ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(4); 305-327

Validation of the Vocabulary Size Test

Mustafa GÖKCAN * Derya ÇOBANOĞLU AKTAN **

Abstract

The Vocabulary Size Test (VST) is one of the most commonly used assessment tools for measuring English vocabulary size in the field of language testing. Despite its common usage, only a limited number of validity and reliability studies have been carried out with regard to the VST. Besides, they were mostly predicated on the Rasch model. This validation study has attempted to reveal evidence for construct validity for the VST, and to this end, item response theory (IRT) analyses were performed based on the three-parameter logistic model (3PLM). The assumptions of IRT were investigated via factor analysis (unidimensionality) and Yen's Q3 statistic (local independence). Detailed differential item functioning (DIF) analyses were conducted with Mantel-Haenszel, Lord's chi-square test, and Logistic regression methods to add evidence based on internal structure and to check fairness as a lack of measurement bias. The validation results with IRT showed that the 3PLM fitted the data better than the one- and the two-parameter logistic models. DIF results indicated that 10 items exhibited large DIF (seven favoring males and three favoring females). The results further showed that the guessing effect was not negligible for the VST.

Keywords: Language testing, vocabulary assessment, Vocabulary Size Test, item response theory, differential item functioning

Introduction

Vocabulary size is of pivotal importance in almost every aspect of learning a foreign language (Daller et al., 2007). As also echoed by Alderson (2005), "language ability is to quite a large extent a function of vocabulary size" (p. 88). Despite its importance, vocabulary size has been an oft-neglected aspect of language learning (Meara, 1980), and only recently has it drawn attention in applied linguistics and language teaching (Nation, 2013). A growing body of studies conducted on the vocabulary size of English learners indicated that it is a significant indicator of language ability (Milton, 2009). Significant positive correlations were found between English vocabulary size and listening (Li, 2019; Noreillie et al., 2018), reading (Zhang & Zhang, 2020), speaking and writing skills in English (Milton, 2013; Miralpeix & Muñoz, 2018) and, especially related to reading comprehension, it was emphasized that vocabulary size was the most significant predictor (Stæhr, 2008). Although the number of studies investigating vocabulary size has recently seen a significant increase, new assessment tools for measuring English vocabulary are rarely seen in the field (Mizumoto et al., 2019). There are also very few studies examining the validity and practicality of the available tools.

According to Standards for Educational and Psychological Testing, validity means "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests." (AERA, APA & NCME, 2014, p. 11). Moreover, validity is the most fundamental characteristic of a test. It is also important to note that validity is a property of the test scores, not the test itself; in other words, having a particular validity study for a specific test does not guarantee validity in other contexts. For instance, a test can have high validity for a certain group of examinees but can have lower-level validity for other groups. It has been suggested that researchers should collect evidence for validity before they use a particular test's results for their research purposes (AERA, APA & NCME, 2014).

In language testing, the importance conferred to test validity has increased considerably in recent years. Schmitt et al. (2020) put forward that early examples of vocabulary tests generally lack appropriate

* Research Assistant, Hacettepe University, Faculty of Education, Ankara-Türkiye, gokcan.m@gmail.com, ORCID ID: 0000-0002-2284-9967
** Assistant Professor, Hacettepe University, Faculty of Education, Ankara-Türkiye, coderya@gmail.com, ORCID ID: 0000-0002-8292-3815

To cite this article:
Gökcan, M., & Çobanoğlu-Aktan, D. (2022). Validation of the vocabulary size test. *Journal of Measurement and Evaluation in Education and Psychology, 13*(4), 305-327. https://doi.org/10.21031/epod.1144808

Received: 18.07.2022
Accepted: 7.11.2022

validation examinations. They also state that "the typical practice seems to be to develop a test, get a journal article published on it, and then move on to the next project" (p. 2). If tests developed in this manner are utilized in low-stakes contexts, a lack of validation studies may not lead to any significant problems. However, since these tests are used in studies focusing on second language and foreign language acquisition, they affect theoretical and pedagogical developments (Schmitt et al., 2020).

Item response theory and language testing

Briefly stated, item response theory (IRT) models show the connection between a test item and an ability or a latent trait (indicated by the symbol "θ") measured by that test (DeMars, 2010). The first IRT model is the normal ogive model, and the response function used in this model is given in equation 1. Birnbaum (1968) changed the normal ogive function given in equation 1 with the logistic model (equation 2), which is more statistically applicable (van der Linden & Hambleton, 1997). Working with logistic functions is easier than the normal ogive ones because the latter require mathematical integration (De Mars, 2010).

$$P_i(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/z} dz. \quad (1)$$

$$P_i(\theta) = \frac{1}{1+\exp\{-a_i(\theta - b_i)\}}. \quad (2)$$

In the function given in equation 2, P indicates the probability of responding correctly to item "i" at a given "θ" ability level. Parameter b is the point on the ability scale where the probability of a correct response is 0.5. This parameter is also called "location parameter" and shows the position of the item characteristic curve (ICC) on the ability scale. The parameter "a" gives the curve of the ICC at the point where parameter b is located on the ability scale.

Later on, factor D was added to equation 2, and the function was formed like in equation 3. Factor D is a scaling factor and is used to make the logistic function estimates as similar as possible to normal ogive function estimates (de Ayala, 2009; Hambleton et al., 1991). If the value of factor D is equated to the constant 1.7, the logistic function is located on the same metric with the normal ogive function. By this way, for all values of θ, it is possible to get estimates differing in absolute value by less than 0.01 (Camilli, 1994).

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad (3)$$

Birnbaum suggested adding a third parameter to explain the performances of low ability individuals different from zero in multiple-choice items or tests. According to him, the scores different from zero do not result from the possibility of responding correctly. After adding the third parameter, "c", the equation is formed as in equation 4.

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad (4)$$

Despite the fact that equation 4 does not indicate a logistic function anymore, the model is still known as the three-parameter logistic model (van der Linden & Hambleton, 1997). It is important to know that parameter c does not vary as a function of "θ". For this reason, the probability of responding correctly by guessing is the same for low and high-ability individuals (Baker, 2001).

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad (5)$$

The introduction of IRT to the field of language testing essentially came too late. This introduction took place with the Rasch model, and today, it can be seen that this one-parameter model is broadly used in the field (Aryadoust et al., 2020). The function that forms part of the Rasch model (Rasch, 1980) is given in equation 5. This model is less complicated than the two- or three-parameter logistic models and only comprises the difficulty parameter. The utilization of IRT models like the two- or three-parameter logistic models is more limited than the Rasch model. The Rasch – one parameter – model has generally been considered inadequate to indicate the item characteristics in full measure and has been found too simple in the community of educational measurement researchers. However, applied linguistics researchers immediately embraced the Rasch model, whose simplicity is actually quite deliberate (McNamara & Knoch, 2012).

The researchers in language testing generally do not enter the field as graduates of statistics or psychometrics but as graduates of language teaching. For this reason, their background in mathematics and statistics is not so strong, and, accordingly, it can be said that the use of IRT analyses in language testing was a little bit delayed. Moreover, unidimensionality, an IRT assumption, held language researchers back from IRT because the fact that language proficiency is multidimensional in nature and that there are lots of variables intervening in the language learning process gave them the impression that the use of IRT is not appropriate for language studies (McNamara & Knoch, 2012). However, after the introduction of multidimensional IRT models (Reckase, 2009), an approach to the effective use of IRT in language testing was opened (Ockey & Choi, 2015).

Vocabulary Size Test

The Vocabulary Size Test (VST) was developed by Nation and Beglar in 2007 to measure English vocabulary size. The VST was formed by selecting 140 words from the most frequently used 14,000 words according to the British National Corpus (BNC). Firstly, 14,000 words were split into 14 levels (1000 words in each level), and then a sample of 10 words was selected from each level. The words in the BNC are ordered according to the frequency of use in English texts. The frequency of use of a word decreases as its order in the list increases. Thus, among the 14 levels of the VST, the items in the first level are envisaged to be easier than the ones in later levels. The VST items are provided in multiple-choice format. The item stems are kept short so that any variable other than vocabulary knowledge does not affect the examinees' responses. Here is an example item from the first level.

4. FIGURE: Is this the right **figure**?

a. answer

b. place

c. time

d. number

Bilingual versions of the VST have been developed in various languages to date. Nevertheless, other than the original form of the VST, only a few studies have examined its reliability and validity. One such is a Rasch-based study carried out by Beglar (2010). Different versions of the VST, relying on the study by Beglar, have not sought further evidence for the validity of the original VST in their works. Thus, the issues in the original version have not been fully handled, and these issues have also remained

in these different versions (Schmitt et al., 2020). Information related to these bilingual versions and Beglar's study is presented in the next section.

Previous Research

There are bilingual versions of the VST in Arabic, Gujarati, Japanese, Korean, Mandarin, Persian, Russian, Tamil, Thai, and Vietnamese. The development and validation studies of some of these bilingual versions have also been published in various journals with a high impact factor (Elgort, 2012; Karami, 2012; Nguyen & Nation, 2011; Zhao & Ji; 2018). Two sample items from these bilingual versions are presented below. As seen in the items, the bilingual versions have the same question stems, but the choices are in the native tongue of the respondents.

Russian Version (Elgort, 2012)	Vietnamese Version (Nguyen & Nation, 2011)
4. FIGURE: Is this the right **figure**?	4. FIGURE: Is this the right **figure**?
a. ответ	a. câu trả lời
b. место	b. địa điểm
c. время	c. thời gian
d. цифра	d. con số

As stated earlier, these bilingual studies searched for evidence of the validity of their bilingual versions, but not for the original version. They mostly relied on Beglar's Rasch-based validity study. In his study, Beglar (2010) carried out a detailed investigation into the validity of the VST, the findings of which demonstrated that most of its items showed acceptable fits to the Rasch model, and that the VST had a high degree of psychometric unidimensionality when item residuals were analyzed. The VST possesses a high degree of measurement invariance, as evidenced by the similar ability parameters produced by different forms of the VST.

Different from previous studies, Zhang (2013) investigated whether the addition of an "I don't know" option affected the scores of the VST. To this end, he applied three different versions of the VST to 150 university students in China. The first version was the original VST. The "I don't know" option was added to the second version. Additionally, in the third version, a penalty for incorrect answers was also added to the scale. The penalty comprised a one-point deduction for each wrong answer. Zhang (2013) found that the number of guesses significantly decreased in the second and third versions of the scale. But this also decreased the number of correct responses given with partial knowledge. Based on the findings, Zhang suggested that the second or the third versions of the scale be adopted, rather than the original version, to measure the precise word knowledge in order to eliminate the guessing effect.

Purposes of the Study

The purpose of this study is to collect evidence related to the validity of the Vocabulary Size Test (VST) developed by Nation and Beglar (2007) by comparing different item response theory (IRT) models. There is a particular need for a three-parameter logistic model (3PLM) validation study for the VST to examine the guessing effect, which might be conducive to overestimation in VST results (Stewart, 2014). In previous studies, the one-parameter logistic model and the Rasch model were used to validate the VST and to analyze the results. In this study, by comparing different IRT models with the three-parameter model which considers the chance factor, related gaps in the literature have been addressed.

Moreover, there is no detailed differential item functioning (DIF) study for the VST in the literature. DIF occurs when the possibility of responding correctly to a particular item differs as a function of a specific group membership. According to Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014), DIF poses a major threat to fairness in testing because it can lead to biased ability

estimations. The first step in detecting item bias is detecting potentially biased items by making DIF analyses (Çepni & Kelecioğlu, 2021; Uysal et al., 2019). Historically, most DIF studies have focused on group differences based on gender or race (Kıbrıslıoğlu Uysal & Atalay Kabasakal, 2017; Zumbo, 2007). In our study, we also investigated gender-related DIF, and by detecting potentially biased items of this significant test, suggestions have been made for further studies and to improve the quality of the test.

Method

Study Group

At the beginning of the study, the intention was to collect data with a paper-pencil format VST. However, due to the pandemic, paper-pencil format data collection was not possible. For this reason, the VST data were collected in an online form by sharing a link via e-mail. The link was shared with 4500 university students from seven different universities (four state – three private). Eight hundred and fifty-four students voluntarily responded to the test. Since this number is not enough for our study, research assistants who are students of Master's and Ph.D. programs were also added to the study group. Then, 4000 research assistants were sent e-mails, 781 of whom responded to the VST. In this way, we reached a total number of 1622 voluntary students.

Data Collection

After obtaining the required permissions from the Hacettepe University Ethics committee (Document number: 35853172-300-E.00001113493) for this study, the data were collected via the 140-item version of the VST. This version can be found by following the link below: (https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/Vocabulary-Size-Test-14000.pdf). The online version of the VST was generated using Google Forms. The link for the online VST version was sent to undergraduate and graduate students via e-mail. In the online form, some information about the VST was provided prior to the test. It was also stated that they could skip any questions that included items that were unfamiliar to them and that there was no time limitation. Since the participants volunteered to learn their own vocabulary levels, it was assumed that they did not cheat as they took the online test. The data collection tool also included items about the participants' demographics, such as their level of education, gender, and English proficiency test score (TOEFL or the Foreign Language Proficiency Exam).

Data Analysis

We carried out various analyses to collect evidence of the validity of our assessment tool, the VST. Throughout our validation study, we followed the suggestions offered in "Standards for Educational and Psychological Testing" (AERA, APA & NCME, 2014). As stated in the Standards, there are various sources of validity evidence that can be used to evaluate the validity of an intended interpretation of test scores for a specified use. In different settings, varying combinations of these sources might be used. There is no requirement that every single source should be used in all the validation processes. The sources of evidence listed by the Standards are content, relation to other variables (convergent validity), internal structure, response processes, and consequences of testing. In this study, we gathered evidence for all of these sources, with the exception of the consequences of testing.

Evidence Regarding Internal Structure Validity

With regard to internal structure validity, we carried out 3PLM-based IRT and DIF analyses. Internal structure validity refers to construct validity evidence. By analyzing the internal structure of a test, we examine the relationships between the test items which conform to a construct. In this study, the

measured construct is the vocabulary size. By finding the best-fitting IRT model, we confirm that the items in the VST measure the vocabulary size construct.

For internal structure, IRT analyses were conducted using the R software with the "mirt" package (Chalmers, 2012). The model that the data fitted the best was tested with the ANOVA function in the same package. The unidimensionality and local independence assumptions of IRT were checked prior to IRT analyses of the VST.

In the literature, there are three commonly used methods for determining dimensionality, namely the Kaiser rule (Kaiser, 1960), parallel analysis (Horn, 1965), and scree plot (Cattel, 1966; Cho et al., 2009). Weng and Cheng (2005) showed that parallel analysis produced good estimates in the dimensionality analyses with dichotomous items, although there was a risk of obtaining meaningless dimensions. However, Tran and Formann (2009) found the reliability of parallel analysis to be too low when they worked with dichotomous items and Pearson correlation. Moreover, no improvement was observed when tetrachoric correlation was used. For this reason, for the dimensionality analysis of VST, parallel analysis was not preferred. Instead, the number of dimensions was decided by examining the scree plot and the associated eigenvalues. It was investigated whether there was a dominant dimension.

An explanatory factor analysis (EFA) was carried out and weighted least square mean and variance adjusted (WLSMV) was selected as the estimator. WLSMV utilizes tetrachoric correlation for factor extraction. When the factor analysis is carried out with continuous variables, and the data meet the assumption of univariate and multivariate normality, maximum likelihood (ML) estimation methods should be used, and when it is conducted with categorical variables, the least squares methods are recommended (Koyuncu & Kılıç, 2019). It has been found that, when compared to ML methods, WLSMV is better with large models that include categorical or binary data in terms of statistical performance and duration of the analysis (Muthen et al., 1997), and indeed that it can make less biased estimations (Li, 2016).

Yen's (1993) Q3 statistics between item pairs were calculated to test the local independence assumption. De Ayala's (2009) suggestions were followed to determine a cutoff value for the Q3 statistic. A 140x140 matrix was examined to detect potentially dependent item pairs.

Differential item functioning (DIF) analyses were run through the "difR" package (Magis et al., 2010) of R with the Logistic regression method, Lord's chi-square test, and the Mantel-Haenszel method. The difLogistic, difLord, and difMH functions were employed, and then the dichoDif function was run to make comparisons to determine the items that are flagged as DIF items by all of the three methods. The DIF statistics of the items showing large DIF are given in a table. Moreover, item characteristic curves (ICC) of these large DIF items were drawn using the R software.

Evidence Regarding Content Validity

The VST items were written as representative as possible of the English vocabulary corpus by the developers of the original version, and by this way, they provided content-related validity evidence in their work. We also investigated other sources of content validity and generated a person-item map (Wright map) in R to check whether there is a sufficient number of items in the VST and whether they spread fairly on the ability scale of the IRT model.

Evidence Regarding Convergent Validity

For evidence considering relations to other variables (convergent validity), correlations between the VST scores and the scores from two English proficiency exams, namely TOEFL and the Foreign Language Proficiency Exam (FLPE), were examined. The FLPE is a national English proficiency exam applied in Turkey. Evidence regarding response processes examines whether participants answer the questions the way the test developers intended. Although this requires collecting evidence through think-aloud processes, in this study, we indirectly collected evidence to probe the impact of responding correctly by chance by including the guessing effect in the IRT model.

Descriptive Statistics

Before presenting the findings of the study, it will be of value to review some descriptive statistics briefly. Since 165 of the participants completed the test after responding to just the first few questions, the data related to those 165 respondents were removed, and the descriptive statistics of the remaining 1457 students are presented in Table 1, Table 2, and Figure 1.

According to Table 1, 823 of the respondents are female, and 634 of them are male students. In the study group, there are 49 preparatory, 690 undergraduate, 181 master's degree, and 537 Ph.D. students. Descriptive statistics and box-plot for participants' VST scores are shown below.

Table 1

The Students' Educational Levels by Gender

	Preparatory	Undergraduate	Master	PhD	Total
Female	27	395	93	308	823
Male	22	295	88	229	634
Total	49	690	181	537	1457

According to Table 2, the mean score of females is 68.51, and it is 70.5 for males. The mean score of the entire group is calculated as 69.38. When we examine the values of skewness and kurtosis, we can see that the test score distribution does not depart from the normal distribution too much.

Table 2

Scores by Gender

	Minimum	Maximum	Mean	Standard Error	Skewness	Kurtosis
Female	2	135	68.51	26.65	- 0.08	- 0.42
Male	5	136	70.50	27.97	- 0.09	- 0.63
Total	2	136	69.38	27.24	- 0.08	- 0.52

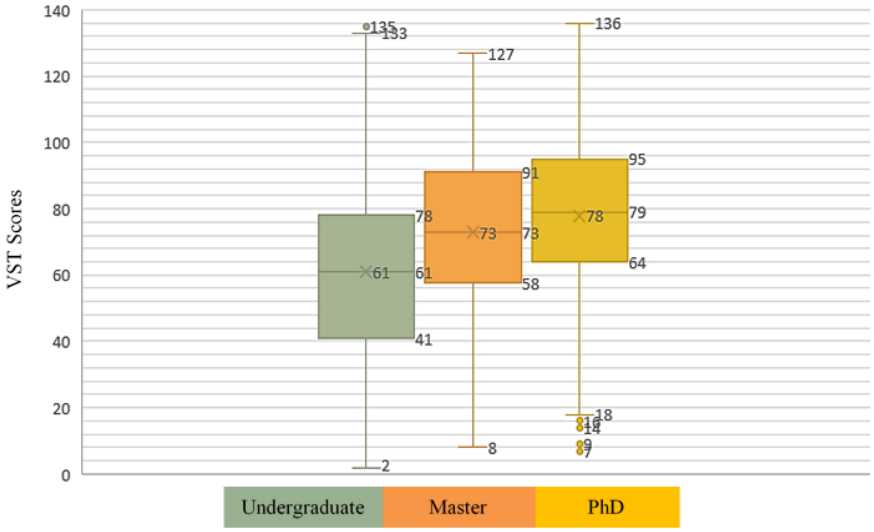
In Figure 1, the box and whisker plot of the VST scores of the respondents is presented. For this plot, the preparatory students are included in the undergraduate group. In the figure, the upper part of a box is the third quartile, the lower part is the first quartile, the number next to the x is the mean score, and the line represents the median.

As may be seen from the plot, all scores (quartiles, means, and medians) increase based on the educational levels of the participants. The means for undergraduate, master and Ph.D. students were found at 61, 73, and 78, respectively. This finding can be considered evidence for the fact that the VST distinguishes students from different education levels. Education level reflects the students' English proficiency to some extent because to become a research assistant and to study in graduate programs,

different levels of English proficiency are required in Turkey. The proficiency level demanded for PhD programs is higher than that of master's programs, and for undergraduate programs it is too much lower. Based on these results, it is clear that the VST is also able to distinguish students at different English proficiency levels.

Figure 1

Box and Whisker Plot of the Scores and Education Levels of the Students



Data screening and cleaning were carried out prior to validation analyses. First of all, missing data and out-of-range values were checked. The questions skipped were regarded as incorrect answers because skipping a question means that the respondent does not know the meaning of the word used in the item. Since the data were collected via online forms, no missing values and no out-of-range values were found, and accordingly, there were no univariate outliers either. The Mahalanobis distance was calculated for each respondent to detect multivariate outliers. We also calculated *p*-values for every distance to see if any were statistically significant. It was found that 170 observations had p-values less than .001, and they were considered to be a multivariate outliers. They were excluded from the data, and the remaining analyses were carried out with the data of 1287 respondents.

Findings

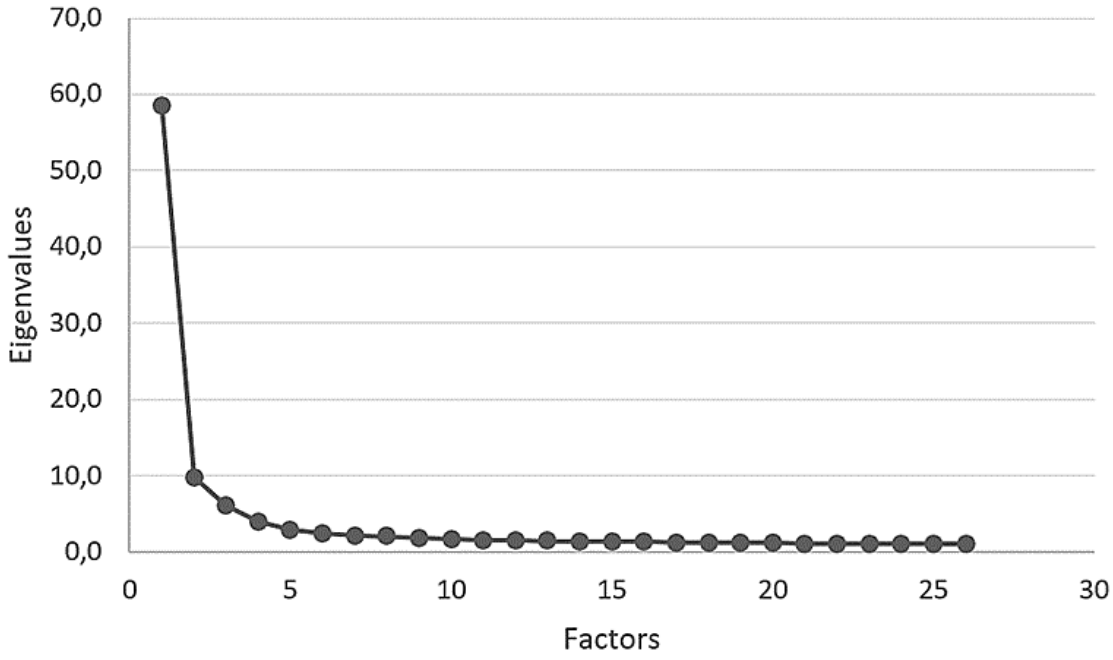
Findings of Evidence Regarding Internal Structure Validity – IRT

Before conducting the IRT analyses of the VST, we tested unidimensionality and local independence, which are two main IRT assumptions.

After the factor analysis was performed to investigate the dimensionality of the VST, it was observed that there was a dominant dimension. A dominant dimension has been considered sufficient for meeting the unidimensionality assumption in IRT analyses (Hambleton & Swaminathan, 1985). In Figure 2, the scree plot showing the eigenvalues of the factor analysis is shown. It is seen that the eigenvalue of the first dimension is almost six times bigger than the eigenvalue of the second one.

Figure 2

The Scree Plot of Eigenvalues and Factors



Moreover, the model fit indices in Table 3 illustrate that the unidimensional model fits the data well. Although the two, three, and four-dimension models had better fit indices than the unidimensional model, the VST was reckoned as unidimensional.

Table 3

Exploratory Factor Analysis Model Fit statistics

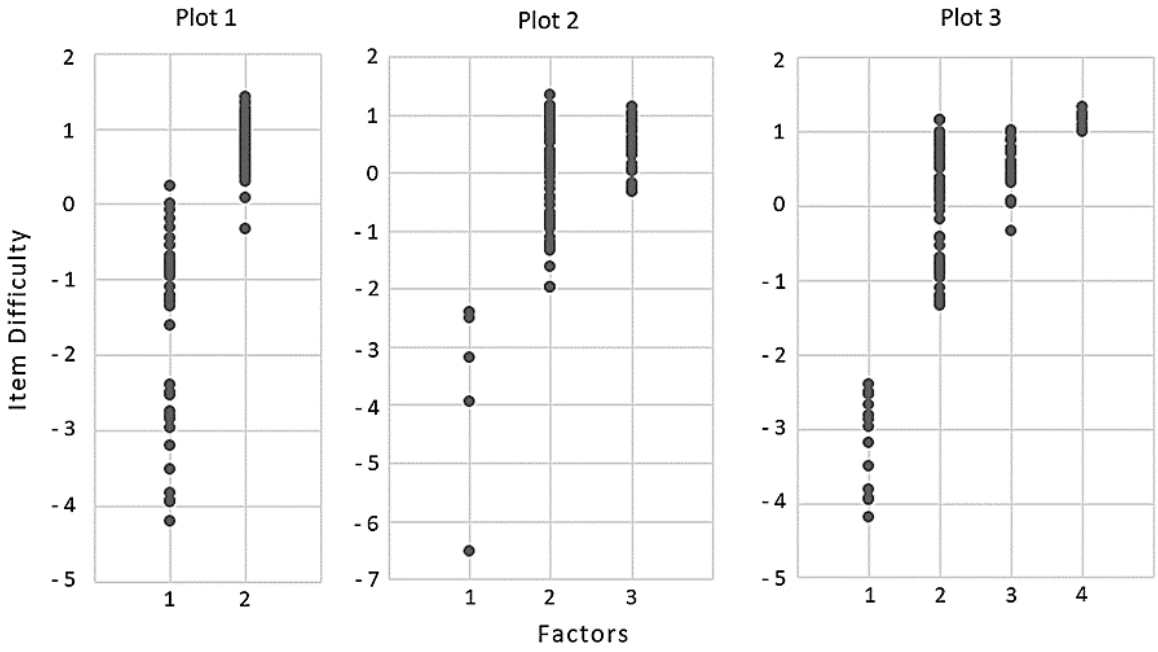
MSodel	X^2	df	X^2/df	RMSEA	CFI	TLI	SRMR
1 Factor	19167.1	9590	1.99	0.028	0.934	0.933	0.101
2 Factor	14514.7	9451	1.53	0.020	0.965	0.964	0.074
3 Factor	11842.6	9313	1.27	0.015	0.983	0.982	0.060
4 Factor	11118.9	9176	1.21	0.013	0.987	0.986	0.055

The reason behind the multidimensional findings is difficulty factors. The history of the problem of "difficulty factors" dates back to almost a century ago (Spearman, 1927; Hertzman, 1936), and it is encountered frequently in factor analyses of binary-scored items (see Hattie, 1985, for more detail). It is known that when the items of a test vary in difficulty parameter to a large extent, "spurious" factors are extracted according to item difficulty regardless of item content (McDonald & Ahlawat, 1974; Yang & Xia, 2015). This problem is generally named "spurious/artificial factors" or "difficulty factors", and it sometimes causes simple constructs like vocabulary knowledge to seem multidimensional (Reckase et al., 1988).

In Figure 3, there are scatter plots that show the relations between each of the two, three, and four-dimensional models and the difficulty parameters of the items in related dimensions. From plot 1, it is seen that the difficulty parameters of the items in the first dimension are mostly between -4 and 0, while the ones in the second dimension are between 0 and 1.5. In the two-dimensional model, the mean of the item difficulties for the first dimension is -1.57, and it is 0.76 for the second dimension. When Plot 2 and Plot 3 were examined, it was seen that the intervals of the difficulty parameters belonging to the first dimension have lower values than those belonging to other dimensions and that the values of the intervals increase respectively for other dimensions.

Figure 3

The Scatter Plots of Item Difficulties and Factors



In the three-dimensional model, for the first dimension, the mean of item difficulties was calculated as -3.70. It is 0 for the second dimension and 0.45 for the third dimension. In the 4-dimensional model, -3.18, 0, 0.54, and 1.17 are the mean of item difficulties for the first, second, third, and fourth dimensions, respectively. Briefly, when we examine both the outputs related to dimensionality and the item contents, the reason for multidimensionality can be explained as difficulty factors.

Table 4

Locally Dependent Item Pairs

Item	Items	Item	Items	Item	Items	Item	Items
1	2,5,6,8	17	19,21	38	35		
2	1,5,6,8,19,21	19	2,17,21	42	34,43	103	26,30,70,74, 94,105,107
5	1,2,6,8	21	2,6,17,19,22	43	34,42		
6	1,2,5,8,21	22	21	62	65	105	70,94,103, 107,117
7	10	26	103	65	62	107	70,103,105
8	1,2,5,6	30	103	70	103,105,107	116	140
10	7,11	34	42,43	74	103	117	105
11	10	35	38	94	103,105	140	116

Q3 statistics, which show the relations between item residuals, were estimated to investigate the assumption of local independence. Since Q3 is a correlational statistic, its value ranges between -1 and +1, and a high absolute value of Q3 indicates a significant violation of local independence (Paek & Cole, 2020). As a cutoff value for the Q3 statistic, de Ayala (2009) suggested $|Q3| \geq \sqrt{0.5}=.2236$. Since the VST has 140 items, a 140x140 matrix (16900 cells) was examined for detecting potentially dependent item pairs. The item pairs which have Q3 statistics above .2236 were flagged as locally dependent items. It was found that among the 16900 cells, 74 of them have Q3 values higher than the cutoff value and that the 74 cells belong to 30 different items. In Table 4, these 30 items are shown in the "Item" columns, and the items which have high Q3 values with those 30 items are in the "Items" columns. We also reviewed item pairs that are potentially dependent, but couldn't see any cause of dependency. Incorrect or correct replies to a VST item should not result in incorrect or correct replies to another VST item. This is because VST items are vocabulary items, each of which asks for a different vocabulary. The item stems are very short, and there isn't any item pair which has the same item stem. Having a common passage or item stem is not the only source of dependency. According to Ackerman (1987), item parameters (i.e., discrimination and difficulty) and the order of the items (e.g., easy to hard or hard to easy) can also lead to local independence. In our case, the VST items are ordered from easy to hard, and it can be seen from Table 4 that the locally dependent item pairs are mostly neighboring items. If one wants to examine the contents of the item pairs which violate local independence, s/he can reach VST by clicking the link in the section "Assessment Tool".

Table 5

The Comparison of the 1PLM with the 2PLM

	AIC	AICc	SABIC	HQ	BIC	logLik	X^2	Df	p
1PLM	155622.6	155657.6	155902.3	155895.7	156350.2	-77670.29	NaN	NaN	NaN
2PLM	152677.1	152833.5	153232.5	153219.5	154121.9	-76058.56	3223.472	139	0

After testing the assumptions of unidimensionality and local independence, IRT analyses were carried out to determine which IRT model fits the data best. Firstly, the estimations were made with the one- and the two-parameter logistic models. Then two models were compared by conducting a likelihood ratio test and by examining AIC, AICc, SABIC, and BIC model fit indices with "ANOVA". Table 5 displays the results of the likelihood ratio test. In the two-parameter model, there are decreases in the values of AIC, SABIC, and BIC. Besides, a smaller logLik value was calculated. The *p*-value of the likelihood ratio test was estimated as zero, and this means that the 2PLM fits the data better than the 1PLM.

Table 6

The Comparison of the 2PLM with the 3PLM

	AIC	AICc	SABIC	HQ	BIC	logLik	X^2	Df	p
2PLM	152677.1	152833.5	153232.5	153219.5	154121.9	-76058.56	NaN	NaN	NaN
3PLM	152254.2	152662.5	153087.3	153067.7	154421.4	-75707.08	702.946	140	0

After it was found that the two-parameter model had a better fit than the one-parameter model, the same test was carried out to compare the 2PLM with the 3PLM. Table 6 exhibits the results of this comparison. Although there are decreases in the values of AIC, SABIC, and BIC model indices again, these decreases are not as large as in the comparison of the 1PLM and the 2PLM. Moreover, the increase in the value of logLik is not too much, but the *p*-value of the likelihood test is significant, and this indicates that the 3PLM fits the data better than the 2PLM.

Findings of Evidence Regarding Internal Structure Validity – DIF

Differential item functioning analyses were carried out with Logistic regression, Lord's chi-square test, and Mantel-Haenszel methods. There are 34 items that were flagged as DIF items by all three methods. These items are listed in Table 7, and the results of the three DIF methods are visualized in the plots given in Appendix A.

According to the results of the Logistic regression and Lord's chi-square methods, there isn't any item showing a large DIF. However, the Mantel-Haenszel results indicate that, among the 34 items, 24 items show negligible or moderate DIF, and 10 items show large DIF. If the absolute value of the Δ MH for a particular item is higher than 1.50, the item is considered to exhibit a large DIF (Magis et al., 2010). The large DIF items are items of 3, 7, 17, 20, 63, 72, 74, 98, 104, and 138 as displayed in bold in Table 7. It also shows the DIF statistics for the 34 items. The LRT statistics of the DIF items, the *p*-value related to that statistic and Nagelkerke's R^2 (Nagelkerke, 1991) are given in the results of the Logistic regression method. In the results of Lord's chi-square method, Lord's χ^2 statistic, and the *p*-value of that statistic are provided. In the results of Mantel-Haenszel, on the other hand, besides chi-square and *p*-values, α MH and Δ MH values are also presented. As shown in Table 7, the *p* values calculated in all three methods of items showing DIF are smaller than .05. We can conclude the items that show DIF in favor of males and females by examining the deltaMH values (Magis et al., 2010). When the deltaMH (Δ MH) value is negative, it indicates DIF in favor of the reference group, and when it is positive, DIF is in favor of the focal group (Holland & Thayer, 1988). Females were predetermined as the reference group in the codes written for the DIF analysis. It is seen that, among the items which exhibit large DIF, the items which have negative Δ MH values are the items of 72, 74, and 138. These items exhibit DIF in favor of females, and the vocabulary included in these items are *palette, kindergarten, and erythrocyte*, respectively. Moreover, the ICCs of these DIF items are shown in Appendix B. When the ICCs are examined, it is seen that, for females, the possibility of responding correctly to these items is higher on almost every level of the ability scale. Items 3, 7, 17, 20, 63, 98, and 104, which have positive Δ MH values, show large DIF in favor of males. The vocabulary asked in these items are *period, jump, pub,*

pro, stealth, crowbar, and counterclaim, respectively, and the ICCs of these items are presented in Appendix C.

Table 7

Items Showing DIF and their DIF Statistics

	Logistic Regression			Lord's chi-square		Mantel-Haenszel			
ITEM	LRT Statistic	p-value	R^2	Lord's χ^2	p-value	MH χ^2	p-value	α MH	Δ MH
ITEM 3	**20.1307**	**0.0000**	**0.0305**	**14.3176**	**0.0008**	**15.5451**	**0.0001**	**0.3939**	**2.1892**
ITEM 7	**22.9162**	**0.0000**	**0.0288**	**10.7711**	**0.0046**	**15.7926**	**0.0001**	**0.4775**	**1.7370**
ITEM 14	13.8811	0.0010	0.0096	8.9157	0.0116	9.9325	0.0016	0.6513	1.0078
ITEM 16	10.7591	0.0046	0.0075	7.9392	0.0189	10.0675	0.0015	0.6629	0.9663
ITEM 17	**14.9093**	**0.0006**	**0.0297**	**9.6579**	**0.0080**	**9.9347**	**0.0016**	**0.2531**	**3.2290**
ITEM 20	**56.2356**	**0.0000**	**0.0381**	**33.9760**	**0.0000**	**49.5375**	**0.0000**	**0.4083**	**2.1048**
ITEM 25	10.8871	0.0043	0.0139	9.1278	0.0104	9.7025	0.0018	0.5439	1.4310
ITEM 31	12.1880	0.0041	0.0137	13.2703	0.0033	7.6450	0.0087	0.5606	1.3603
ITEM 32	9.3430	0.0094	0.0060	10.8923	0.0043	8.2690	0.0040	1.4813	-0.9234
ITEM 34	19.5076	0.0001	0.0107	21.1253	0.0000	15.2666	0.0001	1.7712	-1.3434
ITEM 55	12.6630	0.0018	0.0074	9.8334	0.0073	7.9740	0.0047	0.6683	0.9472
ITEM 59	14.4812	0.0007	0.0077	20.3274	0.0000	13.8536	0.0002	1.7509	-1.3163
ITEM 62	11.0667	0.0040	0.0098	12.7461	0.0017	11.9657	0.0005	1.8575	-1.4552
ITEM 63	**59.7718**	**0.0000**	**0.0313**	**42.5657**	**0.0000**	**48.6764**	**0.0000**	**0.3481**	**2.4799**
ITEM 69	8.2869	0.0159	0.0047	10.1917	0.0061	6.0796	0.0137	1.4578	-0.8857
ITEM 72	**39.5102**	**0.0000**	**0.0323**	**49.7523**	**0.0000**	**24.5789**	**0.0000**	**2.3489**	**-2.0068**
ITEM 74	**19.3315**	**0.0001**	**0.0149**	**31.8527**	**0.0000**	**15.2801**	**0.0001**	**2.5784**	**-2.2258**
ITEM 82	7.8601	0.0196	0.0037	13.1187	0.0014	5.7513	0.0165	1.4983	-0.9501
ITEM 90	6.1641	0.0459	0.0056	7.7200	0.0211	6.1266	0.0133	1.4817	-0.9241

Table 7

Items Showing DIF and their DIF Statistics (Continued)

ITEM 92	13.3572	0.0013	0.0079	10.6864	0.0048	13.1646	0.0003	0.5860	1.2560
ITEM 93	9.6763	0.0079	0.0053	12.1010	0.0024	6.4793	0.0109	1.4568	-0.8842
ITEM 98	**41.6161**	**0.0000**	**0.0234**	**30.4000**	**0.0000**	**38.0141**	**0.0000**	**0.3789**	**2.2806**
ITEM 104	**31.2653**	**0.0000**	**0.0153**	**20.8562**	**0.0000**	**33.5489**	**0.0000**	**0.4063**	**2.1163**
ITEM 109	17.3575	0.0002	0.0089	11.0125	0.0041	16.6053	0.0000	0.5471	1.4175
ITEM 114	16.1258	0.0003	0.0080	8.8901	0.0117	14.5232	0.0001	0.5423	1.4380
ITEM 115	9.4762	0.0088	0.0049	13.0773	0.0014	6.6750	0.0098	1.4819	-0.9243
ITEM 116	11.5260	0.0031	0.0113	9.2755	0.0097	7.4143	0.0065	1.6128	-1.1232
ITEM 122	9.6359	0.0081	0.0056	7.7227	0.0210	8.7972	0.0030	0.6240	1.1084
ITEM 123	9.7038	0.0078	0.0050	5.9939	0.0499	7.7186	0.0055	0.6616	0.9709
ITEM 124	10.1171	0.0064	0.0058	12.0558	0.0024	8.3024	0.0040	1.5783	-1.0724
ITEM 128	11.6936	0.0029	0.0061	16.1651	0.0003	7.1759	0.0074	1.5005	-0.9537
ITEM 130	6.0630	0.0482	0.0037	9.7588	0.0076	3.9663	0.0464	1.3197	-0.6519
ITEM 138	**24.9563**	**0.0000**	**0.0156**	**23.2532**	**0.0000**	**20.7391**	**0.0000**	**1.9070**	**-1.5170**
ITEM 140	10.6448	0.0049	0.0116	10.8352	0.0044	11.0913	0.0009	1.8808	-1.4845

Findings of Evidence Regarding Content Validity

After observing the 3PLM fits the data best, item and person parameters were estimated with the three-parameter logistic model to obtain content validity evidence. In Appendix D, the person-item map (wright-map) in which the difficulty parameters of the VST items and ability parameters of the respondents are located on the same scale is given. On this map, it is observed that the VST has a sufficient number of items in every level of ability parameter, meaning that the VST, with its 140 items, is able to measure the vocabulary size of both low and high-proficiency individuals. Among 140 items, the easiest ones are items 6., 2., and 1. Moreover, for these items, the b parameters were estimated as -6.50, -6.23, and -5.45, respectively. The most difficult items are items 96, 58, and 68, and b parameters for these items were found as 3.14, 3.09, and 3.01, respectively. The locations of these items can be seen on the person-item map.

As it has been stated before, when the number of an item increases, the frequency of the word used in this item decreases, and therefore, in theory, the difficulty of the item increases, as well. On the person-item map (Appendix D), we can easily see that this theory is valid to some extent. The first questions are located on the left part of the scale, and when the sequence number of the items increases, they gradually move to the right side. However, there are some exceptions. Firstly, there are some questions which are more difficult than expected. These are items 4, 16, 58, and 68, and the vocabulary used in these items are *figure, nil, cavalier,* and *azalea*, respectively. These four items have difficulty parameters which are quite higher than the other items in their 10-word group. For instance, the mean of the difficulty parameters of the first 10 items is -3.80; however, the b parameter of the 4th item is 0.55. The departure of item 4 can be clearly seen from the person-item map. Secondly, there are approximately 20 questions which are easier than expected. These are the items of 35, 46, 47, 50, 54, 56, 61, 67, 70, 72,

74, 83, 88, 94, 103, 105, 107, 117, and 126, and the words used in these items are *quiz, cube, miniature, bacterium, accessory, thesis, olive, demography, yoghurt, palette, kindergarten, monologue, octopus, mystique, yoga, puma, aperitif, caffeine,* and *plankton*. Except for *olive* and *kindergarten*, all those easy words are loan words in Turkish, and thus these words were answered correctly quite more than the other items in their group. To illustrate, item 126 included *plankton*, and its parameter b was calculated as -1.30. However, other words in the same group have a mean of 1.70 for the same parameter. Likewise, the departures of these items from their groups can be seen on the person-item map (Appendix D).

Findings of Evidence Regarding Convergent Validity

While collecting data with the VST, we also asked the respondents the last score they obtained from an English proficiency test. Approximately 600 students responded to that question. The responses included scores on two English proficiency tests, namely TOEFL and the FLPE. We examined the relationship between their VST scores and the scores from those two English proficiency tests. One hundred and sixty of the students provided their TOEFL scores.

Figure 4

Scatter Plots of the VST Scores and Two Language Tests

The relation between the VST and TOEFL scores, and the VST and the FLPE scores can be seen in Figure 4, where scatter plots of the VST scores, the TOEFL scores, and the FLPE scores are illustrated. As may be seen from the first plot in Figure 4, there is a positive correlation between the VST and

TOEFL scores. The correlation coefficient for these variables was calculated as 0.60 (95% CI = 0.49, 0.69). Three hundred and sixty-eight of the respondents reported their FLPE results, and the positive correlation between the VST and the FLPE results can be seen in the second plot. The correlation coefficient was found as 0.53 (95% CI = 0.45, 0.60) for these two. Two high correlations (Cohen, 1992) indicate that VST scores relate closely to other measures of English proficiency, and this provides convergent evidence for the validity of the VST.

Discussion and Conclusion

In this study, to validate the VST, we collected validity evidence based on Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014). The sources of evidence for validity that we investigated included content, relation to other variables, internal structure, and response processes.

In order to examine content validity through a person-item map, we checked whether there were a sufficient number of items in the VST and whether they were distributed moderately on the ability scale. The results showed that the VST items possessed a wide variety of difficulty parameters which were located on almost every level of the ability scale, meaning that the VST distinguished high-proficiency students from the low-proficiency ones and had an appropriate number of questions for every level of θ. For internal structure validity, we carried out 3PLM-based IRT and DIF analyses. The internal structure of the VST was found to be unidimensional with the EFA. Moreover, the data and model fit of the VST scores were modelled via the IRT. In the IRT analyses, to determine the best-fitting model, the log-likelihood, its *p*-value, and other model fit indices were considered. The 3PLM IRT model was found to be the best-fitting model. DIF results revealed that there were 10 items which showed large DIF. In addition to the items which showed large gender-related DIF, some questions were identified as potentially problematic because they were easier than their difficulty level. These questions included loan words like yoghurt, microphone, or kindergarten. After appealing to expert opinions, the removal of these questions from the test might be considered to avoid inaccurate estimations of students' vocabulary size and item parameters. For relations to other variables' validity evidence (convergent validity), correlations between the VST scores with the TOEFL and the FLPE scores were examined. Convergent validity analysis revealed that there were high positive correlations between the VST scores and both of these exams. By using the 3PLM model, which also investigated the guessing effect, we indirectly gathered evidence for the response processes.

One of the fundamental properties that a measurement tool should have is validity. By collecting the validity evidence provided above to validate the VST, we contribute to the literature. In his study, Beglar (2010) administered the whole VST to high-proficiency students, but the middle- and low-proficiency groups took different versions of the VST which had fewer items. In our study, we gave the 140-item version of the VST to all participants regardless of their English proficiency levels. In our conditions, the VST was found to represent a valid measurement tool. When the VST is intended to be used in a computer adaptive test, in line with our findings, it is suggested that the 3PLM should be used for the CAT estimations and calculations.

The finding that the 3PLM fitted the VST data better than the one- and two-parameter models also indicates that the guessing effect does exist in answering the VST items, and some precautions suggested in the literature (Stewart, 2014; Zhang, 2013) like increasing the number of distractors, or adding an "I don't know" option, should be considered to decrease this effect.

Declaration

Author Contribution: Mustafa Gökcan: Conceptualization, methodology, investigation, formal analysis, data curation, writing - original draft. Derya Çobanoğlu Aktan: Conceptualization, methodology, investigation, data curation, validation, supervision, writing - review & editing.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: This study was approved by the Ethics Boards and Commissions of Hacettepe University (date: 16.06.2020, document number: 35853172-300-E.00001113493). This paper presents some of the results obtained during the Doctoral Thesis process under the supervision of Asst. Prof. Derya Çobanoğlu Aktan.

References

Ackerman, T. A. (1987, April). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence.* Paper presented at the annual meeting of the American Educational Research Association. Washington, DC.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. Continuum. https://doi.org/10.5040/9781474212151

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Baker, F. (2001). *The basics of Item response theory*. ERIC Clearinghouse on Assessment and Evaluation.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing, 27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Birnbaum A. (1968) Some Latent Trait Models, In Lord F.M., & Novick M.R. (eds.), *Statistical Theories of Mental Test Scores.* Addison-Wesley.

Camilli, G. (1994). Origin of the Scaling Constant d = 1.7 in Item Response Theory. *Journal of Educational and Behavioral Statistics, 19*(3), 293-295. https://doi.org/10.2307/1165298

Cattell, R. B. (1966). The Scree Test for The Number of Factors. *Multivariate Behavioral Research, 1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Cho, S.-J., Li, F., & Bandalos, D. (2009). Accuracy of the Parallel Analysis Procedure With Polychoric Correlations. *Educational and Psychological Measurement, 69*(5), 748–759. https://doi.org/10.1177/0013164409332229

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. https://doi.org/10.1037//0033-2909.112.1.155

Çepni, Z. & Kelecioğlu, H. (2021). Detecting Differential Item Functioning Using SIBTEST, MH, LR and IRT Methods. *Journal of Measurement and Evaluation in Education and Psychology, 12*(3), 267-285. https://doi.org/10.21031/epod.988879

Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge*. Cambridge University Press. https://doi.org/10.1017/CBO9780511667268

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the vocabulary size test. *Language Testing, 30*(2), 253–272. https://doi.org/10.1177/0265532212459028

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: principles and applications*. Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139–164. https://doi.org/10.1177/014662168500900204

Hertzman, M. (1936). The effects of the relative difficulty of mental tests on patterns of mental organization. *Archives of Psychology, 197*.

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Lawrence Erlbaum.

Horn, J. L. (1965). A Rationale and Test for The Number of Factors in Factor Analysis. *Psychometrika, 30*, 179–185. https://doi.org/10.1007/BF02289447

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement, 20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Karami, H. (2012). The development and validation of a bilingual version of the vocabulary size test. *RELC Journal, 43*(1), 53–67. https://doi.org/10.1177/0033688212439359

Kıbrıslıoğlu Uysal, N., & Atalay Kabasakal, K. (2017). The effect of background variables on gender related differential item functioning. *Journal of Measurement and Evaluation in Education and Psychology, 8*(4), 373-390. https://doi.org/10.21031/epod.333451

Koyuncu, İ., & Kılıç, A. (2019). The use of exploratory and confirmatory factor analyses: A document analysis. *Education and Science, 44*(198). http://dx.doi.org/10.15390/EB.2019.7665

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*, 936–949. https://doi.org/10.3758/s13428-015-0619-7

Li, C. H. (2019). Using a Listening Vocabulary Levels Test to explore the effect of vocabulary knowledge on GEPT listening comprehension performance. *Language Assessment Quarterly, 16*(3), 328–344. https://doi.org/10.1080/15434303.2019.1648474

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27*(1), 82–99. https://doi.org/10.1111/j.2044-8317.1974.tb00530.x

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555–576. https://doi.org/10.1177/0265532211430367

Meara, P. (1980). Vocabulary acquisition: A neglected aspect of language learning. *Language Teaching*, 13(3-4), 221-246. https://doi.org/10.1017/S0261444800008879

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Maters. https://doi.org/10.21832/9781847692092

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer. (Eds.), *L2 vocabulary acquisition, knowledge, and use: New perspectives on assessment and corpus analysis* (pp. 57-78). Eurosla Monographs Series. https://www.eurosla.org/monographs/EM02/Milton.pdf

Miralpeix, I. & Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching, 56*(1), 1-24. https://doi.org/10.1515/iral-2017-0016

Mizumoto, A., Sasao, Y., & Webb, S. A. (2019). Developing and evaluating a computerized adaptive testing version of the word part levels test. *Language Testing, 36*(1), 101–123. https://doi.org/10.1177/0265532217725776

Muthén, B., du Toit, S.H.C. & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished technical report. https://www.statmodel.com/download/Article_075.pdf

Nagelkerke, N. J. D. (1991). A note on the general definition of the coefficient of determination. *Biometrika, 78*(3), 691-692. https://doi.org/10.1093/biomet/78.3.691

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13. https://jalt-publications.org/tlt/issues/2007-07_31.7

Nation, I. S. P. (2013). *Learning vocabulary in another language (2nd ed.).* Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nguyen, L. T. C., & Nation, P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal, 42*(1), 86–99. https://doi.org/10.1177/0033688210390264

Noreillie, A. S., Kestemont, B., Heylen, K., Desmet, P., & Peters, E. (2018). Vocabulary knowledge and listening comprehension at an intermediate level in English and French as foreign languages. *International Journal of Applied Linguistics, 169*(1), 212-231. https://doi.org/10.1075/itl.00013.nor

Ockey, G. J., & Choi, I. (2015) Item Response Theory. *The Encyclopedia of Applied Linguistics*. 1-8. https://doi.org/10.1002/9781405198431.wbeal1476

Paek, I., & Cole, K. (2020). *Using R for item response theory model applications*. Routledge.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* University of Chicago Press.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25*(3), 193–203. https://doi.org/10.1111/j.1745-3984.1988.tb00302.x

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer. https://doi.org/10.1007/978-0-387-89976-3

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. https://doi.org/10.1017/S0261444819000326

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* MaCmillan.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal, 36*(2), 139–152. https://doi.org/10.1080/09571730802389975

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly, 11*(3), 271–282. https://doi.org/10.1080/15434303.2014.922977

Tran, U. S., & Formann, A. K. (2009). Performance of Parallel Analysis in Retrieving Unidimensionality in the Presence of Binary Data. *Educational and Psychological Measurement, 69*(1), 50–61. https://doi.org/10.1177/0013164408318761

Uysal, İ., Ertuna, L., Ertaş, F., G. & Kelecioğlu, H. (2019). Performances based on ability estimation of the methods of detecting differential item functioning: A simulation study. *Journal of Measurement and Evaluation in Education and Psychology, 10*(2), 133-148. https://doi.org/10.21031/epod.534312

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer.

Weng, L.-J., & Cheng, C.-P. (2005). Parallel Analysis with Unidimensional Binary Data. *Educational and Psychological Measurement, 65(*5), 697–716. https://doi.org/10.1177/0013164404273941

Yang, Y., & Xia, Y. (2015). On the number of factors to retain in exploratory factor analysis for ordered categorical data. *Behavior Research Methods, 47*(3), 756–772. https://doi.org/10.3758/s13428-014-0499-2

Yen, W.M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zhang, X. (2013). The "i don't know" option in the vocabulary size test. *TESOL Quarterly, 47*(4), 790–811. https://doi.org/10.1002/tesq.98

Zhang, S., & Zhang, X. (2020). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research,26*(4), 696–725. https://doi.org/10.1177/1362168820913998

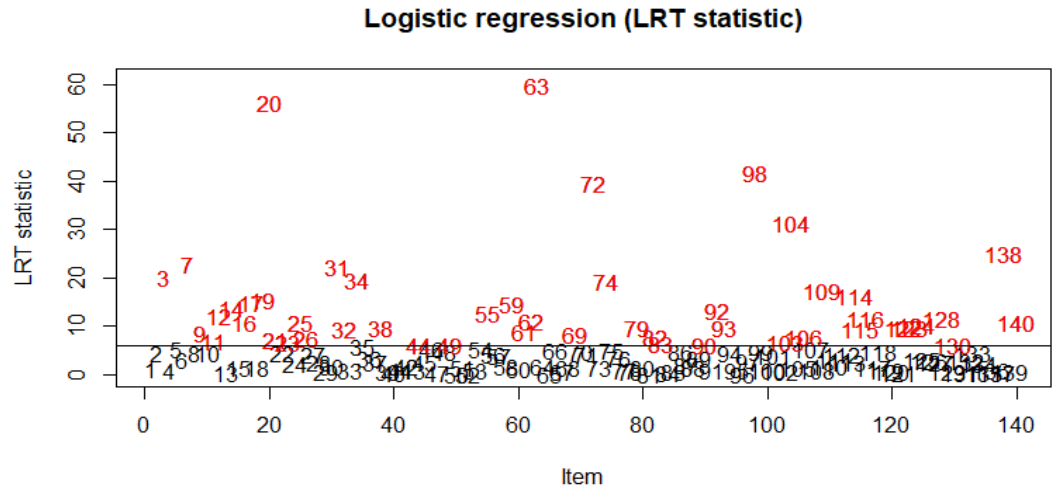
Zhao, P., & Ji, X. (2018). Validation of the Mandarin version of the vocabulary size test. *RELC Journal, 49*(3), 308–321. https://doi.org/10.1177/0033688216639761

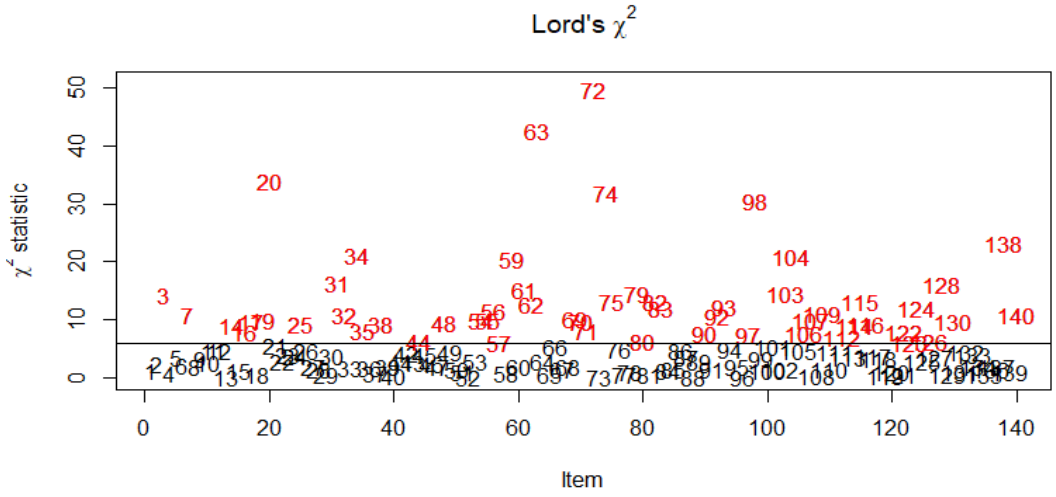
Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223–233. https://doi.org/10.1080/15434300701375832

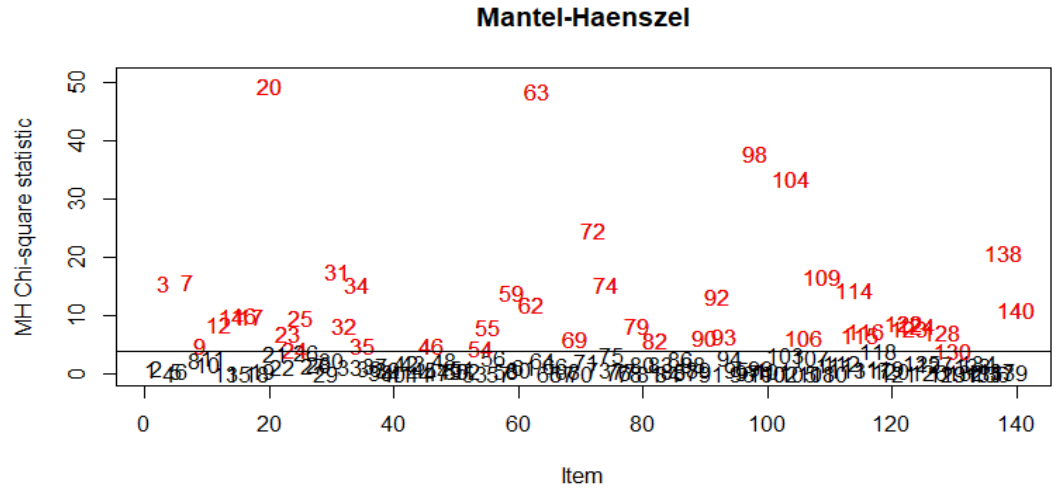
Appendices

Appendix A

Plots of DIF Results

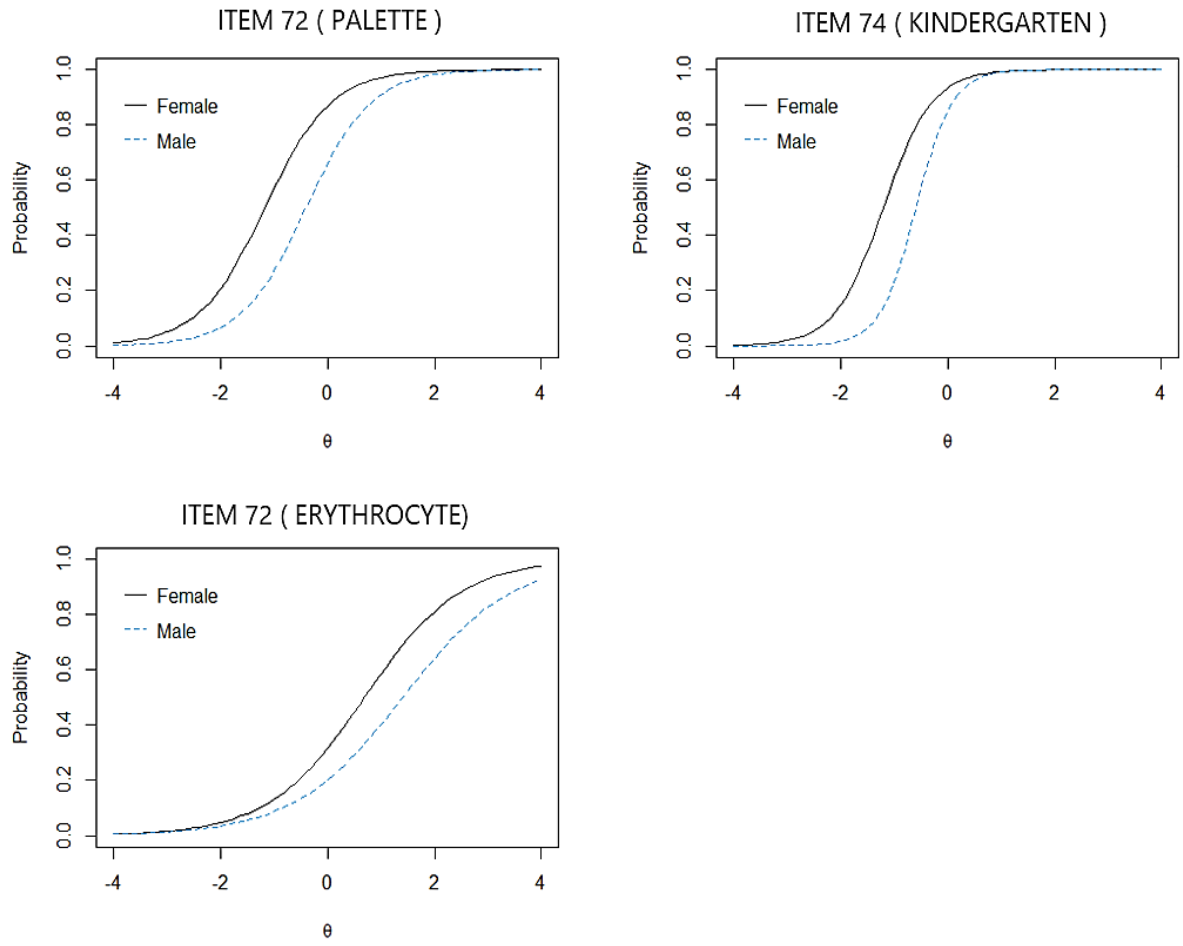






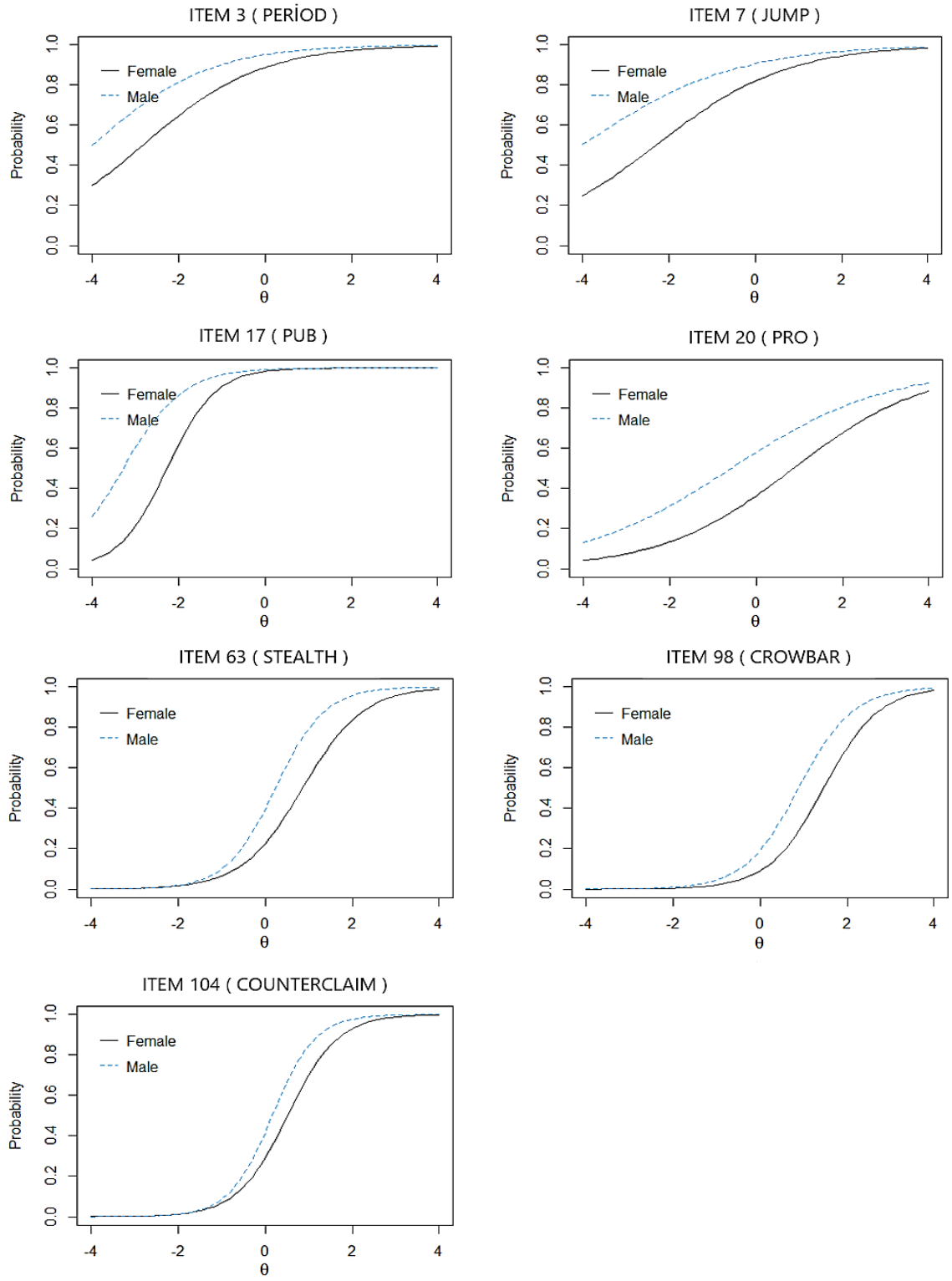
Appendix B

ICCs of the DIF Items Favoring Females



Appendix C

ICCs of the DIF Items Favoring Male



Appendix D

Person-Item Map (Wright Map)