Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 27 Number 28, November 2022 ISSN 1531-7714

The Argument-Based Validation of a Large-Scale High-Stakes Vocabulary Test

Elaheh Rafatbakhsh, *Shiraz University*
Alireza Ahmadi, *Shiraz University*

The purpose of this study was to investigate the validity of the vocabulary subsection of a high-stakes university entrance exam for Ph.D. programs using the argument-based approach. All the three different versions of the test administered in a period of five years and the responses of 12,500 test-takers were studied. The study focused on four inferences of domain definition, evaluation, generalization and explanation mainly using corpus linguistics, the Rash measurement model and factor analysis. The results indicated substantial threats to the validity of the test in terms of vocabulary choice, item difficulty, item discrimination, construct representation, and reliability.

Keywords: argument-based validity, Corpus of Contemporary American English (COCA), vocabulary test, Rasch model, high-stakes assessment

Introduction

Vocabulary assessment is an inseparable component of vocabulary acquisition as it serves several important purposes such as evaluating different aspects of lexical knowledge, studying the results of different treatments and modeling their impacts, assessing vocabulary growth, observing the results of various pedagogical interventions and estimating learners' strengths and weaknesses (Beglar & Nation, 2013). However, despite the great deal of research on vocabulary assessment, the literature suffers from a paucity of research on validation issues in this field (Schmitt et al., 2020).

Validation methods have constantly evolved with argument-based validation (Bachman & Palmer, 2010; Chapelle et al. 2008, Kane, 1992, 2001, 2006, 2013) being the most recent framework. It has been widely used as it involves a fairly simple and systematic process. It insists less on formal scientific theories than on a model of argumentation as used in real-world contexts: "informal logic" and "presumptive reasoning" (Kane, 2004, p. 145). In this approach to validation, a clear claim is made and justifiable evidence and comprehensive interpretations are provided accordingly. Considering testing contexts and test uses, researchers can now flexibly select what claims to make and what evidence to gather for their support (Chapelle, 2012; Chapelle et al., 2010; Kane, 2013).

In a recent paper, Schmitt et al. (2020) presented step-by-step guidelines for vocabulary test validation through argument-based approach in order to encourage more systematic and rigorous procedures for test development. We followed their guidelines for vocabulary test validation which are based on Chapelle et al.'s (2008) framework involving six inferences/steps. The first inference, domain definition, 'links performance in the target domain to the observations of performance in the test domain' (Chapelle, Enright, & Jamieson, 2008, p. 14). This step requires a careful analysis of the domain involved in item selection and relevance/effectiveness of the test design methodology or item format. The evaluation step involves the analysis of test scores. This includes providing evidence for the adequacy of test items and scoring procedures through statistical analysis. The generalization step mainly focuses on the reliability and generalizability of the vocabulary test scores. The consistency of the scores and the ability of the test to discriminate between different groups of test-takers are among the evidence that should be provided in this step. Then, the explanation step connects the items and scores to the construct definition. In this step, the relationship between the test and other tests with similar constructs or skill areas is explored. The next step, extrapolation, connects the test scores to the ability of the test-takers outside the test setting by comparing the test-takers' performance beyond the test situation but in a relevant domain. In the final step, the utilization and impact of the test should be studied. This step is more related to the purpose and use of a test, whether a test is useful exactly in the domain that it claimed to be in.

Recently, various validation studies in both high-stakes and low-stakes assessment contexts have used argument-based approach to validate language tests (e.g., Aryadoust, 2013; Brooks & Swain, 2014; Chapelle, et al., 2008; Cheng & Sun, 2015; Crosthwaite & Raquel, 2019; Hsu, 2012; LaFlair & Staples, 2017; Liu, 2014; Pan & Qian, 2017; Staples, et al., 2018; Sun, 2016; Youn, 2015). However, very few of such studies (e.g., Beglar, 2010; Fitzpatrick, & Clenton, 2010; Karami, 2012; McLean et al. 2015; Schmitt et al., 2011) have focused on validating vocabulary tests.

For instance, Beglar (2010) validated the Vocabulary Size Test which evaluates written receptive knowledge of the first 14,000 words of English. This study focused on various aspects of Messick's validation framework mainly using the Rasch model. The findings indicated that the Vocabulary Size Test was a valid test as the vast majority of the items showed good fit to the Rasch model, strong degree of unidimensionality and measurement invariance. The test also had low standard errors and high reliability estimates.

Within the Iranian context, national university entrance exams are the most important high-stakes largescale exams which are held annually at three levels of bachelor, master, and PhD, whose results greatly define individuals' lives. These exams include items assessing domain-specific knowledge and also general English knowledge of the candidates. Various validation studies have been conducted to validate Iranian National University Entrance Exams (INUEE) at different levels of education (e.g., Ahmadi et al. 2015; Ahmadi & Thompson, 2012; Barati et al., 2006; Darabi Bazvand, & Ahmadi, 2020; Darabi Bazvand et al., 2019; Ravand & Firoozi, 2016; Ravand et al., 2018; Razavipur, 2014; Razmjoo & Heydari Tabrizi, 2010). These studies have specifically focused on the washback effect, differential item functioning, content and construct of the test.

For example, Ravand and Firoozi (2016) exploring the validity of the general English proficiency sections of INUEE for master's program found that the majority of items of this section showed good fit to the Rasch model. However, the lack of invariance in person measures displayed threats to construct validity. Also, the difficulty of the items seemed to be much above the ability of the test-takers and the test displayed low Rasch reliability estimates for all the sections of reading, grammar, and vocabulary.

However, few of these studies have used the argument-based framework, and to the best of the researchers' knowledge, no study has specifically focused on the vocabulary section of these high-stakes university entrance exams which are annually used to screen a large number of candidates seeking admission to the university. Also, the knowledge of academic vocabulary which is directly connected to academic success, societal well-being, and economic opportunity (Goldenberg, 2008; Ippolito et al. 2008; Jacobs, 2008) for both native and non-native speakers of English, has not received enough attention. There is still a need for more explicit and focused academic vocabulary instruction (Gardner & Davies, 2014). As such, the current study was aimed at validating the vocabulary section of the INUEE for PhD candidates designed in three versions for the fields of humanities, engineering and English language. For the current study, the first four steps of Chapelle et al.'s (2008) framework were investigated following the guidelines provided by Schmitt et al. (2020) for vocabulary test validation through argument-based approach. Since we no longer had access to the test-takers, we did not explore the extrapolation and utilization/impact inferences.

Method

Data

INUEE for PhD is a high-stakes exam, run annually to screen candidates for admission into universities to pursue Doctor of Philosophy. Participants with higher scores go through the second phase which involves an interview and the assessment of educational and research background. Each year approximately 170,000 test-takers nationwide register to take part in this exam in the seven broad categories of natural and physical sciences, humanities, engineering, agriculture, languages, arts, and veterinary science. Each of these categories includes various disciplines receiving the same version of the test. For example, the engineering category includes about 40 different disciplines, and all should sit for the same version of the test. The exam includes items on different subjects such as an English proficiency section, assessing general proficiency knowledge of the test-takers, which consists of overall 30 items, including eight items of grammar, 12 items of vocabulary, and two reading passages each with five comprehension items. Each year, three versions of the English proficiency test are administered to different fields of study.

In this study, the vocabulary subsection was explored since it constitutes 40% of the items and therefore determines a large proportion of the overall proficiency score. For this purpose, the vocabulary subsections of the three test versions in five years, from 2015 to 2019, were examined. Therefore, for each test version, 60 vocabulary items (12 items each test in five years) were studied. The items were all in multiple-choice format.

For the data, the responses of the test-takers in the INUEE PhD were requested from the Iranian National Organization for Educational Testing. For each test version, we were granted access to the responses of one field of study, i.e., engineering (including about 40 disciplines), humanities (including about *8*0 disciplines), and English language (including 4 disciplines). The data included the responses of 5,000 test-takers in the fields of engineering, 5,000 in the fields of humanities and 2,500 in the fields of English language from 2015 to 2019 (nearly 7% of the whole population each year). The participants were female and male non-native speakers of English with various levels of English proficiency and their ages varied, ranging from 23 to 71 years old.

Overall, this study examined a total of 180 vocabulary items in the vocabulary subsection, and a total number of 12,500 test-takers' responses to these items over five years.

Data analysis procedure

For the purpose of test validation, Chapelle et al.'s (2008) framework was used in the current study. We followed the suggestions that Schmitt et al. (2020) presented for evidence-gathering for each step of the validation process. From among the six steps of the approach, the four steps of domain definition, evaluation, generalization, and explanation were addressed.

In the domain definition step, we first defined the domain and the context of the mentioned tests and analyzed the results using the data from a corpus and wordlists. Since the test designers had not claimed to employ any specific sources, corpora, or frequency lists from which they designed the tests, we selected the widely accepted sources that best represent the English language to investigate the test domain.

The corpus used in this study was the Corpus of Contemporary American English (COCA) created by Mark Davies (2008-present). This corpus is the first large genre-balanced corpus which well-represents the English language and models changes in the real world. The purchased version of COCA that we used in this study includes over 520 million words in 220,225 texts from 1990 to 2015 which is evenly divided between five genres of spoken, fiction, popular magazines, newspapers, and academic journals. We extracted all the options in all the items of the tests and searched them and their lemmas (sets of lexical forms) in the corpus. Normally, searching corpora for frequency data is done by concordancers which are computer programs for text analysis. However, because of their limitations, scholars propose researchers to develop their own tools for text analysis based on their specific needs and purposes (e.g., Anthony, 2009). Therefore, due to the large size of the corpus and the large number of items (720 options), we used a computer program specifically written for this study by an expert programmer using Hypertext Preprocessor (PHP) scripting language. The system was designed with the capability of receiving a large wordlist as the input, searching all the items in the list along with their lemmas in the corpus, and giving a spreadsheet of items and their frequencies as its output.

We also made use of two academic wordlists, the Academic Word List (AWL) (Coxhead, 2000) and the Academic Vocabulary List (AVL) (Gardner & Davies, 2014). AWL contains 570 word-families and AVL 2,000 word-families. To calculate the number and percentages of common items between the options of the tests and the two wordlists, AntWordProfiler version 1.5.1 created by Laurence Anthony (2021) was employed.

Then, in the evaluation step, the main focus was on the examinees and their scores. The Rasch measurement model (Wolfe & Smith, 2007a, 2007b), was chosen for this step because it is "the most simple and robust model" of IRT (Luoma, 2004) to measure person ability and item difficulty. In situations where item discrimination and guessing are also significant factors to consider, other IRT models such as the two-parameter and three-parameter models would be better options than the Rasch model. In this study, however, the Rasch model was a good choice since only item difficulty was our concern. Furthermore, Rasch indices (e.g., infit and outfit mean square values) revealed a good data-model fit. The Rasch model provides evidence for how the test functions in measuring the intended construct by analyzing item and person measures and their relationships, therefore suitable for validity arguments. The Rasch model was also employed for the generalization step which was concerned with the reliability and the generalizability of the scores. Here, the reliability and separation values for both items and persons were measured. For both inferences, we used Winsteps software version 3.68.2.

Finally, in the explanation step, we studied the construct of vocabulary which the tests were intended to measure. First, we ran factor analysis to check the number of constructs being tested. For factor extraction, all three eigenvalue-based procedures including Kaiser's (1974) criterion, Cattel's (1966) scree plot, and Horn's (1965) parallel analysis were studied to reach a satisfactory result.

Then, the correlations between test-takers' scores in the vocabulary subsection and the two subsections of grammar and reading comprehension were measured to find out whether the scores yielded by the test can be attributed to the theoretical construct of vocabulary. For both factor analysis and correlation studies, we used IBM SPSS software version 26. The overall research design for each inference is presented in Table 1.

Table 1. Inferences and sources of evidence

Inference	Aim	Methods and sources for backing
Domain definition	Analysis of the domain	• Frequency search of the options in COCA o Developing a tailor-made computer program • Option search in two wordlists of AVL and AWL
Evaluation	Analysis of test scores	• Item and person statistics • Item difficulty • Misfitting items o Rasch analysis > Winsteps version 3.68.2
Generalization	Analysis of reliability and generalizability	• Item and person reliability and separation o Rasch analysis > Winsteps version 3.68.2
Explanation	Linking the items and scores to the construct definition	• Factor analysis > Eigenvalue-based methods: Kaiser's criterion, Cattel's scree plot, & Horn's parallel analysis • Correlation between test-takers' scores in vocabulary subsection and other subsections o IBM SPSS software version 26

Results and Discussion

Domain definition

Domain definition is the first step of the chain of reasoning in validity argument. INUEE for PhD admission is a screening test, and assessing test-takers' English proficiency is a part of this test that contributes to the selection of the candidates. As a result, we can identify the academic domain as the dominant context of use in which particular linguistic knowledge is required to perform university tasks. In this respect, we examined the domain of the vocabulary items to see if it is consistent with the domain in which the test is expected to be. We extracted all the four options of each item from the three test versions designed for the fields of humanities, engineering, and English language (each version included 60 items, i.e., 12 items in five years) and searched their frequency in COCA. Overall, 720 words (240 words for each test version) along with their lemmas were searched in all the five genres of the corpus. To make the test versions more comparable, we separately extracted the option frequencies for each test version each year and calculated the average frequencies over the five years. Figure 1 shows the average sums of option frequencies per million for each test version in 5 years separated by genres.

Option frequencies are the highest in the academic genre in all three test versions (Figure 1). This means that that the tests are in line with the purpose they have been intended for. To study in PhD programs, the candidates enter an academic context that requires the knowledge of academic language more than the other genres.

Another finding is that the average frequencies are higher for the fields of humanities and engineering compared to the English language which is logical since the fields of English language demand higher levels of vocabulary knowledge. This is because English is the teaching medium for PhD programs in the fields of English language including Teaching English as a Foreign Language, English Literature, English Translation Studies, and Linguistics. As such, PhD candidates in these fields may encounter less frequent words compared to other fields of study.

Also, the difference between the coverage of the academic genre and the other four genres is less in the fields of English language. This can be explained by the fact that the academic genre is not the only genre that students of these fields are exposed to during their PhD studies. They may also face other genres such as fiction and spoken as a part of their education.

As such proficiency tests for high-stakes admission into universities should indicate whether test-takers can handle the linguistic demands of their academic studies, examining the lexical coverage of the existing academic wordlists in these tests can provide evidence for the suitability of the tests for university admission purposes. Moreover, such analysis can also help compare different versions of the test in terms of similarity and consistency in the use of vocabulary. Therefore, for further analysis, we examined the occurrence of items from the AWL and AVL in the 240 options in each test version. The percentages of the common items in the wordlists and the options are presented in Table 2.

Figure 1. The average sums of option frequencies (pm) in COCA

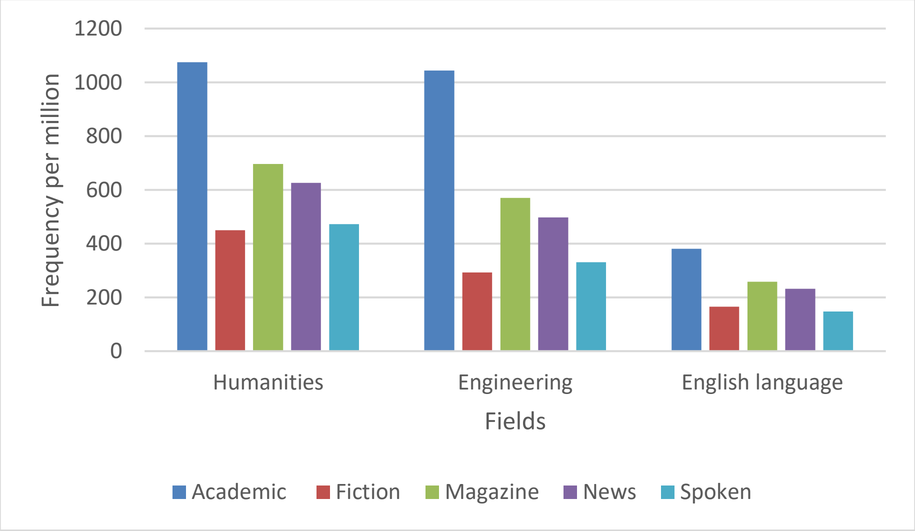

Table 2. The percentages of AVL and AWL items in the options

	AVL	AWL
Humanities	47.9%	17.8%
Engineering	50.6%	21.3%
English language	23%	2.9%

As indicated, from all the 240 options in the items designed for the fields of engineering and humanities, 50.6% and 47.9% were found in the AVL, respectively. However, in the test designed for the fields of English language, only 23% of the options occurred in the AVL. The coverage percentages of AWL have the same order, the highest in engineering (21.3%) and the lowest in the English language test version (2.9%). However, the percentages of AWL items in the options of the tests were much lower in all three versions. This difference in the percentages might be due to the differences between the two wordlists. The AVL is based on the 120-million-word academic subcorpus of COCA which is a much larger corpus than BNC, based on which the AWL was created. Also, AVL's coverage in COCA academic and BNC academic is approximately 13.8% whereas AWL's coverage is about 7% (Gardner & Davies, 2014). The AVL represents different conceptualizations of 'core academic' and includes more high-frequency words than the AWL.

In a similar study, 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. Compared to this study, the academic vocabulary coverage in our study is much lower in the AWL.

There were no studies on the general English proficiency section to compare the results with. However, a study was carried out by Darabi Bazvand and Ahmadi (2020) that focused on the domain description inference of the subject matter section of the PhD Entrance Exam of English language teaching. A similar conclusion was reached; the test tasks in that section were not fully represented in the postgraduate syllabus and PhD course objectives.

We next had a look at the item format which was multiple-choice fill-in-the-blank. According to Schmitt (2019), this is the most typical item format for vocabulary assessment which can only assess the recognition level of mastery. Alderson and Kremmel (2013) claimed that for diagnostic assessment, it is not desirable to separate grammar and lexicosemantic knowledge and perhaps it is more rational for these two to be considered as one unitary component of reading ability. In line with this claim, Schmitt (2019) suggested assessing the target words in various authentic reading and listening contexts besides placing them in typical test formats which measure them in isolation. In the tests under study, grammar and vocabulary assessment constituted different subsections of the general proficiency section and the reading comprehension part did not assess the vocabulary knowledge. Whereas, one option for such vocabulary tests is the use of reading comprehension questions which require direct knowledge of target words to be answered.

Evaluation

Next in Chapelle et al.'s (2008) framework is the evaluation step where test-takers' responses and scores are interpreted. We examined item difficulty as well as response behaviors and examinees' abilities for the three test versions in 5 years using the Rasch model. Table 3 presents the summary of the results of item and person statistics for the fields of humanities.

From the 1000 responses that we analyzed for each year in the fields of humanities, 333 to 445 (33.3% to 44.5%) of the test-takers chose not to answer the English proficiency at all and they left all the 30 items in their answer sheets blank. We eliminated these test-takers from our analysis and only considered the participants who answered at least one item of this section. From the population who had answered at least one item in the English proficiency section, 23.9% to 34% did not respond to the vocabulary subsection. This means that overall, 53.2% to 60.7%, in different years, chose not to answer the vocabulary subsection altogether.

The mean scores of the test-takers in the vocabulary subsection over the 5 years were 1.5 and 1.6 out of 12 (Table 3). In addition, the average person ability to answer vocabulary items in all 5 years, was far less than the easiest items. There were floor effects present for the test-takers in all the tests; in other words, most of the scores were near the bottom because even the easiest item was too difficult for the population.

In the fields of engineering (Table 4), the percentages of the test-takers who had not answered any items in the English proficiency section ranges from 24.8% to 28.3% of 1000 participants over the 5 years, which was lower than those in the fields of humanities. From the examinees who answered at least one item of this section, 17.2% to 27.5% did not answer any of the vocabulary items. Therefore, overall 39.4% to 47.8% left the vocabulary section empty. In addition, the mean scores in the fields of engineering are slightly higher compared to the fields of humanities; however, they do not exceed the score of 2.4 out of 12. The average person ability and item difficulty range also indicate that the easiest items are more difficult than the average abilities of the examinees.

We also studied the scores of the candidates participating in the PhD entrance exam in the fields of English language (Table 5).

From the data of 500 examinees that we studied each year, 2.4% to 21.6% had not answered the general English proficiency section and 16.8% to 26.5% of the test-takers who answered this section, left the vocabulary subsection blank. The percentages of the test-takers who were not able to answer the vocabulary section were lower than those in the other two fields of study, 9.4% to 14.7% of the overall population. However, the mean scores were similar, from 1.5 to 2.7 out of 12 and the ability estimates of the majority of the test-takers were below the lowest item difficulty measures.

Table 3. Item and person statistics for the fields of humanities

Year	Participants who answered the English section	Participants who answered the vocabulary subsection	Mean score out of 12	Average person ability	Person ability range	Item difficulty range
2015	65.8%	46.8%	1.5	-2.45	1.68 to -3.76	.90 to -.83
2016	66.7%	50.7%	1.5	-2.46	2.48 to -3.77	.75 to -.98
2017	59.6%	39.3%	1.5	-2.49	1.64 to -3.70	.30 to -.68
2018	55.5%	39.6%	1.6	-2.40	2.45 to -3.71	.46 to -.65
2019	61.6%	42.3%	1.5	-2.52	1.66 to -3.73	.54 to -.64

Table 4. Item and person statistics for the fields of engineering

Year	Participants who answered the English section	Participants who answered the vocabulary subsection	Mean score out of 12	Average person ability	Person ability range	Item difficulty range
2015	71.7%	53.3%	1.7	-2.43	1.77 to -3.86	1.20 to -.97
2016	73.7%	53.4%	1.6	-2.72	4.03 to -4.27	1.07 to -2.32
2017	75.2%	55.5%	2	-2.29	2.58 to -3.96	1.10 to -1.86
2018	71.7%	52.2%	1.8	-2.40	3.95 to -3.91	1.32 to -1.18
2019	73.2%	60.6%	2.4	-2.20	4.19 to -4.16	1.56 to -1.42

Table 5. Item and person statistics for the fields of English language

Year	Participants who answered the English section	Participants who answered the vocabulary subsection	Mean score out of 12	Average person ability	Person ability range	Item difficulty range
2015	78.4%	57.6%	1.9	-2.20	3.69 to -3.69	.47 to -.58
2016	97.6%	81.2%	2.7	-1.83	3.89 to -3.88	1.50 to -1.46
2017	95%	78%	2	-2.24	2.64 to -3.99	1.43 to -1.54
2018	96%	70.6%	1.9	-2.33	2.60 to -3.90	1.41 to -1.31
2019	95.6%	73.4%	1.5	-2.44	3.75 to -3.75	.97 to -.89

From the data of 500 examinees that we studied each year, 2.4% to 21.6% had not answered the general English proficiency section and 16.8% to 26.5% of the test-takers who answered this section, left the vocabulary subsection blank. The percentages of the test-takers who were not able to answer the vocabulary section were lower than those in the other two fields of study, 9.4% to 14.7% of the overall population. However, the mean scores were similar, from 1.5 to 2.7 out of 12 and the ability estimates of the majority of the test-takers were below the lowest item difficulty measures.

The vocabulary tests were also assessed through an inspection of misfitting items using the Rasch model. Infit and outfit-mean square (MNSQ) values between 0.5 and 1.5 are productive for measurement (Green, 2013). Values higher than this range are considered underfit which means the persons or items behaviors are too unpredictable. On the contrary, MNSQ values lower than 0.5 are overfit which in this case a person or an item is behaving too predictably. This happens when a person answers all the easy items correctly and the difficult ones incorrectly. ZSTD values are expected to be within the range of -2 to +2. However, if infit MNSQ values are within the acceptable range, ZSTD can be ignored (Green, 2013).

In the current study, we assessed all MNSQ and ZSTD values in the three test versions over the 5 years, and the result indicated no MNSQ values outside the acceptable range. There were items with ZSTD values higher than +2 (3.3% in the test versions designed for humanities, 6.6% in engineering, & 11.6% in English language) and lower than -2 (3.3% in humanities, 11.6% in engineering, & 13.3% in English language). However, since the infit MNSQ values are acceptable, ZSTD values are not a threat to the quality of the tests.

In sum, all the items in the three test versions were fit to the Rasch model, however, the difficulty levels of the tests were much higher than the average ability estimates of the population who took the tests. An inspection of the number of students who left this section blank and did not respond at all, confirms that test was unduly difficult for this population. The results confirm the findings of similar validation studies. For instance, according to the results of the study on the general English proficiency sections of INUEE for master's program conducted by Ravand and Firoozi (2016), although the items displayed good fit to the Rasch model, the difficulty levels of the items were very much above the abilities of the examinees. Also, Darabi Bazvand et al. (2019) and Darabi Bazvand and Ahmadi (2020) studied the items of the subject matter section (as opposed to the general section) that measure the applicants' expertise in the field of English language teaching in the INUEE for PhD, and according to the results of surveys and statistical analysis, the test and all its subsets were considered very difficult for the population and best reliable for high-ability test-takers.

One reason for the development of such difficult items might be the fact that each year a lot of candidates participate in entrance exams and therefore the competition is high among the test-takers while only a limited number can enter the universities. That might be the reason behind ignoring the ability level of the majority of the test-takers to be able to filter the most capable candidates.

Generalization

Generalization is the next step in the process of validation which focuses on the reliability and generalizability of the vocabulary test scores. In this respect, we calculated the reliability and separation indexes for both items and persons for all three test versions from 2015 to 2019 using the Rasch model (Table 6).

Person separation value indicates the extent to which the instrument is sensitive to classify the test-takers and distinguish between performers with different levels of ability. As stated by Linacre (2012), person separation value below 2 and person reliability below 0.8 show lack of sensitivity of the test to separate different levels. Item separation, on the other hand, verifies item hierarchy. According to Linacre (2012), item separation lower than 3 and item reliability below 0.9 implies that the sample of people is not large enough to confirm the item difficulty hierarchy of the instrument. As shown in Table 6, all person separation values are below the acceptable value of 2, and the person reliability estimates are below 0.8. This means that the test is only able to separate 1 or at most 2 levels of vocabulary knowledge. Therefore, the discrimination ability of the test is limited and so the test cannot serve its purpose effectively, i.e., to differentiate among test-takers for screening purposes.

The test should include either more items, more categories, or better sample-item targeting in order to have higher person reliability (Linacre, 2012). In this test, 12 items with the mentioned difficulty level for vocabulary assessment might not be a good indication of the test-takers' knowledge. Therefore, for this vocabulary test adding more items is one possible way to help improve the person reliability.

In addition, having a look at Table 6, one can say that the vocabulary tests for the fields of engineering have acceptable item separation and reliability values. However, three tests designed for the fields of humanities and two tests for the fields of English language seem to have item separation and reliability below the acceptable values. This means that the sample is not big enough to confirm the item difficulty hierarchy of the instruments and we do not have a reasonable amount of confidence in the replicability of the performance of these items on another similar test population. To have higher item reliability, the tests should either include a wider difficulty range or a larger sample size (Linacre, 2012). Therefore, including items of various difficulty levels can enhance the reliability of the vocabulary tests. Frequency lists can be employed to develop vocabulary test items to systematically include items with specific difficulty levels.

Table 6. Item and person reliability and separation

		Item reliability	Item separation	Person reliability	Person separation
Humanities	2015	*.91*	*3.24*	.00	*.04*
	2016	*.91*	*3.18*	.13	.38
	2017	.77	*1.84*	.09	.32
	2018	.68	1.45	.00	.00
	2019	.86	2.53	.16	.44
	Average	0.826	2.448	0.076	0.236
Engineering	2015	.96	4.59	.29	.64
	2016	.98	7.01	.10	.34
	2017	.97	5.76	.29	.65
	2018	.97	5.54	.32	.69
	2019	.99	8.37	.45	.90
	Average	0.974	6.254	0.29	0.644
English language	2015	.67	1.44	.38	.78
	2016	.96	4.71	.55	1.11
	2017	.96	5.09	.20	.50
	2018	.95	4.32	.28	.62
	2019	.87	2.56	.05	.23
	Average	0.882	3.624	0.292	0.648

Explanation

Following the generalization step, the explanation step connects the items and scores to the construct definition. Following the suggestions by Schmitt et al. (2020), we first examined the internal structure of the tests using factor analysis. We then studied the relationship between the vocabulary subsections and other similar subsections of the same tests.

The tests under the study aimed to assess the vocabulary knowledge of the test-takers, therefore, we ran factor analysis to check if the tests assess the intended construct. Although researchers (e.g., Conway & Huffcutt, 2003; Fabrigar et al., 1999) have suggested that factor analysis should be done using a combination of procedures and many (e.g., Dinno, 2009; Schmitt, 2011) identified parallel analysis as the most effective method, few studies in applied linguistics have considered a combination of the three eigenvalue-based criteria in factor selection and mostly ignored the importance of parallel analysis (Karami, 2015). Therefore, for more satisfactory results, we conducted factor analysis using the three eigenvalue-based methods, i.e., Kaiser's criterion, scree plot, and parallel analysis.

Results of the factor analysis of the 15 tests indicated that the vocabulary tests designed for humanities except for the year 2015, for which we extracted two factors, assess only one factor. The factors extracted for the tests developed for the fields of engineering were on average 1.8 since all but one test (the year 2018 with one factor extracted) seem to include two underlying factors. All the tests developed for the fields of English language also indicated to measure two factors. Upon further content analysis of the tests measuring more than one factor, we could not find a logical pattern for the factors depicted. Neither were we able to find any differences between the two vocabulary factors extracted. In other words, similar vocabulary items were found in different test factors, which were not logically acceptable. We argue this may be due to the problems in test designing. As explained, the test overall showed problems in word choice based on the appropriate frequency levels and genres.

The second way in which the validation of the test was assessed in this step was through exploring the correlation between test-takers' scores in the vocabulary subsection and in other subsections. Alderson and Kremmel (2013) argue that the constructs of grammar and potentially reading ability are inseparable from the construct of vocabulary and they are "highly patterned structure of language" (p. 549). Therefore, in the current study, we examined the correlation between the scores of the vocabulary subsection, and the grammar and reading comprehension subsections of the same tests. To this end, Pearson correlation was calculated for all 5 tests in the three test versions and the average correlation values of the 5 tests are displayed in Table 7.

The results indicated significant positive correlations between the mentioned constructs in all the test forms. The average correlation between the scores of vocabulary and reading comprehension subsections is slightly higher than that of vocabulary and grammar subsections. The average correlation values in the tests for the fields of English language are overall lower than those of the other two test versions.

The results are partially in line with the studies which proved a high positive correlation between the construct of vocabulary and reading comprehension (Laufer & Ravenhorst-Kalovski, 2010; Qian, 2008; Stæhr, 2008; Zhang, 2012) and between lexis and syntax (Alderson & Kremmel, 2013; Romer, 2009; Shiotsu & Weir, 2007). However, in this study, while there exist positive correlations between these subsections, the values (< 0.5) show weak correlations in almost all cases. Table 8 summarizes the evidence for and threats to the validity of the vocabulary subsection.

Table 7. The average correlation of vocabulary with grammar, and reading comprehension

	Average correlation with grammar	Average correlation with reading comprehension
Humanities	0.445**	0.493**
Engineering	0.493**	0.538**
English language	0.333**	0.395**

**. Correlation is significant at the 0.01 level (2-tailed).

Table 8. Overview of the validation framework, the evidence and the threats

Inference	Claim	Evidence for validity	Threats to validity
Domain definition	The domain that the test is intended to assess and the test format are in line with the test purpose which is admission for PhD program.	#Option frequencies are the highest in the academic genre in all three test versions. #The parallel test versions represent the academic genre similarly. # Option frequencies in the tests for the fields of English language are lower than those in the other versions.	#The coverage of the academic wordlists is lower than that in the other studies. #Separating grammar and lexicosemantic knowledge is not desirable for assessment.
Evaluation	Observations of performance on the vocabulary subsection are evaluated to provide observed scores with intended characteristics.	#There were not misfitting items according to the Rasch analysis.	#The difficulty levels of the items in all three versions were much higher than the average ability level of the test-takers. #A large percentage of test-takers in the fields of humanities and engineering did not answer this subsection.
Generalization	Observed scores are reliable and generalizable and the test has discriminating power.	#The test version for the fields of engineering has acceptable item separation and reliability values.	#The three test versions are only able to separate 1 or at most 2 levels of vocabulary knowledge #The tests designed for the fields of humanities and English language have item separation and reliability below the acceptable values.
Explanation	Expected scores are attributed to the construct of vocabulary.	# The test version designed for the fields of humanities on average assess 1.2 factors. #The vocabulary subsection had significantly positive correlations with both the grammar and reading comprehension subsections.	#There were on average 1.8 factors extracted for the engineering test version. #The English language test version indicated to tap two factors. #There existed significant but weak correlations between the vocabulary subsection and the grammar and reading comprehension subsections.

Overall, the outcomes indicated significant problems in a variety of areas for this vocabulary test. No claims seem to be fully supported by the evidence as severe threats exist to their validity. Therefore, this vocabulary test is not entirely a valid assessment tool for evaluating test-takers' lexical knowledge for academic purposes.

Conclusions and implications

The use of a single exam to make decisions about the examinees is not uncommon in higher education. Validating such tests is crucial and valuable as the results of these tests directly affect individuals' life prospects both socially and financially. Despite the importance of validation processes and the insistence of scholars on validating tests, not many studies have endeavored to meaningfully validate these high-stakes tests. Therefore, the purpose of this study was to present validity evidence for the vocabulary subsection of the high-stakes PhD university entrance exam using the first four steps of Chapelle et al.'s (2008) argument-based framework. The tests under examination included the vocabulary subsections designed for the three fields of humanities, engineering and English language from 2015 to 2019.

The results of this validation study show substantial problems in the functioning of the tests and accordingly provide insights into the solutions for the improvement of these tests. The problems with the three test versions mainly include testing unnecessary vocabulary items, including extremely difficult and low discriminating items, misinterpreting the vocabulary construct and in some cases having low item separation and reliability among others. On the other hand, items fit to the Rach model, the dominance of the academic genre, significant positive correlations with the grammar and reading comprehension subsections, and acceptable item separation and reliability for the engineering test versions are among the strengths of the mentioned tests.

Besides the effects on test-takers' future, such a high-stakes test can have a very strong washback effect (the impact of testing on teaching and learning practices). Therefore, everything teachers do in their preparatory classes for this test, in terms of the skills they focus on, their teaching method as well as students' learning strategies, are highly affected by this test (e.g., Farhady & Hedayati, 2009; Riazi & Razavipour, 2011). As such, these problems may have severe harmful consequences for teachers, test-takers and the whole educational system. Some measures can be taken before the administration of such vocabulary tests. For instance, data from corpora, including wordlists and word-families, are considered as yardsticks for the selection of the words to be tested. This information can logically complement the intuition and the knowledge of the experts in the process of test design and validation. With regards to the test format, improvements can be applied by merging the vocabulary section with grammar or reading comprehension sections according to the previous research. Also, adding more items with various difficulty levels can significantly enhance the reliability, generalizability and discrimination power of the test.

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.

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im) possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, *30*(4), 535-556.

American Psychological Association. (1996). Standards for educational and psychological tests and manuals. Washington DC.

Anthony, L. (2009). Issues in the design and development of software tools for corpus studies: The case for collaboration. In P. Baker (ed.), *Contemporary corpus linguistics,* (pp. 87-104). London, UK: Continuum Press.

Anthony, L. (2021). AntWordProfiler [computer software]. Retrieved from https://www.laurenceanthony.net

Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency*. Cambridge Scholars Publishing.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford University Press.

Barati, H., Ketabi, S., & Ahmadi, A. (2006). Differential item functioning in high stakes tests: The effect of field of study. *IJAL*, *19*(2), 27-42.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*(1), 101-118.

Beglar, D., & Nation, P. (2013). Assessing vocabulary. *The companion to language assessment*, *1*, 172-184.

Brooks, L., & Swain, M. (2014). Contextualizing performances: comparing performances during TOEFL iBT TM and real-life academic speaking activities. *Language Assessment Quarterly, 11*, 353–373.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245–276.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…. *Language Testing*, *29*(1), 19-27.

Chapelle, C. A., Chung, Y.-R., Hegelheimer, V., Pendar, N., & Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. *Language Testing, 27*(4), 443–469.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Cheng, L., & Sun, Y. (2015). Interpreting the impact of the Ontario Secondary School Literacy Test on second language students within an argument-based validation framework. *Language Assessment Quarterly*, *12*(1), 50-66.

Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods, 6*(2), 147-168.

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

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.

Darabi Bazvand, A., & Ahmadi, A. (2020). Interpreting the validity of a high-stakes test in light of the argument-based framework: Implications for test improvement. *Journal of Research in Applied Linguistics*, *11*(1), 66-88.

Darabi Bazvand, A., 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.

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

Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of simulated data. *Multivariate Behavioral Research, 44*, 362-388.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272-299.

Farhady, H., & Hedayati, H. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics*, 29, 132–141.

Fitzpatrick, T., & Clenton, J. (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary. *Language Testing*, *27*(4), 537-554.

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

Goldenberg, C. (2008). Teaching English language learners: What the research does—and does not—say. *ESED 5234 - Master List 27*. 8-44.

Green, R. (2013). *Statistical analyses for language testers*. Springer.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179–185.

Hsu, H.-L. (2012). *The impact of world Englishes on language assessment: Rater attitude, rating behavior, and challenges (IELTS)*. (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses database. (3571158).

IBM Corp. (2019). IBM SPSS Statistics for Windows, Version 26.0. [Computer software]. Armonk, NY: IBM Corp.

Ippolito, J., Steele, J. L., & Samson, J. F. (2008). Introduction: Why adolescent literacy matters now. *Harvard Educational Review, 78*(1), 1-6.

Jacobs, V. (2008). Adolescent literacy: Putting the crisis in context. *Harvard educational review, 78*(1), 7-39.

Kaiser, H. (1974). An index of factorial simplicity. *Psychometrika, 39*, 31–36.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of educational Measurement*, *38*(4), 319-342.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement*, *2*(3), 135-170.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17–64). American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Karami, H. (2012). The development and validation of a bilingual version of the Vocabulary Size Test. *RELC Journal*, *43*(1), 53-67.

Karami, H. (2015). Exploratory factor analysis as a construct validation tool: (Mis) applications in applied linguistics research. *TESOL Journal*, *6*(3), 476-498.

LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing*, *34*(4), 451-475.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15-30.

Lerdorf, R. (1995). PHP: Hypertext Preprocessor. http://php.net

Linacre, J. M. (2012). *Practical Rasch measurement.* Retrieved from www.winsteps.com/tutorials.htm

Linacre, J. M., & Wright, B. D. (2000). Winsteps. http://www.winsteps.com/index.htm

Liu, H.-m. (2014). *Investigating the relationships between a reading test and can-do statements of performance on reading tasks.* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses database. (3607916).

Luoma, S. (2004). *Assessing speaking.* Cambridge University Press.

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*(6), 741-760.

Medical Outcomes Trust Scientific Advisory Committee (1995). *Instrument Review Criteria. Medical Outcomes Trust Bulletin, 3*, 1–4.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing.* John Benjamins.

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.

Qian, D. D. (2008). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. Language Learning, 52(3), 513–536.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedogogiske Institut.

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.

Razmjoo, S. A., & Heydari Tabrizi, H. (2010). A content analysis of the TEFL MA Entrance examinations (Case study: Majors courses). *Journal of Pan-Pacific Association of Applied Linguistics*, *14*(1), 159-170.

Riazi, A.M. & Razavipour, K. (2011). Agency of EFL teachers under the negative backwash effect of centralized tests. *International Journal of Language Studies (IJLS), 5*(2), 122-142.

Romer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, *7*(1), 141–163.

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*(2), 261-274.

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *53*(1), 109-120.

Schmitt, N., Ng, J. W. C., & Garras, J. (2011). The word associates format: Validation evidence. *Language Testing*, *28*(1), 105-126.

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment, 29*, 304–321.

Shiotsu, T., & Weir, C.J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, *24*(1), 99-128.

Stæhr, L. S. (2008). Vocabulary size and the skills of reading, listening and writing. *Language Learning Journal, 36*(2), 139–152.

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.

Sun, Y. (2016). *Context, construct, and consequences: Washback of the college English test in China.* Retrieved from ProQuest Dissertations & Theses database. (10155303).

Wolfe, E. W. & Smith Jr., E. V. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I – Instrument development tools. *Journal of Applied Measurement, 8*(2), 97-123.

Wolfe, E. W. & Smith Jr., E. V. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II – Validation activities. *Journal of Applied Measurement, 8*(2), 204-234.

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing*, *32*(2), 199-225.

Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *The Modern Language Journal*, *96*(4), 558-575.

Citation:
Rafatbakhsh, E., & Ahmadi, A. (2022). The argument-based validation of a large-scale high-stakes vocabulary test. *Practical Assessment, Research, & Evaluation*, 27(28). Available online: https://scholarworks.umass.edu/pare/vol27/iss1/28/

Corresponding Author:
Elaheh Rafatbakhsh
Shiraz University

Email: e.rafatbakhsh [at] shirazu.ac.ir