NS-Predicted Difficulty | .615 | – | |
| 3. NNS-Predicted Difficulty | .566 | .896 | – |

**Table 10.** Correlations of difficulty variables of TOEIC reading and listening.

| Variables | 1 | 2 | 3 |
|---|---|---|---|
| *TOEIC Reading* | | | |
| 1. Observed Difficulty | – | | |
| 2. NS-Predicted Difficulty | .544 | – | |
| 3. NNS-Predicted Difficulty | .406 | .914 | – |
| *TOEIC Listening* | | | |
| 1. Observed Difficulty | – | | |
| 2. NS-Predicted Difficulty | .604 | – | |
| 3. NNS-Predicted Difficulty | .594 | .929 | – |

with each other, but that the overall rank order of item difficulty perceived by NSs followed a similar pattern of those by NNSs.

## Examining the difference between NS- and NNS-predicted difficulty

Noting that there was a close relation between NS- and NNS-predicted difficulty, and that most of the predicted difficulties also had a significant correlation, *t*-tests were conducted for each section to determine whether the two different groups' predicted difficulties were significantly different (see Table 11). The *t*-test results indicated that there was a significant difference between NSs (*M* = 2.34, *SD* = 0.71; *M* = 2.27, *SD* = 0.74, respectively) and NNSs (*M* = 3.32, *SD* = 0.77; *M* = 3.53, *SD* = 0.82, respectively) for both the New TEPS reading and listening sections (*t* = – 15.90, *p* < .01; *t* = – 16.934, *p* < .01, respectively). The *t*-test results also showed that there was a significant difference between NSs (*M* = 2.06, *SD* = 0.54; *M* = 1.77, *SD* = 0.45, respectively) and NNSs (*M* = 3.10, *SD* = 0.59; *M* = 2.58, *SD* = 0.50, respectively) for the TOEIC reading and listening sections (*t* = – 20.371, *p* < .01; *t* = – 19.922, *p* < .01, respectively).

The Cohen's *d* was also calculated to investigate the effect size of the mean differences found from the *t*-test. The Cohen's *d* for the New TEPS reading and listening were 1.33 and 1.6, respectively,

**Table 11.** Descriptive statistics and *T*-tests for NS- and NNS-predicted difficulty.

| Section | Group | *N* | Mean | Std.deviation | Std. error mean | Mean difference | Std.deviation | *t* |
|---|---|---|---|---|---|---|---|---|
| *New TEPS* | | | | | | | | |
| Reading | NS | 30 | 2.338 | 0.706 | 0.129 | 0.981 | 0.338 | −15.900** |
| | NNS | 30 | 3.319 | 0.766 | 0.140 | | | |
| Listening | NS | 18 | 2.268 | 0.737 | 0.174 | 1.261 | 0.316 | −16.934** |
| | NNS | 18 | 3.529 | 0.824 | 0.194 | | | |
| *TOEIC* | | | | | | | | |
| Reading | NS | 19 | 2.060 | 0.543 | 0.126 | 1.041 | 0.223 | −20.317** |
| | NNS | 19 | 3.101 | 0.588 | 0.135 | | | |
| Listening | NS | 23 | 1.773 | 0.449 | 0.093 | 0.808 | 0.195 | −19.922** |
| | NNS | 23 | 2.581 | 0.500 | 0.104 | | | |

indicating that the mean differences were meaningful. In addition, for the TOEIC reading and listening, effect sizes of the Cohen's *d* were 1.84 and 1.7, respectively. The TOEIC tests showed that the mean differences between the NS and NNS predicted difficulties were also meaningful. Thus, the New TEPS and TOEIC tests did have a noteworthy difference between NS- and NNS-predicted difficulty for both reading and listening sections.

When comparing the mean of predicted difficulty, it seems that the NNSs showed a tendency to overrate the predicted item difficulty compared to NSs. This may be because NNSs are able to consider EFL students' English level, whereas NSs will evaluate the items from a native speaker's view. Nevertheless, results from the correlation analysis (see Table 9, 10) demonstrate that, in fact, NS-predicted difficulty tends to show a higher correlation with the actual test difficulties than NNS-predicted difficulty. Only the New TEPS reading section had a higher correlation with NNS-predicted difficulty, although the difference was quite small, while the listening section of the New TEPS and both reading and listening of the TOEIC had a higher correlation with NS-predicted difficulty. The results from the *t*-test and correlation analysis point out the fact that NS- and NNS-predicted difficulty are significantly different, and that NS-predicted difficulty has a similar or higher correlation with the actual item difficulty of both tests.

## Explaining the observed difficulty with predicted difficulties and corpus features

### Comparing the explanatory power of NS- and NNS-predicted difficulty

As a next step, hierarchical regression analyses were conducted with the observed difficulty as the dependent variable for each section of the New TEPS (Table 12) and TOEIC (Table 13). The NS- and NNS-predicted difficulties of each test section were entered in an alternative sequence. For each regression analyses, the Variance Inflation Factor (VIF) was looked upon in order to prevent multicollinearity issues. The VIF values of NS PD and NNS PD entered in the regression model did not exceed 10 for the New TEPS and TOEIC reading and listening tests.

For the New TEPS reading section, NS-predicted difficulty significantly explained 50.7% ($\Delta F$ = 28.83**) and NNS-predicted difficulty explained 53.2% ($\Delta F$ = 31.80**) of the observed difficulty when entered in the first step. For the New TEPS reading, NNS-predicted difficulty had a relatively stronger explanatory power than NS, although the difference was small. However, neither variable contributed additional significant explanatory power when entered in the second step (NS: $R^2$ = 1.7%, $\Delta F$ = 1.02; NNS: $R^2$ = 4.2%, $\Delta F$ = 2.48). In other words, NS- and NNS-predicted difficulties did not account for the variance of the observed difficulty when the other group's predicted difficulty was controlled for.

As for the listening section of the New TEPs, the NS -predicted difficulty demonstrated 43.1% ($\Delta F$ = 12.11**) of the statistically significant variance, while the NNS-predicted difficulty significantly accounted for 30.8% ($\Delta F$ = 7.11*) when entered first. NNS-predicted difficulty ($R^2$ = 1.8%, $\Delta F$ = 0.50) lost its explanatory power when entered after NS, but the NS-predicted difficulty demonstrated a relative influence on the observed difficulty above and beyond the NNS-predicted difficulty, as it

**Table 12.** Hierarchical regression analyses of new TEPS reading and listening observed difficulty.

| Steps | Variables | *R* | $R^2$ | $\Delta R^2$ | $\Delta F$ | *Sig.* |
|---|---|---|---|---|---|---|
| *New TEPS Reading observed difficulty as the outcome* | | | | | | |
| 1 | NS-Predicted Difficulty | .712 | .507 | .507 | 28.83** | .000 |
| 2 | NNS-Predicted Difficulty | .741 | .549 | .042 | 2.48 | .127 |
| 1 | NNS-Predicted Difficulty | .729 | .532 | .532 | 31.80** | .000 |
| 2 | NS-Predicted Difficulty | .741 | .549 | .017 | 1.02 | .321 |
| *New TEPS Listening observed difficulty as the outcome* | | | | | | |
| 1 | NS-Predicted Difficulty | .656 | .431 | .431 | 12.11** | .003 |
| 2 | NNS-Predicted Difficulty | .670 | .449 | .018 | .50 | .492 |
| 1 | NNS-Predicted Difficulty | .555 | .308 | .308 | 7.11* | .017 |
| 2 | NS-Predicted Difficulty | .670 | .449 | .141 | 3.85~ | .069 |

**Table 13.** Hierarchical regression analyses of TOEIC reading and listening observed difficulty.

| Steps | Variables | *R* | *R*$^2$ | *ΔR*$^2$ | *ΔF* | *Sig.* |
|---|---|---|---|---|---|---|
| *TOEIC Reading observed difficulty as the outcome* | | | | | | |
| 1 | NS-Predicted Difficulty | .478 | .229 | .229 | 5.04* | .038 |
| 2 | NNS-Predicted Difficulty | .627 | .393 | .164 | 4.32~ | .054 |
| 1 | NNS-Predicted Difficulty | .289 | .083 | .083 | 1.55 | .230 |
| 2 | NS-Predicted Difficulty | .627 | .393 | .309 | 8.14* | .011 |
| *TOEIC Listening observed difficulty as the outcome* | | | | | | |
| 1 | NS-Predicted Difficulty | .638 | .407 | .407 | 14.41** | .001 |
| 2 | NNS-Predicted Difficulty | .638 | .407 | .000 | .00 | .997 |
| 1 | NNS-Predicted Difficulty | .587 | .345 | .345 | 11.07** | .003 |
| 2 | NS-Predicted Difficulty | .638 | .407 | .062 | 2.08 | .164 |

explained an additional 14.1% (*ΔF* = 3.85~) and its significance (*r* = 0.069) was close to 0.05. Unlike on the reading section, NS-predicted difficulty demonstrated a higher contribution to the observed difficulty than did NNS for the listening section. Thus, for the New TEPS reading section, the NS- and NNS-predicted difficulties showed similar explanatory powers, whereas for the listening section NS-predicted difficulty clearly accounted much more for the variance of the difficulty than did NNS.

Next, for the TOEIC reading section, NS-predicted difficulty significantly explained 22.9% (*ΔF* = 5.04*), while NNS-predicted difficulty ($R^2$ = 8.3%, *ΔF* = 1.55) did not show a significant explanatory power even when entered in the first step. NS-predicted difficulty also had a significant explanatory power when entered after NNS-predicted difficulty was controlled for ($R^2$ = 30.9%, *ΔF* = 8.14*), and NNS-predicted difficulty also showed a large contribution ($R^2$ = 16.4%) when entered after NS since its significance (0.054) was close to 0.05. This clearly shows that NS-predicted difficulty contributes a stronger significant explanatory power than NNS for the TOEIC reading section.

On the other hand, for the listening section of the TOEIC, both NS- and NNS-predicted difficulty significantly explained 40.7% (*ΔF* = 14.41**) and 34.5% (*ΔF* = 11.07**), respectively, when entered in the first step. However, NS- and NNS-predicted difficulty did not add a significant contribution to the variance of the listening section when controlled for each other. Similar to the TOEIC reading section, NS-predicted difficulty was found to have a stronger explanatory power for the listening section of the TOEIC as well. Therefore, NS-predicted difficulty seems to be able to explain the difficulty of the TOEIC better than NNS-predicted difficulty.

## *Examining the explanatory power of corpus features*

In an attempt to explore the explanatory power of corpus features, multiple regression analyses were conducted for each test. Because of the limited number of passage numbers, the enter method was used instead of the stepwise method. The VIF values of the variables entered in the regression models in this section were all within acceptable limits.

First, the regression analysis results for the New TEPS reading are presented in Table 14. The corpus features significantly accounted for 62.5% of the variance of test difficulty (F (3, 26) = 14.438, $p < .01$). The corpus features entered in the analysis were two syntactic features, including FOG and mean length of sentence, with the subjunctive mood related to pragmatic features. All three of these variables significantly accounted for the New TEPS reading test, indicating that syntactic features of the passage may have influenced the test difficulty.

Next, the results for the New TEPS listening are given in Table 15, and the corpus features demonstrated a significant explanatory power for 55.6% of the total variance (F (4, 13) = 4.07, $p < .05$). The corpus features entered were mostly related to vocabulary, such as the number of tokens and verb variation, with the complex t-unit ratio and clause/sentence as well. The moderately significant variables in the regression model were the complex t-unit ratio and the verb variation, which is comparable to the corpus features entered for the New TEPS reading which had more

Table 14. Multiple linear regression analysis of new TEPS reading observed difficulty (corpus features).

| | Standardized Coefficients | | | | | | |
|---|---|---|---|---|---|---|---|
| Variables | *β* | *t* | *p* | *R* | $R^2$ | *ΔF* | *Sig.ΔF* |
| (Constant) | – | 1.464 | .155 | .791 | .625 | 14.438 | .000 |
| FOG | 1.040 | 6.293 | .000 | | | | |
| Subjunctive Mood | .544 | 4.135 | .000 | | | | |
| MLS | −.543 | −3.520 | .002 | | | | |

Table 15. Multiple linear regression analysis of new TEPS listening observed difficulty (corpus features).

| | Standardized Coefficients | | | | | | |
|---|---|---|---|---|---|---|---|
| Variables | *β* | *t* | *p* | *R* | $R^2$ | *ΔF* | *Sig.ΔF* |
| (Constant) | – | −1.225 | .242 | 74.6 | 55.6 | 4.070 | .024 |
| Complex T-Unit Ratio | .987 | 2.096 | .056 | | | | |
| Verb Variation | .581 | 1.897 | .080 | | | | |
| Token | .041 | .134 | .895 | | | | |
| Clause/Sentence | −.481 | −1.057 | .310 | | | | |

syntactic features. It seems that the difficulty of listening tests is influenced by vocabulary and also the presence of complicated t-units.

Next, regression analyses for the TOEIC reading and listening were conducted. Table 16 shows the results for the TOEIC reading, and the corpus features significantly explained 60% of the observed difficulty ($F$ (3, 15) = 7.488, $p$ < .01). Three features were entered in this model: subjunctive mood, clause/sentence, and F minus. All of these variables were significant, which is a similar result with the New TEPS reading test (see Table 14). This indicates that syntactic features and the subjunctive mood are important features that influence the difficulty in reading tests.

For the TOEIC listening section, several corpus features were entered and they accounted for the largest amount of variance, 63.1%, among the four tests ($F$ (6, 16) = 4.562, $p$ < .01) (see Table 17). The corpus features entered in the regression analysis were the number of lexical types, lexical density, and syllable/word of those related to vocabulary; complex nominal and the number of t-units related to syntactic features; and academic/nonacademic feature related to pragmatic features. The number of lexical types, t-units and academic/nonacademic feature were significant variables in

Table 16. Multiple linear regression analysis of TOEIC reading observed difficulty (corpus features).

| | Standardized Coefficients | | | | | | |
|---|---|---|---|---|---|---|---|
| Variables | *β* | *t* | *p* | *R* | $R^2$ | *ΔF* | *Sig.ΔF* |
| (Constant) | – | 4.885 | .000 | .774 | .600 | 7.488 | .003 |
| Subjunctive Mood | −.443 | −2.605 | .020 | | | | |
| Clause/Sentence | .375 | 2.253 | .040 | | | | |
| F minus | .404 | 2.359 | .032 | | | | |

Table 17. Multiple linear regression analysis of TOEIC listening observed difficulty (corpus features).

| | Standardized Coefficients | | | | | | |
|---|---|---|---|---|---|---|---|
| Variables | *β* | *t* | *p* | *R* | $R^2$ | *ΔF* | *Sig.ΔF* |
| (Constant) | – | .541 | .596 | .794 | .631 | 4.562 | .007 |
| Academic/Nonacademic | .402 | 2.423 | .028 | | | | |
| Lexical Type | .799 | 2.182 | .044 | | | | |
| Lexical Density | −.213 | −.986 | .339 | | | | |
| Syllable/Word | .317 | 1.478 | .159 | | | | |
| Complex Nominal | −.089 | −.339 | .739 | | | | |
| T-unit | −.604 | −2.248 | .039 | | | | |

the regression model. The results of the TOEIC listening are also similar to the New TEPS listening in that most of the variables were related to the number of lexical words and t-unit.

The results of the regression analyses indicate that the corpus features accounted for a considerable amount of variance for the four tests in general, as the explanatory power ranged from approximately 55% to 70%. The features that were significant indicators of the test difficulty were similar for each language skill: syntactic features and the subjunctive mood for the reading test, and lexical and t-unit features for the listening test. Also, when the explanatory power of the NS- and NNS-predicted difficulty and that of the corpus features of each test were compared, a combination of corpus features was capable of explaining a larger variance of the observed test difficulty than the predicted difficulties judged by NSs and NNSs.

### *Examining the explanatory power of combining predicted difficulties and corpus features*

In order to investigate the explanatory power of corpus features in addition to the predicted difficulties, regression analyses were conducted for each test section once more. Because of the limited number of passage numbers, the enter method was used instead of the stepwise method. Along with the predicted difficulties by NSs and NNSs, corpus features that had a significant correlation and may influence item difficulty were entered in the regression model with the observed difficulty as the dependent variable. The VIF values of the variables entered in the regression analyses were all within acceptable limits, except for the NS and NNS predicted difficulty entered in the TOEIC listening test. However, given that the VIF values were slightly above 10 (NS PD = 12.074, NNS PD = 10.726), these variables were not removed.

First, as seen in Table 18 above, the NS- and NNS-predicted difficulty and several corpus features together significantly explained 75.5% ($F$ (5, 24) = 14.758, $p$ < .01) of the observed difficulty of the New TEPS reading section. On the basis of the standardized coefficients $\beta$, FOG had the strongest influence on the dependent variable, the observed difficulty of the New TEPS reading section, followed by MLS, the average length of sentences, subjunctive mood, and the NS- and NNS-predicted difficulty. On the other hand, only the three corpus features, FOG, subjunctive mood, and MLS were significant variables, while the NS- and NNS- predicted difficulty were not. The predicted difficulties were significant variables when entered alone (see Table 12), but lost significance when entered together with corpus features.

Next, for the listening section of the New TEPS, the two predicted difficulty variables combined with corpus features significantly explained 67.1% ($F$ (4, 13) = 6.639, $p$ < .01) of the variance of the observed difficulty (Table 19). Based on the standardized coefficients $\beta$, the NS-predicted difficulty had the strongest influence on the dependent variable, followed by complex t-unit ratio, verb variation, and NNS-predicted difficulty. The variables that were previously included in the regression model composed of corpus features (see Table 15) were excluded as they did not add an additional explanatory power when the predicted difficulties were entered together. Similar to the result of the New TEPS reading, the NS- and NNS- predicted difficulty were not significant variables when entered with corpus features.

**Table 18.** Multiple linear regression analysis of new TEPS reading observed difficulty. (NS, NNS-predicted difficulty and corpus features).

| | Standardized Coefficients | | | | | | |
|---|---|---|---|---|---|---|---|
| Variables | *β* | *t* | *p* | *R* | *R*$^2$ | *ΔF* | *Sig.ΔF* |
| (Constant) | – | .635 | .531 | .869 | .755 | 14.758 | .000 |
| NS-Predicted Difficulty | .249 | 1.071 | .295 | | | | |
| NNS-Predicted Difficulty | .215 | .893 | .381 | | | | |
| FOG | .723 | 4.354 | .000 | | | | |
| Subjunctive Mood | .320 | 2.472 | .021 | | | | |
| MLS | −.457 | −3.454 | .002 | | | | |

**Table 19.** Multiple linear regression analysis of the new TEPS listening observed difficulty. (NS, NNS-predicted difficulty and corpus features).

| | Standardized Coefficients | | | | | | |
|---|---|---|---|---|---|---|---|
| Variables | *β* | *t* | *p* | *R* | $R^2$ | *ΔF* | *Sig.ΔF* |
| (Constant) | | −.707 | .492 | .819 | .671 | 6.639 | .004 |
| NS-Predicted Difficulty | .651 | 1.412 | .181 | | | | |
| NNS-Predicted Difficulty | −.215 | −.496 | .628 | | | | |
| Complex T-Unit Ratio | .464 | 2.810 | .015 | | | | |
| Verb Variation | .283 | 1.503 | .157 | | | | |

The regression analysis results for the TOEIC reading section are given in Table 20 above. The NS- and NNS-predicted difficulties were entered with corpus features and significantly explained 70.6% (*F* (5, 13) = 6.258, $p < .01$) of the observed difficulty of the TOEIC reading items. On the basis of the standardized coefficients *β*, the NS- and NNS-predicted difficulty yielded the strongest influence on the dependent variable, followed by F minus, average number of clauses per sentence, and subjunctive mood. The two reading tests of the New TEPS and TOEIC both included a syntactic feature, a readability index, and the subjunctive mood in its regression model which implies the important features of a reading test that may influence the test difficulty. Unlike the results of the New TEPS tests, the NS- and NNS-predicted difficulties were also moderately significant along with the other corpus features in explaining item difficulty.

Finally, the regression analysis for the TOEIC listening section is given in Table 21. The largest number of features was included for the regression analysis. The NS- and NNS-predicted difficulty along with several other corpus features significantly contributed 74.1% (*F* (8, 14) = 5.003, $p < .01$) of the observed difficulty of the TOEIC listening. On the basis of the standardized coefficients *β*, the predicted difficulty of NSs and that of NNSs followed by the number of t-units and the lexical type feature had the strongest influence. The corpus features explaining the observed difficulty of the TOEIC listening can be classified as lexical features, which included the most variables, such as the average number of syllables per word, lexical type, lexical density, and syntactic features such as

**Table 20.** Multiple linear regression analysis of the TOEIC reading observed difficulty. (NS, NNS-predicted difficulty and corpus features).

| | Standardized Coefficients | | | | | | |
|---|---|---|---|---|---|---|---|
| Variables | *β* | *t* | *p* | *R* | $R^2$ | *ΔF* | *Sig.ΔF* |
| (Constant) | – | 4.716 | .000 | .841 | .706 | 6.258 | .004 |
| NS-Predicted Difficulty | .928 | 2.152 | .051 | | | | |
| NNS-Predicted Difficulty | −.840 | −2.085 | .057 | | | | |
| Subjunctive mood | −.333 | −1.869 | .084 | | | | |
| Clause/Sentence | .344 | 2.227 | .044 | | | | |
| F minus | .357 | 2.172 | .049 | | | | |

**Table 21.** Multiple linear regression analysis of TOEIC listening observed difficulty. (NS, NNS-predicted difficulty and corpus features).

| | Standardized Coefficients | | | | | | |
|---|---|---|---|---|---|---|---|
| Variables | *β* | *t* | *p* | *R* | $R^2$ | *ΔF* | *Sig.ΔF* |
| (Constant) | – | 1.948 | .072 | .861 | .741 | 5.003 | .004 |
| NS-Predicted Difficulty | 1.139 | 2.409 | .030 | | | | |
| NNS-Predicted Difficulty | −.880 | −1.976 | .068 | | | | |
| Academic/Nonacademic | .519 | 3.032 | .009 | | | | |
| Lexical Type | .746 | 2.073 | .057 | | | | |
| Lexical Density | −.358 | −1.636 | .124 | | | | |
| Syllable/Word | .175 | .796 | .439 | | | | |
| Complex Nominal | −.073 | −.308 | .763 | | | | |
| T-unit | −.796 | −2.984 | .010 | | | | |

t-unit, and complex nominal along with the topic-related index of academic/nonacademic. The significant variables of this regression model were the NS-predicted difficulty, academic/nonacademic, and t-unit while NNS-predicted difficulty and lexical type were noteworthy. Both TOEIC tests demonstrated that the predicted difficulty can also significantly account for the variance of the test difficulty even when entered with other corpus features.

To sum up, the regression analyses with predicted difficulties and corpus features entered together demonstrated a strong explanatory power for the observed difficulty of all four tests. For these regression models the NS- and NNS- predicted difficulty were significant variables only for the TOEIC tests. In addition, the amount of variance accounted for by the predicted difficulties and corpus features was similar to or greater than that explained by either the NS- and NNS-predicted difficulties or the corpus features alone.

## Conclusion and implications

The current study attempted to seek ways to better improve the robustness of two non C-R high stakes standardized EFL tests by utilizing corpus features along with NS- and NNS-predicted difficulties. The results of this study may be summed up as follows: (1) Several corpus features and all of the difficulties predicted by NS and NNS experts, except for the NNS-predicted difficulty for TOEIC reading, showed a moderate to high correlation with the observed difficulty. (2) The predicted difficulties of NSs and NNSs did demonstrate a significant explanatory power for the observed difficulty, and NS-predicted difficulty explained the observed difficulty better than NNS-predicted difficulty for all test sections except the New TEPS reading, for which NSs and NNSs showed a similar explanatory power. (3) The corpus features explained a larger variance of the observed difficulty of the four tests than the predicted difficulties. (4) The explanatory power for item difficulty was the strongest when the corpus features were entered with the NS- and NNS-predicted difficulty together, as around 70% or more of the variance of the observed difficulty was explained by these variables.

The corpus features successfully explained item difficulty, which is in line with previous studies that have also looked into the relationship between corpus features and item difficulty (Freedle & Kostin, 1993a, 1993b, 1996; Hamada, 2015; Rupp et al., 2001). Also, the features that played a significant role in explaining item difficulty were mostly related to vocabulary and syntax, which is also consistent with past research results that have pointed to vocabulary as the most influential index category (Crossley et al., 2008; Hamada, 2015). As mentioned earlier, in addition to vocabulary features, including type or token numbers and vocabulary variation features, features related to syntax, such as the average number of clauses within a sentence, and readability features have commonly been found to significantly explain item difficulty. These results indicate that not only vocabulary but also syntactic features may significantly influence item difficulty. However, pragmatic features did not seem to have a large impact on item difficulty except for the subjunctive mood and academic characteristic of the passage.

On the other hand, corpus features that had a noteworthy correlation with NS- and NNS-predicted difficulty were similar but somewhat different from the features that correlated with the observed difficulty. Although most of the related corpus features were common for NSs and NNSs, differences were also found. Moreover, according to the *t*-test, the NS- and NNS-predicted difficulties for all test sections were significantly different, with NNS experts consistently judging the test items as more difficult than NS experts. This implies that NS and NNS judgments of item difficulty are not entirely congruent since they focus on different linguistic and propositional aspects and also judge the test items on different scales of difficulty. Considering that NS-predicted difficulty showed stronger explanatory power for most of the test sections than NNS-predicted difficulty, it would be advisable to take NS-predicted difficulty rather than NNS-predicted difficulty into consideration for difficulty adjustment but also include NNS-predicted difficulty for reference. Thus, even for non C-R tests, expert judgment can still be useful for maintaining a consistent difficulty level.

Another noteworthy finding is that the corpus features demonstrated a stronger explanatory power than the NS- and NNS-predicted difficulty for the tests in question. Thus, it would be reasonable to conduct a corpus analysis in the process of improving the robustness of item difficulty estimation when it is not feasible to employ reviewers when developing tests. Also, the regression analysis results of the predicted difficulty and corpus features together accounted for the largest variance for the tests when compared with the regression results with the predicted difficulty and corpus features entered separately. These results indicate that utilizing both predicted difficulty and corpus analysis would be more desirable for maximizing the overall robustness of test difficulty estimation.

Finally, the fact that the NS- and NNS- predicted difficulty were not significant variables for the New TEPS reading and listening tests when entered together with corpus features needs to be further discussed. The regression analysis results for the TOEIC tests indicated otherwise, as the predicted difficulties remained as significant variables of the test difficulty even when entered with several corpus features. This discrepant result between the New TEPs and TOEIC may be attributed to the time restricted characteristic of the New TEPS among other significant test method facets. That is, since the New TEPS test is a speeded test, the experts may not have been able to consider the pressure that test-takers would experience due to the time restraint when predicting the difficulty of test items. This may have caused the predicted difficulties to lose its significance for the New TEPS tests since the time factor can greatly influence the performance of test-takers. Future research seems necessary to investigate the interrelationship among the relevant variables which may have led to such results.

This study is not without limitations, however, in that there was only one set of test items each for the New TEPS and TOEIC that were analyzed. Because of the limited number of test items, the research results should be interpreted with caution. Further research is needed to collect more data, including both test input and difficulty predictions of experts. This would help to ensure more robust statistical analysis results, especially those of regression analysis, which may also lead to an increased generalizability of the research results. Another issue that should be addressed is the fact that the reading and listening tests in this study have been analyzed using the same tools. Although previous research has indicated that the features that are closely related to the item difficulty of reading and listening tests are in common (Freedle & Kostin, 1996), the two language skills have substantial differences which may lead to discrepant results. As the present study analyzed the reading and listening skills in the same manner, it would be desirable that further studies include features that indicate specific characteristics of each language skill in order to explore the extent to which the results obtained from these studies are in line with the current research.

Although there are a handful of other variables that influence item difficulty, namely, Test Method Facet (item type, prompt, attractiveness of choices, etc.), language subskill components (topic, detail, inference, coherence, etc.), and passage topic, communicative functions, and genre, the current study did not include these variables and limited the scope to only analyzing the linguistic aspects of the test contents through corpus analysis, since conducting a quantitative analysis that included all of these variables was beyond the scope of this restricted research. Further research based on a larger amount of test materials that considers not only the linguistic aspects of the test content but also the aforementioned variables for a more in-depth analysis will significantly contribute to improving the robustness of language tests in general. It would be worth conducting a future study to examine the explanatory power of the predicted difficulty and corpus features more closely. Conducting a comparative analysis of the explanatory power focusing on varying item types and characteristics may provide valuable implications for developing more reliable high-stakes EFL tests.

Overall, the results of the current study were able to confirm the significant explanatory power of corpus features for item difficulty, especially vocabulary and syntax-related features. Also, this research is expected to shed light on how the item difficulty explained by NS and NNS experts can contribute to adjusting the difficulty of tests, which has not been frequently dealt with in previous research but is greatly needed for general language tests that require the robustness of item difficulty as a prerequisite for standard setting. Ultimately, it is hoped that maintaining a consistent difficulty level will be a more plausible and viable task by being able to improve the robustness of content difficulty of tests with the aid of corpus analysis and expert judgment.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## References

Ai, H., & Lu, X. (2010, June 8–12). *A web-based system for automatic measurement of lexical complexity*. Paper presented at the 27th Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10). Amherst, MA.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Anthony, L. (2014). *AntWordProfiler (Version 1.4.1) [computer software]*. Tokyo, Japan: Waseda University.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L., Davidson, F., Ryan, K., & Choi, I.-C. (1995). *Studies in language testing 1: An investigation into the comparability of two tests of English as a Foreign Language*. Cambridge: Cambridge University Press.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Beinborn, L., Zesch, T., & Gurevych, I. (2014). Predicting the difficulty of language proficiency tests. *Transactions of the Association of Computational Linguistics*, *2*(1), 517–529. doi:10.1162/tacl_a_00200

Blau, E. K. (1990). The effect of syntax, speed, and pauses on listening comprehension. *TESOL Quarterly*, *24*(4), 746–753. doi:10.2307/3587129

Choi, I. C. (1994). Content and construct validation of a criterion-referenced English proficiency test. *English Teaching*, *48*, 311–348.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, *42*(3), 475–493. doi:10.1002/j.1545-7249.2008.tb00142.x

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, *54*(5–6), 340–359. doi:10.1080/0163853X.2017.1296264

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating* (TOEFL Research Report 10-29). Princeton, NJ: Educational Testing Service.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185. doi:10.1177/0265532207086780

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing*, *19*(4), 347–368. doi:10.1191/0265532202lt235oa

Freedle, R., & Kostin, I. (1991). *The prediction of SAT reading comprehension item difficulty for expository prose passages* (TOEFL Research Report 91-29). Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1993a). *The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items* (TOEFL Research Report 44). Princeton, NJ: Educational Testing Service.

Freedle, R., & Kostin, I. (1993b). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing*, *10*(2), 133–170. doi:10.1177/026553229301000203

Freedle, R., & Kostin, I. (1996). *The prediction of TOEFL listening comprehension item difficulty for minitalk passages: Implications for construct validity*. TOEFL Research Report 56. Princeton, NJ: Educational Testing Service.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.

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. doi:10.3758/BF03195564

Hale, G. A., Rock, D. A., & Jirele, T. (1989). *Confirmatory factor analysis of the test of English as a Foreign Language*. TOEFL Research Report 32. Princeton, NJ: Educational Testing Service.

Hamada, A. (2015). Linguistic variables determining the difficulty of Eiken reading passages. *JLTA Journal*, *18*, 57–77. doi:10.20622/jltajournal.18.0_57

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, *3*(1), 49–68. doi:10.1016/1060-3743(94)90005-1

Harley, B., & King, M. L. (1989). Verb lexis in the written compositions of young L2 learners. *Studies in Second Language Acquisition*, *11*, 415–440. doi:10.1017/S0272263100008421

Hoshino, A., & Nakagawa, H. (2010). Predicting the difficulty of multiple-choice close questions for computer-adaptive testing. *Special Issue: Natural Language Processing and Its Applications*, 279–291.

Kolen, M., & Brennan, R. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed. ed.). New York, NY: Springer.

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal*, *25*(2), 21–33. doi:10.1177/003368829402500202

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). *Standard setting: A Bookmark approach*. In D. R. Green (Chair), *IRT-based standard setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Loukina, A., Yoon, S. Y., Sakano, J., Wei, Y., & Sheehan, K. (2016). Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3245–3253). Osaka, Japan.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. doi:10.1075/ijcl.15.4.02lu

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208. doi:10.1111/j.1540-4781.2011.01232_1.x

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge: Cambridge University Press.

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. New York, NY: Student Achievement Partners.

Perkins, K., Gupta, L., & Tammana, R. (1995). Predicting item difficulty in a reading comprehension test with an artificial neural network. *Language Testing*, *12*(1), 34–53. doi:10.1177/026553229501200103

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.

Sheehan, K. M., Kostin, I., Futagi, Y., & Flor, M. (2010). Generating automated text complexity classifications that are aligned with targeted text complexity standards. *ETS Research Report Series*, *2010*(2). doi:10.1002/j.2333-8504.2010.tb02235.x

Sydorenko, T. (2011). Item writer judgments of item difficulty versus actual item difficulty: A case study. *Language Assessment Quarterly*, *8*(1), 34–52. doi:10.1080/15434303.2010.536924

Templin, M. (1957). *Certain language skills in children: Their development and interrelationships*. Minneapolis: The University of Minnesota Press.

# Appendix

Correlation Analysis of 53 Corpus Features with Observed Difficulty

| | | New TEPS | | TOEIC | |
|---|---|---|---|---|---|
| | Corpus Feature | R | L | R | L |
| 1 | Token | 0.234 | .547* | 0.239 | 0.356 |
| 2 | Sophisticated Word Tokens | 0.219 | .518* | 0.312 | 0.338 |
| 3 | Lexical Token | 0.267 | .513* | 0.271 | .434* |
| 4 | Sophisticated Lexical Tokens | 0.220 | .514* | 0.318 | 0.351 |
| 5 | Type | 0.245 | .504* | 0.228 | .462* |
| 6 | Sophisticated Word Types | 0.288 | .486* | 0.285 | 0.370 |
| 7 | Lexical Type | 0.289 | .522* | 0.254 | .498* |
| 8 | Sophisticated Lexical Types | 0.307 | .508* | 0.298 | 0.383 |
| 9 | Number Of Different Words | 0.245 | .504* | 0.228 | .462* |
| 10 | Syllable/Word | .483** | 0.234 | .474* | .551** |
| 11 | TTR | −0.253 | −0.304 | −0.227 | 0.212 |
| 12 | Standardized TTR | −0.262 | −0.197 | −0.129 | .422* |
| 13 | Lexical Word Variation | −0.115 | 0.137 | 0.058 | .453* |
| 14 | Verb Variation 1 | 0.343 | 0.451 | 0.168 | .423* |
| 15 | Verb Variation 2 | 0.108 | −0.232 | −0.337 | −0.051 |
| 16 | Noun Variation | −0.240 | −0.096 | −0.221 | −0.088 |
| 17 | Lexical Density | 0.139 | 0.135 | 0.299 | 0.391 |
| 18 | Lexical Sophistication | −0.032 | 0.383 | 0.208 | 0.161 |
| 19 | Word Concreteness | −0.065 | 0.149 | −0.232 | −0.148 |
| 20 | Sentence | 0.195 | 0.177 | −0.002 | 0.119 |
| 21 | Verb Phrase | 0.191 | 0.268 | 0.047 | 0.047 |
| 22 | Clause | 0.230 | 0.404 | 0.167 | 0.078 |
| 23 | T-Unit | 0.179 | 0.223 | 0.049 | 0.009 |
| 24 | Dependent Clause | 0.190 | 0.291 | 0.359 | 0.108 |
| 25 | Complex T-Unit | 0.092 | 0.237 | 0.321 | 0.077 |
| 26 | Coordinate Clause | −0.164 | 0.382 | 0.283 | 0.218 |
| 27 | Complex Nominal | 0.205 | 0.237 | −0.059 | .475* |
| 28 | Syntactic Complexity | 0.285 | 0.187 | −0.134 | 0.009 |
| 29 | Voice | 0.135 | 0.268 | −0.189 | −0.030 |
| 30 | Mean Length Sentence | 0.010 | 0.350 | 0.117 | 0.221 |
| 31 | Mean Length T-Unit | 0.019 | 0.236 | −0.013 | 0.383 |
| 32 | Mean Length Clause | −0.169 | −0.093 | −0.200 | 0.202 |
| 33 | Clause/Sentence | 0.112 | .478* | 0.391 | 0.017 |
| 34 | Complex T-Unit Ratio | 0.114 | .508* | 0.278 | 0.212 |
| 35 | Dependent Clause Ratio | 0.023 | 0.088 | 0.331 | 0.124 |
| 36 | Coordinate Phrases Per Clause | −0.325 | 0.160 | 0.117 | 0.180 |
| 37 | Sentence Coordination Ratio | 0.027 | 0.203 | 0.312 | −0.271 |
| 38 | Complex Nominals Per Clause | −0.119 | −0.162 | −0.210 | 0.306 |
| 39 | Complex Nominals Per T-Unit | 0.001 | 0.014 | −0.101 | 0.400 |
| 40 | Verb Phrases Per T-Unit | 0.040 | 0.128 | 0.078 | 0.049 |
| 41 | F-Minus | .475** | 0.254 | .565* | .519* |
| 42 | FOG | .488** | 0.258 | 0.364 | 0.287 |
| 43 | Narrativity | −0.295 | −0.200 | −0.383 | −0.265 |
| 44 | Referential Cohesion | 0.106 | −0.221 | 0.293 | −.462* |

(*Continued*)

(Continued).

| | | New TEPS | | TOEIC | |
|---|---|---|---|---|---|
| | Corpus Feature | R | L | R | L |
| 45 | Deep Cohesion | −0.183 | 0.037 | −.572* | 0.292 |
| 46 | Verb Cohesion | −0.177 | −0.152 | 0.313 | −.524* |
| 47 | Connectivity | 0.082 | −0.373 | 0.254 | −0.148 |
| 48 | Temporality | 0.056 | −0.145 | 0.219 | −0.128 |
| 49 | Text Concreteness | −0.179 | 0.025 | −0.257 | −0.077 |
| 50 | Negation | 0.175 | 0.306 | −0.267 | −0.374 |
| 51 | Subjunctive Mood | 0.226 | −0.323 | −.507* | 0.238 |
| 52 | Non/Academic Context | .369* | 0.032 | 0.112 | .472* |
| 53 | Academic/Practical Context | −0.069 | −0.203 | - | - |