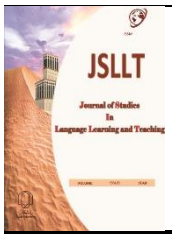
E-ISSN: 2981-1686



Journal of Studies in Language Learning and Teaching

www.jsllt.yazd.ac.ir

JSLLT, Vol. 1, No. 2, 211–225.

Paper Type: Original Article

A Corpus-based Evaluation of a High-stakes EFL Exam

Elaheh Rafatbakhsh[1,*], Alireza Ahmadi[2]

[1] Department of Foreign Languages and Linguistics, Shiraz University, Shiraz, Iran; e.rafatbakhsh@shirazu.ac.ir;
[2] Department of Foreign Languages and Linguistics, Shiraz University, Shiraz, Iran; arahmadi@shirazu.ac.ir;

Received: 30 December, 2023 Revised: 25 April, 2024 Accepted: 11 July, 2024

Abstract

High-stakes assessments play a significant role in people's lives, and their results greatly define individuals' future social and financial prospects. Corpus linguistics has recently been used to inform the development and validation of such tests. This study aimed at identifying the degree of typicality of vocabulary items tested in the English proficiency subtest of the Master of Arts/Science Iranian University Entrance Exam. To this end, the vocabulary options and collocations in 20 test versions were extracted, and their frequency of occurrence in the Corpus of Contemporary American English was examined using a specially written computer program. The results indicated that the frequency of the options in the academic genre was not as dominant as expected in a test designed for academic purposes. The findings also revealed some inconsistencies among the different parallel test versions in terms of their option frequencies. Furthermore, for some options and collocations, atypicality was observed as zero or close to zero instances in the corpus. The current study suggests the inclusion of frequency information from corpora and various wordlists to accompany test developers' intuition for more robust vocabulary assessment.

Keywords: Corpus linguistics, High-stakes exam, Lexical coverage, The Corpus of Contemporary American English (COCA), Vocabulary assessment.

I | INTRODUCTION

Tests in general serve different purposes for learners, teachers, parents, and policymakers (Bovaird et al., 2011). Whether a test is low or high-stakes depends on the importance of the decisions made about test takers based on their test results (Shohamy, 2001). High-stakes test results are used to make significant decisions about test takers or others in the educational system including administrators, teachers, schools, and communities. However, their interpretation and level of importance vary in different settings (Shohamy, 1996; Davis, 2006; Bovaird et al., 2011; Lin & Gao, 2020).

Assigning grades to test takers and classifying them based on their proficiency levels, comparing their scores, and accepting or rejecting them are performed based on tests and their results. That is why, to increase their scores on high-stakes tests, test takers go through a substantial amount of test preparation activities (Gebril & Eid, 2017). So high-stakes tests function as national policies which, upon their introduction, affect the way teachers teach (Alderson & Wall, 1993; Shohamy, Donitsa-

Corresponding Author: e.rafatbakhsh@shirazu.ac.ir
10.22034/jsllt.2024.21052.1030

Schmidt & Ferman, 1996, Larsson & Olin-Scheller, 2020) and, therefore, exhibit strong washback effects (Shohamy et al., 1996).

With recent advances in technology, corpus linguistics can help to enhance the development and validation processes of such important tests. Given that, this study aims at using corpus linguistics to assess the content of an Iranian national test as a high-stakes test, namely the Master of Arts/Science Iranian University Entrance Exam.

1. Corpus Linguistics and Test Validation

Language education is one of the fields that has benefitted the most from the study of corpora. Johansson (2009) has summarised the many uses of corpora in connection with language teaching including the use of corpora in classroom activities, testing, basic research, syllabus design, and developing teaching materials, textbooks, grammars and dictionaries.

The role of corpora in assessment was first introduced by Alderson (1996). However, the potential uses of corpus linguistics in language assessment have remained underexplored compared to other aspects of language education. How the study of corpora can be usefully applied to test construction, scoring, and validation is still under investigation (Taylor & Barker, 2008; Egbert, 2017; Pan & Qian, 2017).

Studies have shown that the development of language tests and automatic item generation systems has been informed by corpora in different ways (e.g., Miktov & Ha, 2003; Brown et al., 2005; Lin et al., 2007; Rafatbakhsh et al., 2021). For instance, corpora are used in task and item design to decide on the critical language features of different proficiency levels, language features responsible for the difficulty of a reading text or a listening section, and plausible distractors based on the learners' error types (Cushing, 2017). Additionally, information from corpora can help to score tests as well as develop scoring scales using linguistics features, frequencies, and difficulty indices (Crossley et al., 2011; Isaacs et al., 2018; Monteiro et al., 2020).

Moreover, some studies have utilized corpora to validate different types of language tests assessing various skills and subskills (e.g., Staples et al., 2018; Beigman Klebanov et al., 2019; Crosthwaite & Raquel, 2019). Corpus linguistics is used to determine the authenticity of test materials in validation processes, (Biber et al., 2002). Authenticity is defined by Bachman and Palmer (1996) as "the degree of correspondence of the characteristics of target language use (TLU) tasks and test tasks" (p. 23). Staples et al. (2018) have summarized different ways of comparing tests and TLU contexts as a part of validation processes. These ways are as follows:

> (a) comparisons of the score level on a test and the holistic scores on the texts/speech produced by the same participant in the TLU, (b) comparisons of the score level on a test and the scores reflecting broader performance in the TLU, (c) comparisons of the language used in a test and the language used in the target domain by the same participants, (d) comparisons of the language used in a test and the language used in the target domain more generally (p.2)

In this regard, corpus linguistics is employed to compare and analyze two or more corpora to investigate the degree of the correspondence between the characteristics of test tasks and the characteristics of target language tasks. The concept of content typicality is introduced to refer to frequently occurring instances of language in the reference corpora (Pan & Qian, 2017). Pan and Qian also used the term "atypical" for instances with extremely low or zero frequencies in corpora. According to them, grammaticality does not necessarily mean correctness; an item can be grammatical yet unusual and unnatural in a language. The topics of authenticity and typicality gain more importance in contexts where English serves as a second or foreign language. Systematic study of corpora makes it possible to determine the degree of typicality using information such as frequency. Therefore, corpus data can complement test developers' intuition. Test developers can be well aware of the lexico-grammatical

characteristics of native speakers' language usage by studying corpora. The information on the relative frequency of words and phrases, patterns, collocations and colligations, grammatical constructions, formulaic expressions, and lexical sequences is what test constructs emerge from (Park, 2014).

Some studies have used frequency lists and other information from corpora to measure test takers' various aspects of vocabulary knowledge at different levels (e.g., Beglar & Nation, 2007; Sasao & Webb, 2017) or assess their writing (e.g., Goodfellow et al., 2002; Staples et al., 2018). A few studies have also examined the content of multiple-choice items to see if they are actually used in real-life language (e.g., Weir & Milanovic, 2003; Bai, 2005). In this context, Paribakht and Webb (2016) evaluated the academic vocabulary coverage and its relationship with test takers' scores in a standardized test. Studying 12 versions of an English proficiency test used for admission purposes at Canadian universities revealed that the coverage of the Academic Word List (AWL) in the tests was substantial. In another study, Vu (2019) focused on the coverage of GSL (Geneal Service List; West, 1953) and AWL to examine the lexical profiles of university admission and high-school graduation exams in Vietnam. The results showed a mismatch between the policies and the practice as the lexical demands of the exams were far more than the set target. Also, Pan and Qian (2017) measured the frequencies of the tested items in a standardized test as a part of the validation process. According to the findings, content typicality was problematic as some grammar items did not conform to native language production. However, to the best of the researchers' knowledge, no research has been conducted to evaluate the content typicality of vocabulary items in high-stakes language proficiency tests through the use of a comprehensive corpus such as COCA. The findings of such a study can be useful for test validation by providing evidence about the relevance of the vocabulary used in the test to the target domain.

2. Study Context

This study is conducted in the Iranian EFL context where large-scale high-stakes tests play a significant role in test takers' lives. The Iranian National University Entrance Exams are run to screen candidates for admission into universities at the three levels of Bachelor of Arts/Science (BA/BS), Master of Arts/Science (MA/MS), and Doctor of Philosophy (PhD).

To pursue undergraduate studies in Iranian public universities, on average about 1.1 million candidates annually take part in the National University Entrance Exam for BA/BS. Also, to gain admission to higher education, on average 800,000 and 200,000 candidates participate in the National University Entrance Exam for MA/MS and PhD, respectively (Iranian National Organization for Educational Testing, 2020).

The focus of this study is on the English proficiency subtest of the Iranian National University Entrance Exam for MA/MS. Similar to the other two exams, MA/MS entrance exam is in the form of a multiple-choice test with a negative score system to avoid guessing and random answering by candidates. According to this scoring system, for every three wrong answers, one correct answer will be eliminated. Table 1 depicts the number of test takers for this test in the years 2015-2019.

Table 1. The number of candidates in the Iranian MA/MS Entrance Exam.

Year	2015	2016	2017	2018	2019
Participants	813013	761273	878388	735734	614833

In the MA/MS entrance exam, the English proficiency subtest assesses test takers' general language proficiency. This subtest includes multiple-choice items assessing vocabulary, grammar, and reading comprehension. No specific sources are determined for test preparation for this exam. So, students are expected to enhance their overall language proficiency to sit and pass this test successfully. They can also have access to the previous versions of the test to increase their preparation for the test.

There have been some validation studies on different Iranian National Exams exploring washback effect, differential item functioning, and content and construct analysis (e.g., Ravand et al., 2008; Ahmadi &

Thompson, 2012; Razavipur, 2014; Ahmadi et al., 2015; Ravand & Firoozi, 2016; Bazvand et al., 2019). However, to the best of the researchers' knowledge, no study has focused on the content typicality of these tests so as to see whether the words and phrases tested are used in real language contexts.

Taking into account the above gap in the literature, this study aimed at examining the content of vocabulary items of the general English proficiency subtest of the Iranian MA/MS Entrance Exam in terms of typicality. To examine the content typicality, the frequency of all the options of the vocabulary items was extracted from the five genres of COCA, and the results were compared across different genres and years. In addition, the collocations found in the items were searched in the corpus to find out their degree of typicality. Collocations were formed using the combination of words from the stems and options. With these objectives in mind, we addressed the following research questions on content typicality in this study.

2. Research Questions

1. How frequent are the vocabulary options of the General English subtest of the MA/MS Iranian National University Entrance Exam, overall and across different genres?
2. Are there any differences in the degree of typicality of vocabulary options among different versions of the test over the five years from 2015 to 2019?
3. How typical are the collocations extracted from each item?

II. METHOD

1. Reference Corpora

The research was based on COCA, created by Davies (2008), Professor of Corpus Linguistics at Brigham Young University. This corpus is the only comprehensive genre-balanced corpus of American English, composed of more than one billion words of text from 1990 to 2019. For this study, a purchased version of the corpus was used with more than 520 million words in 220,225 texts, including 20 million words each year from 1990 to 2015. This corpus is divided evenly among five genres including spoken genre, fiction, popular magazines, newspapers, and academic journals gathered from various authentic sources.

2. Data

The data for this study were collected from the vocabulary section of the general English proficiency subset of the MA/MS Iranian National University Entrance Exam in five years, from 2015 to 2019. The MA/MS exam is administered to candidates of 140 fields of study. Every year, parallel versions (usually seven) of the test are designed and randomly assigned to candidates of different fields except for foreign languages (such as English, German and French) which have their own specific versions of the test. As the tests become publicized after administration, new tests are developed annually.

The vocabulary section of the test comprises 30 multiple-choice items, including 10 vocabulary items, a cloze passage of five items on grammar, and two reading passages each with five comprehension items. For the English language studies group, consisting of the three fields of English Literature, Teaching English as a Foreign Language, and Linguistics, this test version is different and has 60 items including 10 structure items, 20 vocabulary items, one cloze passage of 10 items, and three reading passages with 20 items.

For this study, four versions of the test were selected, including three versions randomly selected from the seven available parallel versions and one version specifically designed for the candidates of English language studies. Therefore, a total of 250 multiple-choice vocabulary items were studied from the

selected MA/MS entrance exams. As a result, 1000 options and 1000 collocations formed with the options were extracted from the items to be searched in the corpus for their frequencies.

3. Data Collection and Analysis

To find out the degree of the typicality of the tested vocabulary items, their frequencies were checked in COCA. First, all the four options (the key and the three distractors) of all the items were extracted. Then, the possible collocations were formed, including words from the stem with the answer and with each distractor. For instance, the options "eccentric, equivocal, exuberant, and exorbitant" and the possible collocations that the options could form, i.e., "eccentric prices, equivocal prices, exuberant prices, and exorbitant prices", were listed as follows:

Item: Although no one was interested in buying Vincent Van Gogh's paintings during his lifetime, they now sell for ……. prices.

1) eccentric 2) equivocal 3) exuberant 4) exorbitant

After that, the frequency of these options was counted in COCA. Considering the objective of the study, writing a specific script for this particular purpose was preferred to using the existing concordancers. Therefore, to find out the frequency of each option and the collocations in COCA, a script was written by a computer programmer. The designed concordancer made it possible to search a long list of 2000 words and collocations in COCA within a short time. All the lemmas of the verbs and the different variations of the phrases were also included in the search. Comparisons were then made among the exam versions and years across the five corpus genres using descriptive statistics.

III. RESULTS

In line with the first research question concerning the frequency of the options, all the options in the items of the selected test versions (250 multiple-choice vocabulary items) were extracted and searched in the corpus. Table 2 reports the total frequency of the options per million. As it can be seen, the exam years 2018 and 2015 had the highest and lowest frequencies of the options, respectively, with a noticeable difference between the two.

Table 2. Overall option frequencies in each exam year.

Year	2015	2016	2017	2018	2019	Total
Frequency (pm)	2762.69	5031.22	2920.57	13514.18	5334.85	29563.51

Additionally, the frequencies of the options in each genre of COCA were counted separately, the results of which are demonstrated in percentages in Figure 1. As it can be seen, the overall coverage of the options is the highest in the spoken genre (23.22%). The academic (21.16%), magazine (18.89%), fiction (18.76%) and news (17.99%) genres occupy the next ranks, respectively. However, it seems that the highest frequencies of the options in each year belong to the genre of academic, except in 2018. In this year, the percentage of the frequencies in the spoken genre is the highest, i.e., 27.65%. Since the 2018 version had the highest frequency among the five versions, it has affected the total results considerably. So, while the academic genre enjoys the highest frequency in four versions, in the 2018 version and the total results, the spoken genre appears with the highest frequency.

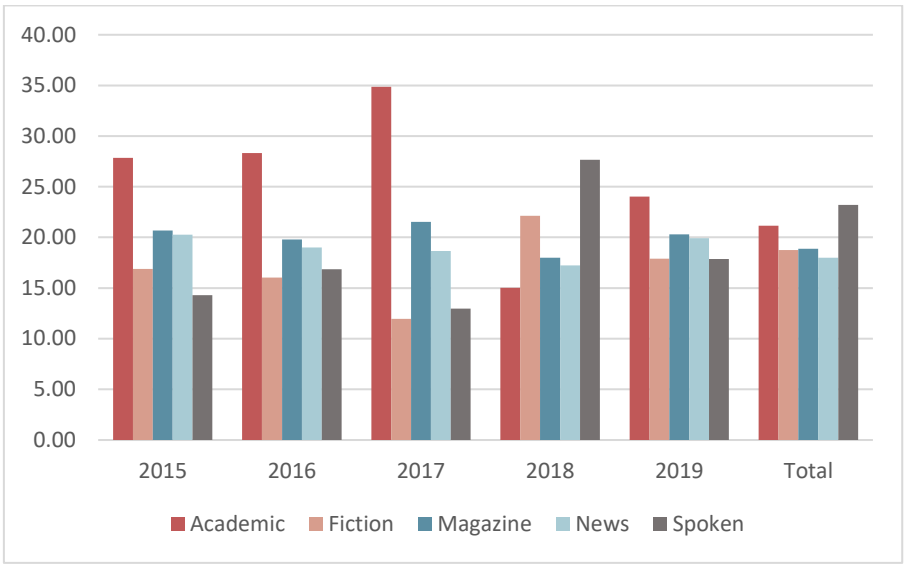


Figure 1. The percentage of the option frequencies across the genres.

To answer the second research question, concerning the typicality of the vocabulary options in different test versions, the overall frequencies of the options in each test version of each year were extracted across the genres. The total frequencies were then calculated for each test version and the percentages of the frequencies in each genre were computed accordingly. Table 3 presents the results.

Table 3. The percentages of the option frequencies in different test versions across the genres.

	Field	Academic	Fiction	Magazine	News	Spoken
2015	Version 1	26.96	13.25	21.46	21.43	16.90
	Version 2	42.37	12.54	16.29	16.48	12.31
	Version 3	24.89	25.06	22.97	16.40	10.67
	Eng. version	19.88	19.03	20.71	28.19	12.20
2016	Version 1	20.11	19.35	20.40	19.32	20.82
	Version 2	35.68	11.21	20.39	18.99	13.74
	Version 3	41.85	10.29	20.08	16.06	11.72
	Eng. Version	23.38	16.76	23.06	24.15	12.65
2017	Version 1	33.40	11.49	23.34	18.51	13.27
	Version 2	36.16	10.66	21.73	18.68	12.77
	Version 3	35.30	11.04	21.63	18.33	13.70
	Version Eng.	25.37	18.93	24.55	19.57	11.58
2018	Version 1	33.94	13.23	19.41	18.96	14.46
	Version 2	35.12	12.53	20.40	18.33	13.62
	Version 3	12.03	22.81	18.32	16.76	30.08
	Eng. Version	30.03	12.14	22.58	18.69	16.55
2019	Version 1	40.33	11.52	20.81	16.72	10.62
	Version 2	38.30	11.72	19.55	17.32	13.11
	Version 3	20.26	18.93	20.80	20.08	19.92
	Eng. Version	28.29	10.98	24.53	23.83	12.37

As indicated, the largest percentage of the option frequencies in all the test versions belonged to the academic genre except for six test versions, highlighted in Table 3. In these tests, the options were more frequent in the fiction, magazine, news and spoken genres. As it means, in 70% of all the cases, the academic genre had the highest frequency. In 30% of the cases, the other genres were the most frequent, and the academic genre had the second rank (in five out of six cases). Exceptionally, in test version 3 in 2018, 30% of the overall frequencies belonged to the spoken genre, and the lowest percentage (12.03%) belonged to the academic genre.

Furthermore, minimum, maximum, and total frequencies were extracted for each test version in each year (Table 4). In the English language studies, the minimum, maximum and overall frequencies were lower than the other versions. However, no specific patterns can be traced in the statistics extracted for

the three parallel test versions. The vocabulary coverage of these three versions in COCA was not similar when compared in each year or across different years. The total frequency of 11843.11 pm was the highest among all in the test version 3, 2018, while the lowest was 203.8 pm in 2019 for the options tested in the field of English language studies.

Table 4. Minimum, maximum and total frequencies (pm) of the options in each test.

	Field	Min	Max	Totals
2015	Version 1	0.33	235.43	1666.16
	Version 2	0.02	52.99	225.98
	Version 3	0.16	189.02	715.24
	Eng. Version	0.02	25.89	155.31
2016	Version 1	0.1	1601.45	2757.37
	Version 2	0.23	246.44	1244.13
	Version 3	0.05	135.99	849.9
	Eng. Version	0.02	26.84	179.82
2017	Version 1	0.05	101.22	694.05
	Version 2	0.01	49.48	434.8
	Version 3	0.75	222.63	1550.23
	Eng. Version	0	27.63	241.49
2018	Version 1	1.11	154.97	580.11
	Version 2	0.22	150.8	809.5
	Version 3	0.03	6128.04	11843.11
	Eng. Version	0	40.37	281.46
2019	Version 1	0.31	221.85	571.5
	Version 2	0.23	201.76	441.35
	Version 3	0.33	1601.45	4118.74
	Eng. Version	0	46.76	203.8
Total Mean	Versions 1, 2, & 3	0.26	752.90	1900.14
	Eng. Version	0.00	33.49	214.52

On the other hand, there were just three options in the 20 tests which had absolute zero occurrences in COCA. The options were "nefandous", "saporous" and "containerport". They were all in the items belonging to the fields of English language studies in 2017, 2018, and 2019, respectively. In addition, the total number of options with less than 10 instances in the whole corpus was 13, all belonging to the vocabulary test for the fields of English language studies except for one which was in the test version 2, 2016.

To answer research question 3 on the typicality of the collocations, combinations of words from the stem and options were extracted for all the items. This yielded all the six main types of collocations consisting of adjective + noun, noun + noun, verb + noun, adverb + adjective, verb + prepositional phrase, and verb + adverb. The collocations were then searched in COCA for their frequencies. Table 5 shows the overall results.

Table 5. Collocation frequencies (pm) in different tests across genres.

Year	Academic	Fiction	Magazine	News	Spoken	Total
2015	8.48	2.43	4.12	3.41	1.61	3.99
2016	15.85	18.96	17.35	17.33	18.69	17.64
2017	72.14	13.25	30.86	26.34	17.03	31.76
2018	8.76	5.61	7.83	8.21	7.93	4.46
2019	38.63	7.92	14.18	11.05	8.57	15.97

The frequencies of the collocations were the highest in the academic genre in all the exam years except in 2016 in which the highest frequency was found in the fiction genre. Besides, the collocations were the most frequent in total in 2017, i.e., 31.76 pm.

In the search for the collocation frequencies, some very low or even zero frequencies were found, showing atypical examples. To detect the degree of atypicality in the collocations, the ones with absolute zero frequency were extracted. Figure 2 shows the percentages of the collocations formed with the key options with absolutely zero instances in the whole corpus.

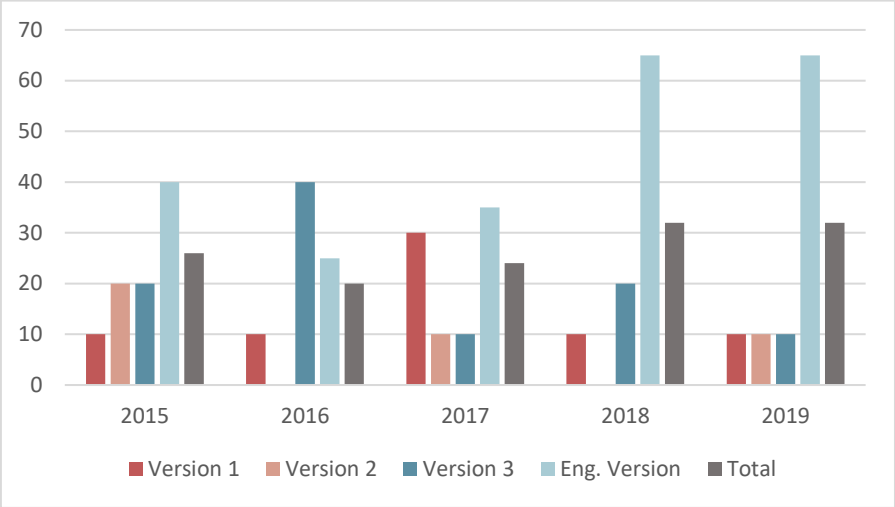


Figure 2. The percentages of the answer collocations with an absolute zero frequency.

The bar chart indicates the zero frequencies across different test versions in different years. The percentages of the answer collocations with a zero frequency are the highest in the test version for English language studies in all the years except in 2016 where test version 3 had the highest percentage of zero frequency collocations (40%). In general, in 2018 and 2019, 32% of the answer collocations could not be found in COCA at all, while, in 2016 the percentage was the lowest at 20%. In contrast, test version 2 included no answer collocations with a zero frequency in 2016 and 2018.

Besides the answer collocations, the frequencies of the distractor collocations in COCA were extracted. There were some instances where the frequency of the distractor collocations was higher than that of the answer collocations and, therefore, more typical in an item. In the years 2015, 2016, 2017, 2018 and 2019, there were respectively 2, 2, 6, 2 and 3 cases in which the distractor collocations showed a higher degree of typicality than the answer collocation. In these items, the collocations that were not much frequent in the corpus were used as the correct answer.

IV. DISCUSSION

In this study, the degree of the typicality of vocabulary items was investigated for four versions of the English proficiency subtest of the MA/MS Iranian National University Entrance Exam. To address the first research question, all the options in the vocabulary subsets were searched in COCA. Differences were depicted among different versions of the test across time. Then, the frequencies were compared across the five genres of COCA for each year. As the most marked point emerging from the data comparison, the academic genre enjoyed the highest frequency in the four years. However, in one year and in the total results, the spoken genre appeared to have the highest frequency.

The items used to assess English proficiency in university entrance exams are usually developed to evaluate whether the test takers are proficient enough to manage the linguistic demands of the academic materials in the academic context of university. Although, in many fields of study, except for language studies, the main language used at Iranian universities is Persian, students face a great number of materials in English. Moreover, the language of most scholarly journals is mainly English. Students are normally required to study the recent articles related to their fields, and they sometimes write papers in

English and publish them in international journals. Therefore, the screening tests should be in line with what is required later in the academic context of universities.

For further analysis, we examined the occurrence of items from the Academic Word List (AWL) (Coxhead, 2000) and the Academic Vocabulary List (AVL) (Gardner & Davies, 2014) in all the options. The results indicated that 194 (19.4%) and 336 (33.6%) options belonged to the AWL and AVL, respectively. Paribakht and Webb (2016) found that 71 out of 144 options (49.30%) of the multiple-choice cloze test of CanTEST (an English language proficiency test used in Canada for university admission purposes and professional certification) existed in the AWL. The academic vocabulary coverage in the current study is much lower. The findings generally reveal that the academic genre is not adequately represented in the MA/MS Iranian exam as expected by the function of this test. This points to the lack of systematic attention to the use of corpora in designing the test.

Research question 2 was concerned with the comparison of different test versions across different genres during the five years. Given the purpose of the tests, it is reasonable to expect that the vocabulary options show the highest frequency in the academic genre. While the results showed that the academic genre had the highest frequency in the majority of cases, the difference depicted was considerable in no case. In some cases, the academic genre was of a similar frequency as the other genres. Also, in about one third of the cases, it was even of lower frequency. The inclusion of the vocabulary frequent in genres other than academic for the fields of English language studies is justifiable because the candidates' overall language proficiency is higher than other fields. Moreover, the language proficiency requirements for their university studies differ from other candidates'. That is, students of language studies are expected to be proficient in all language skills because language is both the medium of instruction and an objective to be studied for them. However, for the students of other fields, reading is the basic language skill needed as they are mostly required to read foreign sources in English. So, while the findings are more or less supported for English language studies, the results are not promising for the other fields and show the need for the inclusion of the academic genre in a more consistent way in this high-stakes test.

Analysis of the test versions in terms of minimum, maximum and mean scores for the frequency of using the options also supported this finding. More differences were found among different versions than what was expected for such parallel tests. One reason for this finding could be that some items aimed to assess the knowledge of collocations or expressions, and the options were mostly chosen from among common verbs such as have, take, make or give, that are more meaningful when they are a part of a collocation.

Furthermore, the fields of English language studies contained less frequent vocabulary as anticipated. As explained earlier, that is because candidates whose major is English are more proficient in different language skills compared to students of other fields. Therefore, the occurrence of low-frequency words or expressions was not unexpected. There are still some doubts on the relationship between word frequencies and the difficulty level; however, many studies suggest that lexical processing depends on the frequency of the words, i.e., high-frequency words are processed faster and more accurately than low-frequency ones, both in the first and second languages (e.g., Laufer et al., 2004; Schmidtke, 2014; Akbari, 2016; Chen et al., 2018). For instance, in a study by Culligan (2015), from among different methods of estimating word difficulty such as frequency, length of word, and the number of syllables, the log of word frequencies extracted from large corpora represented the best estimate of difficulty level and lexical familiarity. The correlation of the word frequency and the difficulty was negative; as the frequencies decreased, the difficulty augmented. Moreover, a study by Choi and Moon (2020) revealed that expert judgment and corpus features cover various aspects of item difficulty. Therefore, frequencies, as one of the influencing factors in word familiarity, should be considered in designing a test.

The frequency in the tests designed for the candidates of English language studies ranged from 0 to 46.76. There were instances where the options had zero or very low frequencies in the corpus, which means they were atypical. The fact that some of the options used in vocabulary assessment were not found in one of the biggest language representatives, i.e. COCA, can question the overall validity of this proficiency test.

Although the test is intended for candidates of English language studies, the use of words such as nefandous, saporous and containerport with absolute zero frequency can be highly controversial. These words were non-existent in the British National Corpus (BNC) as well. Some of the other words with less than 10 instances in the COCA included: defalcate, clamant, banausic, animadversion, torpidity, perorate, depredate, paralogism, and acidulous. Among these words, clamant, animadversion, and depredate were the key options, while the others were used as distractors. These words are not practical in the English language and should be substituted with more practical and functional words.

To answer research question 3 on the typicality of collocations, the possible collocations were manually extracted from the items, formed both with the answer and the three distractors. Their frequencies of occurrence were then extracted from COCA. The findings revealed that the overall frequencies of the collocations differed remarkably across different years and versions. More similar statistics were expected for such parallel tests. Also, although the highest frequency of collocation in the academic genre for each year matched the logic for these tests, the problem was that the frequencies were not necessarily high or different from other genres'. This means the tests were not informed by corpus linguistics.

Despite the grammaticality of the collocations, there were no traces of some collocations in the corpus. There are possible grammatical ways of forming a phrase; however, not all of them sound natural or near-native (Pawley & Syder, 1983). The corpus-based study of collocations is of paramount importance in that they are connected to natural and fluent language production by native speakers (Sinclair, 1991; Ellis, 2002; Schmitt, 2012). Since there are no specific sources for these tests to study, we argue that students may, among other sources, tend to study previous published test samples in their preparation for this test; therefore, the tests simultaneously serve as both testing and teaching materials for the candidates. This calls for further attention to the content development in such tests, and corpus linguistics can be of great importance to this end.

A growing body of literature has evaluated the use of corpora in vocabulary selection for teaching and testing materials, and different threshold levels have been introduced. According to the results of the preliminary work in this field, for university students to read academic texts, a core vocabulary of 3000 word families is essential (Laufer, 1992). Hazenberg and Hulstijn (1996) studied the lexical coverage of university textbooks and placed the threshold at 10,000 word families. More recent evidence suggests 14,000 word families for reading university textbooks (Chujo & Hasegawa, 2003), 8000 to 9000 word families for unassisted comprehension of authentic written texts, and 6000 to 7000 word families for spoken texts (Nation, 2006). Overall, researchers agree that frequencies extracted from corpora can be considered as a yardstick to select teaching and testing materials. The words with frequencies close to 0 per million, which were included in the vocabulary tests of the current study as multiple-choice options, did not belong to any of the aforementioned word families. We believe that the inclusion of words with zero occurrences in the materials used in EFL contexts may not be entirely reasoned. In the development of teaching and testing materials, more systematic data are required besides the experts' intuitions. Even native speakers' judgment cannot be a good yardstick. According to Okamoto (2015), there is a positive correlation between the native speakers' self-reported frequency of word use and the word frequency in corpora only up to the 7000-word level. Above this threshold, native speakers cannot easily categorize the given words. This signifies the value of corpora in designing tests.

V. CONCLUSIONS AND IMPLICATIONS

The current study sought to evaluate the content of the vocabulary items in four versions of the English general proficiency subset of the master's university entrance exam from 2015 to 2019 in Iran. A corpus-based approach was adopted to assess the typicality of the vocabulary tests with the frequencies extracted from COCA. It can be concluded from the findings that the development of vocabulary items in this

test is not informed by the data from corpora or existing wordlists, and genres are not taken into consideration in the process either.

In a major high-stakes exam such as a university entrance exam for master's programs, more academic vocabulary is expected to be observed than the vocabulary from other genres because this test is intended to assess the candidates' academic language capabilities for entering university. Additionally, the vocabularies with zero or extremely low frequencies might be considered unsuitable as they are not essential vocabulary and do not facilitate further studies. Therefore, considering the fact that corpora can manifest information about content typicality, developing and validating vocabulary tests can benefit greatly from corpus linguistics. Checking the typicality of the words tested in an exam can result in more systematic and higher-quality assessments. Test developers are, therefore, advised to refer to well-established corpora as target language representatives to complement the experts' judgments when developing tests. Currently, with the availability of various academic wordlists such as AWL and AVL as well as free online access to corpora such as COCA, frequencies can be taken into account to design more authentic items and more parallel tests.

AUTHORS' BIOGRAPHIES

Elaheh Rafatbakhsh is a postdoctoral researcher of TEFL in the Department of Foreign Languages and Linguistics at Shiraz University. Her main research interests include Computer Assisted Language Learning, Language Assessment and Corpus Linguistics.

Alireza Ahmadi is a professor of TEFL in the Department of Foreign Languages and Linguistics at Shiraz University, Iran. His main interests include Language Assessment and Second Language Acquisition.

REFERENCES

Ahmadi, A. & Thompson, N. A. (2012). Issues affecting item response theory fit in language assessment: A study of differential item functioning in the Iranian national university entrance exam. *Journal of Language Teaching & Research*, *3*(3), 401-412.

Ahmadi, A., Darabi Bazvand, A., Sahragard, R. & Razmjoo, A. (2015). Investigating the validity of PhD entrance exam of ELT in Iran in light of argument-based validity and theory of action. *Journal of Teaching Language Skills*, *34*(2), 1-37.

Akbari, N. (2016). Word frequency and morphological family size effects on the accuracy and speed of lexical access in school-aged bilingual students. *International Journal of Applied Linguistics*, *26*(3), 311-328.

Alderson, C. & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*(2), 115-129.

Alderson, J. C. (1996). Do corpora have a role in language assessment? In Thomas, J. & Short, M. (Eds.), *Using corpora for language research: Studies in the honour of Geoffrey Leech* (pp. 248-259), London: Longman.

Bachman, L. & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Bai, Y. (2005). *Authenticity Assessment of Proofreading in NMET by Corpus-based Approach*. Unpublished master's thesis, Guangdong University of Foreign Studies, Guangzhou, China.

Bazvand, A. D., Kheirzadeh, S. & Ahmadi, A. (2019). On the statistical and heuristic difficulty estimates of a high stakes test in Iran. *International Journal of Assessment Tools in Education*, *6*(3), 330-343.

Beglar, D. & Nation, P. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9-13.

Beigman Klebanov, B., Ramineni, C., Kaufer, D., Yeoh, P. & Ishizaki, S. (2019). Advancing the validity argument for standardized writing tests using quantitative rhetorical analysis. *Language Testing*, *36*(1), 125-144.

Biber, D., Conrad, S., Reppen, R., Byrd, P. & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, *36*(1), 9-48.

Bovaird, J. A., Geisinger, K. F. & Buckendahl, C. W. (2011). *High-stakes Testing in Education: Science and Practice in K-12 Settings*. Washington, DC: American Psychological Association.

Brown, J. C., Frishkoff, G. A. & Eskenazi, M. (2005). Automatic question generation for vocabulary assessment. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver,* 819-826.

Chen, X., Dong, Y. & Yu, X. (2018). On the predictive validity of various corpus-based frequency norms in L2 English lexical processing. *Behavior Research Methods*, *50*(1), 1-25.

Choi, I. C. & Moon, Y. (2020). Predicting the difficulty of EFL tests based on corpus linguistic features and expert judgment. *Language Assessment Quarterly*, *17*(1), 18-42.

Chujo, K. & Hasegawa, S. (2003). Jijieigo no jugyo de motiirareru eibunsozai no goi reberuchousadBNC (British National Corpus) wo kijun ni site [An investigation of vocabulary levels of materials used in current English class: in reference to BNC]. *Jiji Eigogaku Kenkyu, 42*, 439-451.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Crossley, S. A., Salsbury, T., McNamara, D. S. & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561-580.

Crosthwaite, P. R. & Raquel, M. (2019). Validating an L2 academic group oral assessment: Insights from a spoken learner corpus. *Language Assessment Quarterly*, *16*(1), 39-63.

Culligan, B. (2015). A comparison of three test formats to assess word difficulty. *Language Testing*, *32*(4), 503-520.

Cushing, S. T. (2017). Corpus linguistics in language testing research. *Language Testing*, *34*(4), 441-449.

Davies, M. (2008). The corpus of contemporary American English: 450 million words, 1990-present. Available from http://corpus.byu.edu/coca

Davis, A. (2006). High stakes testing and the structure of the mind: A reply to Randall Curran. *Journal of Philosophy of Education, 40*(1), 1-16.

Egbert, J. (2017). Corpus linguistics and language testing: Navigating uncharted waters. *Language Testing*, *34*(4), 555-564.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*, 143-188.

Gardner, D. & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305-327.

Gebril, A. & Eid, M. (2017). Test preparation beliefs and practices in a high-stakes context: A teacher's perspective. *Language Assessment Quarterly*, *14*(4), 360-379.

Goodfellow, R., Lamy, M. -N. & Jones, G. (2002). Assessing learners' writing using lexical frequency. *ReCALL, 14*(1), 133-145.

Hazenberg, S. & Hulstijn, J. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: an empirical investigation. *Applied Linguistics, 17*(2), 145-163.

Iranian National Organization for Educational Testing, (2020). www.sanjesh.org

Isaacs, T., Trofimovich, P. & Foote, J. A. (2018). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*, *35*(2), 193-216.

Johansson, S. (2009). Some thoughts on corpora and second-language acquisition. *Corpora and language teaching*, 33-44.

Larsson, M. & Olin-Scheller, C. (2020). Adaptation and resistance: washback effects of the national test on upper secondary Swedish teaching. *The Curriculum Journal*, *31*(4), 687-703.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In Arnaud, P. J. L. & Bejoint, H. (Eds.), *Vocabulary and Applied Linguistics*, pp. 126-132, London: Macmillan Academic and Professional.

Laufer, B., Elder, C., Hill, K. & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing, 21*(2), 202-226.

Lin, D. & Gao, M. (2020). Book review: Teacher involvement in high-stakes language testing. *Language Testing, 37*(1), 159-162.

Lin, Y. C., Sung, L. C. & Chen, M. C. (2007). An automatic multiple-choice question generation scheme for English adjective understanding. In *Workshop on Modeling, Management and Generation of Problems/Questions in eLearning, the 15th International Conference on Computers in Education (ICCE)*, 137-142.

Mitkov, R. & Ha, L. A. (2003). Computer-aided generation of multiple-choice tests. *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, *2*, 17-22.

Monteiro, K. R., Crossley, S. A. & Kyle, K. (2020). In search of new benchmarks: Using L2 lexical frequency and contextual diversity indices to assess second language writing. *Applied Linguistics*, *41*(2), 280-300.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59-82.

Okamoto, M. (2015). Is corpus word frequency a good yardstick for selecting words to teach? Threshold levels for vocabulary selection. *System*, *51*, 1-10.

Pan, M. & Qian, D. D. (2017). Embedding corpora into the content validation of the grammar test of the National Matriculation English Test (NMET) in China. *Language Assessment Quarterly*, *14*(2), 120-139.

Paribakht, T. S. & Webb, S. (2016). The relationship between academic vocabulary coverage and scores on a standardized English proficiency test. *Journal of English for Academic Purposes*, *21*, 121-132.

Park, K. (2014). Corpora and language assessment: The state of the art. *Language Assessment Quarterly, 11*(1), 27-44.

Pawley, A. and Syder, F. (1983). Two puzzles for linguistic theory. In Richards, J. and Schmidt, R. (eds.). *Language and Communication*. London: Longman.

Rafatbakhsh, E., Ahmadi, A., Moloodi, A. & Mehrpour, S. (2021). Development and validation of an automatic item generation system for English idioms. *Educational Measurement: Issues and Practice*, *40*(2), 49-59.

Ravand, H. & Firoozi, T. (2016). Examining construct validity of the master's UEE using the Rasch model and the six aspects of the Messick's framework. *International Journal of Language Testing*, *6*(1), 1-18.

Ravand, H., Rohani, G. & Faryabi, F. (2018). On the factor structure (invariance) of the PhD UEE using multigroup structural equation modeling. *Journal of Teaching Language Skills*, *36*(4), 141-170.

Razavipur, K. (2014). On the substantive and predictive validity facets of the university entrance exam for English majors. *Research in Applied Linguistics*, *5*(1), 77-90.

Sasao, Y. & Webb, S. (2017). The word part levels test. *Language Teaching Research*, *21*(1), 12-30.

Schmidtke, J. (2014). Second language experience modulates word retrieval effort in bilinguals: Evidence from pupillometry. *Frontiers in Psychology*, *5,* 1-16.

Schmitt, N. (2012). Formulaic language and collocation. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*, pp. 1-10, New York: Blackwell.

Shohamy, E. (2001). *The Power of Tests: A Critical Perspective on The Uses of Language Tests*. Harlow, England: Longman.

Shohamy, E., Donitsa-Schmidt, S. & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing*, *13*(3), 298-317.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford, UK: Oxford University Press.

Staples, S., Biber, D. & Reppen, R. (2018). Using corpus-based register analysis to explore the authenticity of high-stakes language exams: A register comparison of TOEFL iBT and disciplinary writing tasks. *The Modern Language Journal*, *102*(2), 310-332.

Taylor, L. & Barker, F. (2008). Using corpora for language assessment. *Encyclopedia of Language and Education*, *7*, 241-254.

Vu, D. V. (2019). A corpus-based lexical analysis of Vietnam's high-stakes English exams. In *The 20th English in Southeast Asia Conference*. Singapore: National Institute of Education, Nanyang Technological University.

Weir, C. J. and Milanovic, M. (Eds.) (2003). *Continuity and innovation: The History of the CPE, 1913-2002*. Vol. 15, Cambridge, England: Cambridge University Press.