VIETNAM JOURNAL OF EDUCATION, 2023
Volume 7, Issue 1, 63-73
https://doi.org/10.52296/vje.2023.270

ORIGINAL ARTICLE

Assessing the Quality of a Newly Designed Vocabulary Test for Vietnamese EFL Learners: A Rasch-based Analysis

Phuong Kim Thi Bui[1,+], Thanh Quy Nguyen[2], Hung Thai Le[2]

[1]Hanoi University of Science and Technology, Vietnam;
[2]University of Education - Vietnam National University, Vietnam
[+]*Corresponding author ● Email: phuong.buithikim@hust.edu.vn*

Article history
Received: 22 February, 2023
Accepted: 24 March, 2023
Published: 30 March, 2023

Keywords
Test development, English vocabulary test, EFL learners, Rasch model

ABSTRACT
Test development involves many required stages, including examining the psychometric characteristics of the test and each test item. This study describes the development of an English vocabulary test designed for EFL learners in Vietnam and provides an analysis using the Rasch model to evaluate the test quality and test items. The data was collected from 202 students from a Vietnamese university and analysed with the use of the Conquest computer program. The results show that the test had high reliability. Four out of 50 items were removed from the test as not fitting the Rasch model. 36 items were ranked very good, and 10 items require some revision efforts for better distractors. The paper is expected to contribute to the wide application of the Rasch model in the practices of test development and suggest the development of more reliable instruments for EFL learners to assess their vocabulary.

1. INTRODUCTION

Testing vocabulary knowledge is important in language education for both teachers and learners. For language learners, the test of vocabulary knowledge helps to determine their language proficiency because vocabulary serves as the foundation for all activities of language use (Schmitt et al., 2017). A good level of vocabulary knowledge is one of the important prerequisites for successful language learning. For teachers, having reliable estimates of vocabulary knowledge allows them to deliver materials tailored to learners' needs, assess the effectiveness of the learning process, and set appropriate goals so that learners can develop their language skills and competence (Nation, 2001).

For research purposes, vocabulary knowledge is a strong predictor of learners' language proficiency and even their academic achievement (Lin & Morison, 2010). In the opposite direction, learners' vocabulary tends to improve as their language level develops (Zareva et al., 2005), or students' practice in four language skills, such as reading, listening, speaking and writing, further supports the acquisition of newly learned words into students' memory (Laufer, 2013). Furthermore, vocabulary tests can be used to assess the impact of learning experiences on vocabulary development as well as to measure vocabulary development (Stoeckel & Bennett, 2016).

There exists a list of vocabulary tests: Vocabulary Levels Test (Nation, 1983; Schmitt et al., 2001), New Vocabulary Levels Test (McLean & Kramer, 2015), Updated Vocabulary Levels Test (Webb et al., 2017), Productive Vocabulary Levels Test (Laufer & Nation, 1999), Listening Vocabulary Levels Test (McLean et al., 2015), Computer adaptive test of size and strength (CATSS) (Laufer & Goldstein, 2004; Aviad-Levitzkyet al., 2019) and Vocabulary Size Test (Nation & Beglar, 2007) with numerous versions and bilingual adaptations. The practice of English teaching and learning, as well as educational research, is greatly affected by these vocabulary tests. However, no test can be used in all contexts, and the choice of a suitable test depends on the particular purposes of learners, teachers and researchers.

In the context of EFL teaching and learning in Vietnam, vocabulary has always been valued in the English curriculum in Vietnam. After completing the general education program, students are required to have a vocabulary of about 2500 words, which is stipulated in the General Education Curriculum of English issued together with Circular No. 32/2018/TT-BGDĐT dated December 26, 2018 of the Minister of Education and Training. However, according to the results of a small number of recent studies on the Vietnamese EFL learners' vocabulary, high school and university students have a very limited vocabulary and do not achieve the required vocabulary size (Vu & Peters, 2021).

Moreover, it is admitted that the field of vocabulary assessment is underdeveloped and lacking reliable instruments (Ha, 2021). Recent studies rely on available vocabulary tests that are VLT and VST, which cannot meet the growing demands of EFL teachers, learners and researchers in Vietnam. The development of reliable assessment instruments has, therefore, become crucial with more focus on CEFR - a familiar framework that affects English teaching and learning practices in recent years. Future vocabulary tests also need to facilitate a more effective learning process for students, help teachers classify the student' level to support their teaching plans and include validation evidence to guarantee the reliability as suggested by the researchers in the field of vocabulary assessment and language education in general (Schmitt et al., 2020). The study addresses this gap in the field by presenting the author's efforts to develop and validate a new vocabulary test that will help satisfy the ever-growing demands of EFL teachers and learners in Vietnam.

2. LITERATURE REVIEW

This section aims to discuss several related concepts, including receptive vocabulary assessment, the word lists of Oxford 3000 and Oxford 5000, test development, and the Rasch model, in order to provide a theoretical basis for the current study.

Receptive vocabulary assessment

Both the teaching and research fields have seen a rise in vocabulary tests in the past decades. Although vocabulary tests are developed in various ways with different approaches, it cannot be denied that the most important aspect of vocabulary assessment is the relationship between word form and meaning. This aspect forms the basis for conducting learning and understanding other aspects of words (Webb & Chang, 2012).

The two most popular receptive vocabulary tests are Vocabulary Size Test (Nation & Beglar, 2007) and Vocabulary Levels Test (Nation, 1983; Schmitt et al., 2001) while the two most recently built and developed vocabulary tests are CATSS - Computer adaptive test of size and strength (Laufer & Goldstein, 2004; Aviad-Levitzkyet al., 2019) with the application of computer technology and NGSLT - New General Service List Test (Stoeckel & Bennett, 2015) with the updated list of high-frequency words.

These vocabulary tests have practical values and are widely used in English language teaching, learning, and educational research in Vietnam (Bui, 2022). Currently, many studies have been conducted with the use of receptive vocabulary tests to determine Vietnamese EFL learners' vocabulary ability, including studies assessing Vietnamese college and high school students' vocabulary by Le & Nation (2011), Nguyen & Webb (2017), Dang (2020), Duy & Nguyen (2019) and Nguyen (2021).

However, other reliable vocabulary tests are also needed to meet the diverse needs of Vietnamese EFL teachers, learners, and researchers. Schmitt et al. (2020) give valuable recommendations for future vocabulary tests and studies, such as defining a clear test purpose, determining the proper question format, developing a bilingual version, applying computer technology in the test design and administration, and conducting strict procedures of test development and validation. All these suggestions are carefully considered throughout the test development of this study.

Oxford 3000 and 5000

These are the two lists of words classified based on CEFR levels, developed and reviewed by reputable experts in Applied Linguistics. These lists are core words that have been selected based on their frequency in the Oxford English Corpus and their relevancy to English learners (Todd, 2016).

Oxford 3000 and 5000 are selected as suitable word sources for vocabulary tests for EFL learners, specifically Vietnamese students, for the following two reasons. Firstly, they are developed "specifically for the needs of English language learners" (Burkett, 2015). The lists include words that learners may encounter in class and the lessons, then every English learner should acquire. After the test on Oxford 3000 and 5000, test takers consciously improve their vocabulary, then use word lists to identify words to learn, focusing on strengthening their vocabulary. Secondly, CEFR is also the most popular international language reference framework for language proficiency in Vietnam. The Minister of Education and Training promulgates Circular 01/2014/TT-BGDĐT on the foreign language competency framework with six levels compatible with levels from A1 to C2 in CEFR. The wordlists with CEFR-aligned words may provide EFL learners with helpful guides through the most relevant words at levels of A1-C1 to direct their learning activities for vocabulary expansion and language proficiency improvement.

Test development

Test development is a process that involves the investment of time and effort in many stages. Figure 1 shows the stages of test development introduced by Lane et al. (2015). The process includes ten stages, the third of them - item development is the focus of this paper.

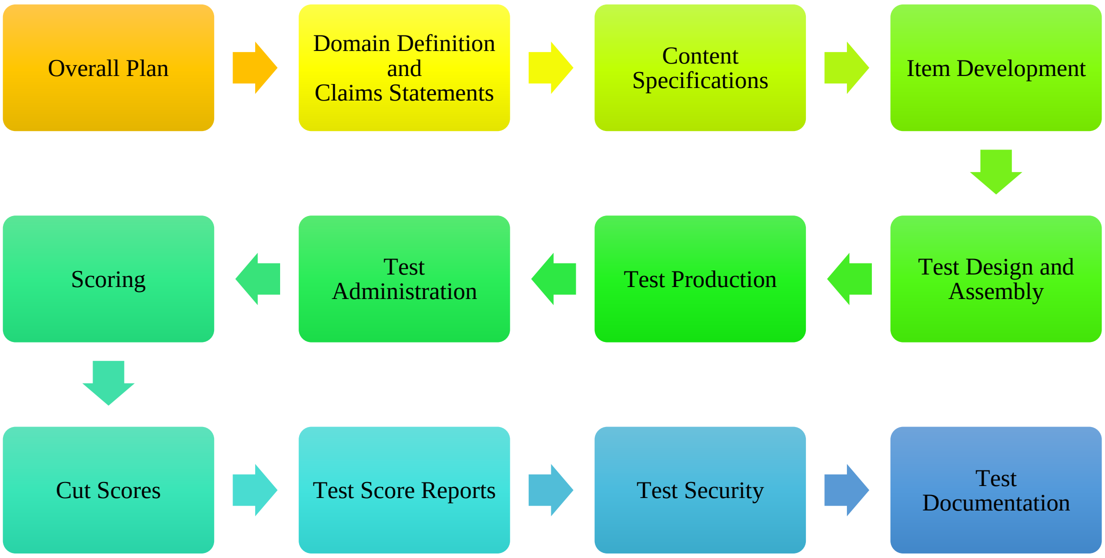


Figure 1. Test development process (Lane et al., 2015)

Item development involves determining the item format, item writing and review, item tryouts, and item banking. In this paper, the authors are going to review key concepts, introduce the test blueprints, including the test purpose, structure, matrix, word list, and sample item, then focus on a small-scale study to try out the items and assess test quality with a Rasch-based analysis for later item revision and other components of the test development process.

Rasch model

The Rasch measurement model, a one-parameter logistic model, is the simplest in the model family of item response theory (IRT) that is used to estimate the probability of a correct response given by a test taker to an item by examining the person’s trait estimate and the item difficulty, with the assumption that "A person having a greater ability than another person should have the greater probability of solving any item of the type in question and similarly, one item being more difficult than another means that for any person the probability of solving the second item correctly is the greater one" (Rasch, 1960, p. 117). The following is the equation for the Rasch measurement model.

$$P(u_i = 1 | \theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

In this equation, θ is the examinee’s trait estimate, P is the probability that a random examinee with ability θ answers item i correctly, u_i is the examinee’s response to the item, e is a transcendental number with the value of 2.718, b_i is the difficulty parameter of item i. The item difficulty parameters and trait estimates are on the same scale, generally ranging from -3 to +3.

The Rasch measurement model has many strengths, which could not be found in classical testing theory, for example, the assessment of the instrument quality targeting a specific group of respondents and the detection of challenges and opportunities for instrument improvement (Aghekyan, 2020). Rasch-based analysis has been performed in many studies of test development and validation. In Vietnam, many recent studies, such as Le et al. (2019), Le & Nguyen (2021), Nguyen et al. (2021) and Nguyen & Nguyen (2020) applied the Rasch measurement model for their test analysis. This study is another example of the application of the Rasch model in test development.

3. MATERIALS AND METHODS

3.1. Instrument

Before the step of test tryouts, all of the decisions for the new test were made, including the test purpose, the target test takers, the source word list, sampling methods for the test items, the item format, and guides for item writing. The test blueprints are presented in the following table. Following the test blueprints, the test designers composed a test with 50 items and had the tests reviewed by two experienced item writers in the field to minimize possible mistakes and improve the item quality of the test.

Table 1. Test blueprints

<table>
<tr><td><b>Purpose</b></td><td>Assessing EFL learner’s vocabulary through the relationship between word form and meaning in written form</td></tr>
<tr><td><b>Target test takers</b></td><td>EFL learners of all ages, beginner to intermediate English language proficiency levels</td></tr>
<tr><td><b>Word list to be assessed</b></td><td>Fifty discrete content words are sampled from Oxford 3000 and 5000, 10 items of each CEFR-based level from A1 to C1, with the ratio of word types in the test corresponding to the ratio of word types in the source word list.
<table>
<tr><th rowspan="2">CEFR</th><th colspan="5">Part of speech</th></tr>
<tr><th>Noun</th><th>Verb</th><th>Adjective</th><th>Adverb</th><th>Total</th></tr>
<tr><td>A1</td><td>5</td><td>3</td><td>1</td><td>1</td><td>10</td></tr>
<tr><td>A2</td><td>5</td><td>3</td><td>1</td><td>1</td><td>10</td></tr>
<tr><td>B1</td><td>5</td><td>3</td><td>2</td><td>0</td><td>10</td></tr>
<tr><td>B2</td><td>5</td><td>3</td><td>2</td><td>0</td><td>10</td></tr>
<tr><td>C1</td><td>5</td><td>3</td><td>2</td><td>0</td><td>10</td></tr>
<tr><td>Total</td><td>25</td><td>15</td><td>8</td><td>2</td><td>50</td></tr>
</table></td></tr>
<tr><td><b>Time allowance</b></td><td>Not limited</td></tr>
<tr><td><b>Item format</b></td><td>Multiple choice items
- An English word is presented in bold format, then a sentence using the word in a non-defining context is followed.
- There are four options, including one correct answer and three distractors.</td></tr>
<tr><td><b>Instructions</b></td><td>Instructions are written in both English and Vietnamese:
- English: Choose the answer that best describes the meaning of the word in “…” of the given sentence.
- Vietnamese: Bạn hãy chọn đáp án miêu tả gần nhất nghĩa của từ trong dấu “…” của câu cho trước.</td></tr>
<tr><td><b>Example test item</b></td><td>She “climbed” a tree.
a. broke</td></tr>
</table>

	b. went up c. drew d. cut down
Scoring	1 point per correct answer
Result report	Total score out of 50 and level score out of 10 for five levels.

3.2. Participants and test administration

The test trialing was carried out in a Hanoi university with 202 participants selected with convenience sampling. They all volunteered to join the study by doing the vocabulary test and could see the results after finishing it. This sample size of more than 150 examinees satisfies the requirements recommended by Şahin & Weiss (2015) to ensure accurate person ability and item parameter estimates. The participants included 104 English-majored students enrolled in the program of English language studies and 98 students of other majors.

Before the test administration, the test instructions were given to the participants, who could ask any related questions before starting the test. The test takers could spend as much time answering the questions as needed to complete the test. They were also encouraged to skip the item instead of guessing the answer if they faced unknown words to reduce the risk of overestimating their vocabulary knowledge. The results from the test trial were collected and analyzed to assess the characteristics of the test.

4. RESULTS AND DISCUSSION

4.1. Results

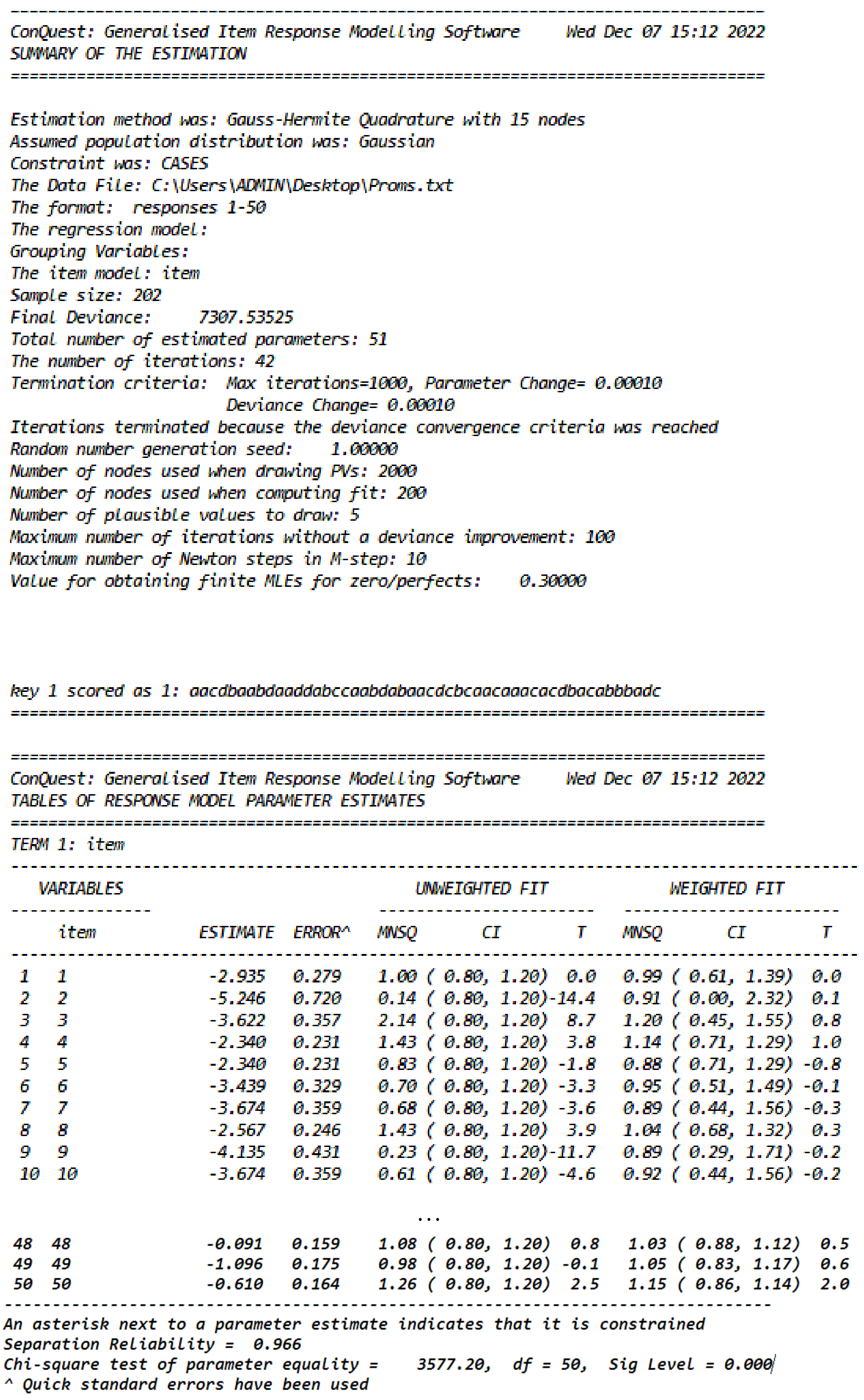
Test reliability

Conquest was adopted to perform an analysis of the designed test to examine the test features. The Separation Reliability was reported to be very high - 0.966, indicating item parameters were very well separated.

Table 2 shows the reliability coefficients for the test with the whole sample of 202 test takers. The coefficient alpha was presented at 0.89, showing the high reliability of the designed test. Moreover, the average score was high for both groups of English-majored and non-English-majored, the maximum score was 100, and the minimum was 36. The test could be evaluated to be easy for the participants of the test tryouts.

Table 2. Descriptive statistics of test scores

Group	n	Mean	Median	Min
All	202	80	82	36
English-majored	104	85	88	46
Non-English-majored	98	74	78	36



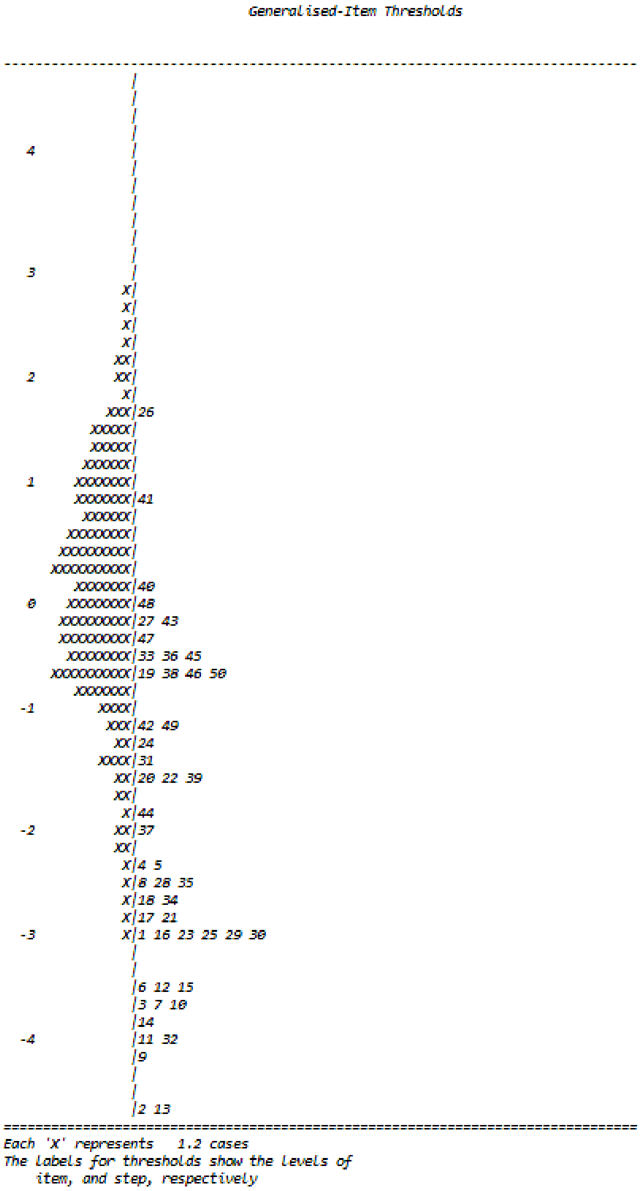
```
--------------------------------------------------------------------------------
ConQuest: Generalised Item Response Modelling Software     Wed Dec 07 15:12 2022
SUMMARY OF THE ESTIMATION
================================================================================

Estimation method was: Gauss-Hermite Quadrature with 15 nodes
Assumed population distribution was: Gaussian
Constraint was: CASES
The Data File: C:\Users\ADMIN\Desktop\Proms.txt
The format:  responses 1-50
The regression model:
Grouping Variables:
The item model: item
Sample size: 202
Final Deviance:     7307.53525
Total number of estimated parameters: 51
The number of iterations: 42
Termination criteria:  Max iterations=1000, Parameter Change= 0.00010
                       Deviance Change= 0.00010
Iterations terminated because the deviance convergence criteria was reached
Random number generation seed:    1.00000
Number of nodes used when drawing PVs: 2000
Number of nodes used when computing fit: 200
Number of plausible values to draw: 5
Maximum number of iterations without a deviance improvement: 100
Maximum number of Newton steps in M-step: 10
Value for obtaining finite MLEs for zero/perfects:    0.30000



key 1 scored as 1: aacdbaabdaaddabccaabdabaacdcbcaacaaacacdbacabbbadc
================================================================================

================================================================================
ConQuest: Generalised Item Response Modelling Software     Wed Dec 07 15:12 2022
TABLES OF RESPONSE MODEL PARAMETER ESTIMATES
================================================================================
TERM 1: item
------------------------------------------------------------------------------------------
   VARIABLES                              UNWEIGHTED FIT            WEIGHTED FIT
---------------                      -----------------------  -----------------------
     item            ESTIMATE  ERROR^  MNSQ       CI        T  MNSQ       CI        T
------------------------------------------------------------------------------------------
 1   1                -2.935   0.279   1.00 ( 0.80, 1.20)  0.0  0.99 ( 0.61, 1.39)  0.0
 2   2                -5.246   0.720   0.14 ( 0.80, 1.20)-14.4  0.91 ( 0.00, 2.32)  0.1
 3   3                -3.622   0.357   2.14 ( 0.80, 1.20)  8.7  1.20 ( 0.45, 1.55)  0.8
 4   4                -2.340   0.231   1.43 ( 0.80, 1.20)  3.8  1.14 ( 0.71, 1.29)  1.0
 5   5                -2.340   0.231   0.83 ( 0.80, 1.20) -1.8  0.88 ( 0.71, 1.29) -0.8
 6   6                -3.439   0.329   0.70 ( 0.80, 1.20) -3.3  0.95 ( 0.51, 1.49) -0.1
 7   7                -3.674   0.359   0.68 ( 0.80, 1.20) -3.6  0.89 ( 0.44, 1.56) -0.3
 8   8                -2.567   0.246   1.43 ( 0.80, 1.20)  3.9  1.04 ( 0.68, 1.32)  0.3
 9   9                -4.135   0.431   0.23 ( 0.80, 1.20)-11.7  0.89 ( 0.29, 1.71) -0.2
 10  10               -3.674   0.359   0.61 ( 0.80, 1.20) -4.6  0.92 ( 0.44, 1.56) -0.2

                                              ...

48  48               -0.091   0.159   1.08 ( 0.80, 1.20)  0.8   1.03 ( 0.88, 1.12)  0.5
49  49               -1.096   0.175   0.98 ( 0.80, 1.20) -0.1   1.05 ( 0.83, 1.17)  0.6
50  50               -0.610   0.164   1.26 ( 0.80, 1.20)  2.5   1.15 ( 0.86, 1.14)  2.0
--------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =  0.966
Chi-square test of parameter equality =    3577.20,  df = 50,  Sig Level = 0.000
^ Quick standard errors have been used
```

Figure 2. Item parameter estimates

Person ability estimates

Map of Latent Distributions and Thresholds was included in the result files of Conquest analysis (Figure 3). In this map, persons' ability estimates were presented on the left, and items' difficulty levels were presented on the right. From this map, the ability that could be measured was from -4 to 2 logits, which suggests that the test could measure a wide range of person abilities. The item difficulty needed more attention when the item of the highest difficulty level could measure the person's ability level of 1. The latent distribution map was consistent with the above-mentioned average score indicating that the ability estimates of the participants were high, or the participants' person ability was higher than the expected ability of the designed test. In other words, the test appeared to be more beneficial in measuring students of lower proficiency levels.

It is also noteworthy that eight items measured the ability of below -3; however, with the purpose of measuring the test takers' vocabulary size, these items are still valid to indicate the vocabulary obtained by the test takers and to test the learners of the lowest level - beginner.



```
                              Generalised-Item Thresholds


------------------------------------------------------------------------------
                |
                |
                |
                |
   4            |
                |
                |
                |
                |
                |
                |
   3            |
               X|
               X|
               X|
               X|
              XX|
   2          XX|
               X|
             XXX|26
           XXXXX|
           XXXXX|
          XXXXXX|
   1     XXXXXXX|
         XXXXXXX|41
          XXXXXX|
        XXXXXXXX|
       XXXXXXXXX|
      XXXXXXXXXX|
         XXXXXXX|40
   0    XXXXXXXX|48
       XXXXXXXXX|27 43
       XXXXXXXXX|47
        XXXXXXXX|33 36 45
      XXXXXXXXXX|19 38 46 50
         XXXXXXX|
  -1        XXXX|
             XXX|42 49
              XX|24
            XXXX|31
              XX|20 22 39
              XX|
               X|44
  -2          XX|37
              XX|
               X|4 5
               X|8 28 35
               X|18 34
               X|17 21
  -3           X|1 16 23 25 29 30
                |
                |
                |6 12 15
                |3 7 10
                |14
  -4            |11 32
                |9
                |
                |
                |2 13
==============================================================================
Each 'X' represents   1.2 cases
The labels for thresholds show the levels of
    item, and step, respectively
```

Figure 3. Map of Latent distributions and thresholds

Item assessment for revision

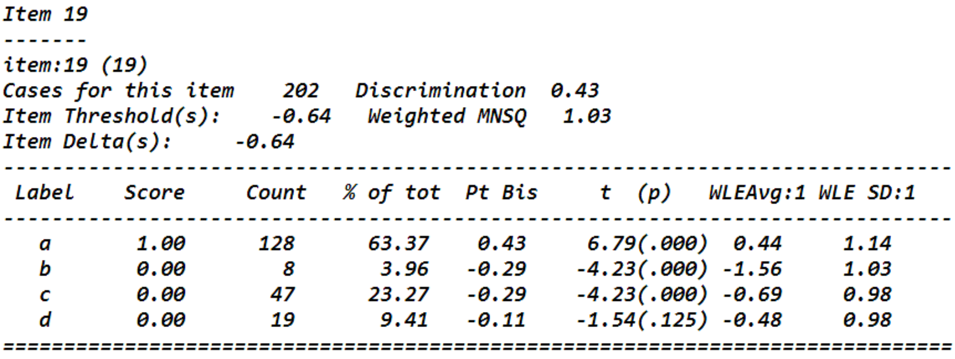
In this part for the stage of item development, the items are classified based on quality with a view to revising the test items.

Regarding the relevancy of test items, the Weighted Fit index was used to detect inappropriate items. From the Conquest analysis results, nine items (Table 3) were found to have MNSQ that are not in the CI range, and their corresponding T statistics had an absolute value that exceeded 2.0. It is concluded that the items did not conform to the model and would be removed from the test in the completed version. The remaining 46 items with the MNSQ statistics fit within the interval were confirmed to fit the model. These questions were then divided into two categories.

Table 3. Analysis result of the 4 misfit items

VARIABLES			UNWEIGHTED FIT			WEIGHTED FIT		
item	ESTIMATE	ERROR^	MNSQ	CI	T	MNSQ	CI	T
20 20	-1.597	0.192	1.63	(0.80, 1.20)	5.4	1.27	(0.79, 1.21)	2.3
24 24	-1.221	0.179	1.34	(0.80, 1.20)	3.1	1.21	(0.82, 1.18)	2.1
36 36	-0.450	0.162	1.21	(0.80, 1.20)	2.0	1.14	(0.87, 1.13)	2.0
50 50	-0.610	0.164	1.26	(0.80, 1.20)	2.5	1.15	(0.86, 1.14)	2.0

The first category of good or satisfactory questions fit the Rasch model and had acceptable difficulty, acceptable discrimination, and choices with similar selection rates. One good item of the test – item 19 is the following (Figure 4).



```
Item 19
-------
item:19 (19)
Cases for this item     202   Discrimination  0.43
Item Threshold(s):    -0.64   Weighted MNSQ    1.03
Item Delta(s):     -0.64
------------------------------------------------------------------------------
 Label     Score      Count   % of tot  Pt Bis      t  (p)   WLEAvg:1 WLE SD:1
------------------------------------------------------------------------------
   a        1.00       128     63.37     0.43     6.79(.000)  0.44     1.14
   b        0.00         8      3.96    -0.29    -4.23(.000) -1.56     1.03
   c        0.00        47     23.27    -0.29    -4.23(.000) -0.69     0.98
   d        0.00        19      9.41    -0.11    -1.54(.125) -0.48     0.98
==============================================================================
```

Figure 4. Analysis result of Item 19

With regards to the Classical Testing Theory, the difficulty of the question was calculated as 0.63, within the range of 0.25-0.75; 63.3% of the test takers answered this question correctly. Discrimination was very good, d=0.43, indicating that the question could distinguish between a highly qualified candidate group and a group of candidates with low ability levels. The correlation coefficient (PT BIS) showed us the distractors with negative indexes and the right option with a positive one. The given choices of Item 19 were valuable in assessing the competency of the candidates.

This result could also be confirmed thanks to the matching plot of the modeled and empirical category characteristic curves of Item 19 (Figure 5), which proved that the item was well designed to assess the test takers' ability.

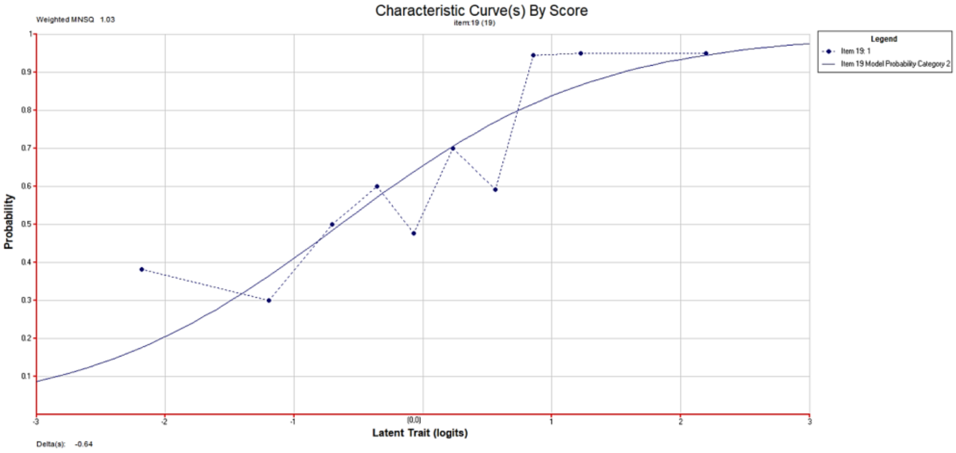
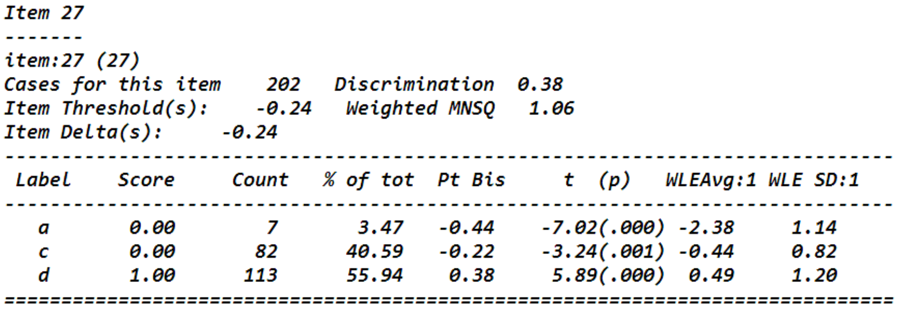


Figure 5. Characteristic curves of Item 19

The second category of unsatisfactory questions should be removed or adjusted to fit the model and the evaluation criteria. Ten items, including Items 1, 2, 3, 9, 11, 13, 14, 26, 27, and 32, were reported to require more investment of time and effort for revision. For example, Item 27 had a good discrimination index of 0.38; and an acceptable difficulty level of 0.56, but none of the test takers chose "b" as their answer. This item, therefore, needs some revisions to edit its distractors.



```
Item 27
-------
item:27 (27)
Cases for this item     202   Discrimination  0.38
Item Threshold(s):    -0.24    Weighted MNSQ   1.06
Item Delta(s):      -0.24
------------------------------------------------------------------------------
 Label     Score      Count   % of tot  Pt Bis      t   (p)   WLEAvg:1 WLE SD:1
------------------------------------------------------------------------------
   a        0.00        7       3.47   -0.44    -7.02(.000) -2.38      1.14
   c        0.00       82      40.59   -0.22    -3.24(.001) -0.44      0.82
   d        1.00      113      55.94    0.38     5.89(.000)  0.49      1.20
==============================================================================
```

Figure 6. Analysis result of Item 27

4.2. Discussion

After reviewing the analysis results for 50 items, it was concluded that four items should be removed from the test. Ten items required revision in order to improve their difficulty or edit their distractors. The process of developing the vocabulary test may continue with the writing of additional items to replace the disqualified items and build an item bank, then to produce parallel tests to support regular assessments and EFL learning and teaching in general.

As "validation is seen as an ongoing process, and so tests can never be 'validated' in a complete and final manner" (Schmitt et al., 2020), this study only provides initial validity evidence, thanks to the Rasch-based analysis, for the new test designed to assess the Vietnamese EFL learner's receptive vocabulary; further studies are always essential before a more confident use of the test in the future. From another perspective, this study, along with previous studies of test development and validation (Le et al., 2019; Le & Nguyen, 2021; Nguyen et al., 2021; Nguyen & Nguyen, 2020), contributes to a positive progress in the field of vocabulary assessment in Vietnam, emphasizing the importance of test validation to guarantee the validity and reliability of vocabulary tests as well as encouraging the use of Rasch measurement model as a good choice to accumulate validity evidence for future tests.

5. CONCLUSION

The study has accomplished the objective of evaluating the quality of the English receptive vocabulary test sampled from Oxford 3000 and 5000 word lists for EFL learners in Vietnam. The findings from Rasch-based analysis suggest that the designed test had a high level of reliability in assessing the test takers' receptive vocabulary knowledge. This study contributes to confirming that the Rasch measurement model is a useful tool in the ongoing test development process and should be used to guarantee the validity and reliability of newly designed instruments in the field.

These positive results may pave the way for the design of valuable and reliable vocabulary testing tools to serve different purposes and needs of learners, teachers and researchers in the context of English language education in Vietnam.

On the other hand, the study has some limitations. Firstly, the number of students participating in the study is still limited. Future studies may expand the sample size and investigate with more students of different language levels and learning contexts. Additionally, the study leaves room for future validation studies using two or three-parameter analytic models, instead of just using the Rasch model, to yield more useful and multidimensional results. Finally, future work could be directed towards building a question bank or applying adaptive assessments to serve regular testing and assessment, supporting EFL teaching, learning and research practices in Vietnam.

Conflict of Interest: No potential conflict of interest relevant to this article was reported.

REFERENCES

Aghekyan, R. (2020). Validation of the SIEVEA instrument using the Rasch analysis. *International Journal of Educational Research*, *103*, 101619. https://doi.org/10.1016/j.ijer.2020.101619

Aviad-Levitzky, T., Laufer, B., & Goldstein, Z. (2019). The new computer adaptive test of size and strength (CATS): Development and validation. *Language Assessment Quarterly*, *16*(3), 345-368.

Bui, T. K. P. (2022). A Review on Vocabulary Tests of High Frequency English Words. *VNU Journal of Science: Education Research*, *38*(3), 51-60.

Burkett, T. (2015). An investigation into the use of frequency vocabulary lists in university intensive English programs. *International Journal of Bilingual & Multilingual Teachers of English*, *3*(2), 71-83. https://doi.org/10.12785/IJBMTE/030202

Dang, T. N. Y. (2020). Vietnamese non-English major EFL university students' receptive knowledge of the most frequent English words. *VNU Journal of Foreign Studies*, *36*(3).

Duy, V. V., & Nguyen, C. N. (2019). An assessment of vocabulary knowledge of Vietnamese EFL learners. *The 20th English in Southeast Asia Conference*, Date: 2019/12/06-2019/12/07, Location: National Institute of Education, Nanyang Technological University, Singapore,

Ha, H. T. (2021). A Rasch-based validation of the Vietnamese version of the listening vocabulary levels test. *Language Testing in Asia*, *11*(1), 16. https://doi.org/10.1186/s40468-021-00132-7

Lane, S., Raymond, M. R., Haladyna, T. M., & Downing, S. M. (2015). Test development process. In *Handbook of test development* (pp. 19-34). Routledge.

Laufer, B. (2013). Lexical thresholds for reading comprehension: What they are and how they can be used for teaching purposes. *Tesol Quarterly*, *47*(4), 867-872. https://doi.org/10.1002/tesq.140

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399-436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33-51. https://doi.org/10.1177/026553229901600103

Le, T. C. N., & 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

Le, T. H. & Nguyen T. H. (2021). Experimental Research and Application of Computerized Adaptive Tests to assess Learners' Competencies. In *2021 3rd International Conference on Computer Science and Technologies in Education (CSTE)* (pp. 69-74). IEEE.

Le, T. H., Tang, T. T, Tran L. A., Nguyen T. D., Nguyen P. A & Nguyen T. Q. G. (2019). Developing Computerized Adaptive Testing: An Experimental Research on Assessing the Mathematical Ability of 10th Graders. *VNU Journal of Science: Education Research*, *35*(4).

Lin, L. H., & Morrison, B. (2010). The impact of the medium of instruction in Hong Kong secondary schools on tertiary students' vocabulary. *Journal of English for Academic Purposes*, *9*(4), 255-266. https://doi.org/10.1016/j.jeap.2010.09.002

McLean, S., & Kramer, B. (2015). The creation of a New Vocabulary Levels Test. *Shiken*, *19*(2), 1-11.

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*(6), 741-760. https://doi.org/10.1177/1362168814567889

Nation, I. S. P. (2001). How many high frequency words are there in English. *Language, Learning and Literature: Studies Presented to Hakan Ringbom*, *4*, 167-181.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9-13. http://www.jalt-publications.org/archive/tlt/2007/07_2007TLT.pdf

Nation, P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*, 12-25.

Nguyen, C. D. (2021). Lexical Features of Reading Passages in English-language Textbooks for Vietnamese High-school Students: Do they Foster both Content and Vocabulary Gain?. *RELC Journal*, *52*(3), 509-522. https://doi.org/10.1177/0033688219895045

Nguyen, T. H., Vu , T. L., Le , T. H., & Pham , V. H. (2021). Designing adaptive multiple-choice questions to assess the mathematical ability of 12th grade students. *Journal of Education*, *508*(2), 33-40.

Nguyen, T. M. H., & Webb, S. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research*, *21*(3), 298-320. https://doi.org/10.1177/1362168816639619

Nguyen, V. C., & Nguyen P. H. (2020). Analysing and selecting multiple choice test items based on classical test theory and item response theory. *Ho Chi Minh City University of Education Journal of Science*, *17*(10), 1804-1818.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

Şahin, A., & Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. *Educational Sciences: Theory & Practice*, *15*(6), 1585-1595. https://doi.org/10.12738/estp.2015.6.0102

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching*, *50*(2), 212-226. https://doi.org/10.1017/S0261444815000075

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

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55-88. https://doi.org/10.1177/026553220101800103

Stoeckel, T., & Bennett, P. (2015). A test of the new General Service List. *Vocabulary Learning Instruction*, *4*(1), 1-8.

Todd, V. (2016). Learn The Oxford 3000™. *English Australia Journal*, *31*(2), 95-97.

Vu, D. V., & Peters, E. (2021). Vocabulary in English language learning, teaching, and testing in Vietnam: A review. *Education Sciences*, *11*(9), 563. https://doi.org/10.3390/educsci11090563

Webb, S. A., & Chang, A. C. S. (2012). Second language vocabulary growth. *RELC Journal*, *43*(1), 113-126. https://doi.org/10.1177/0033688212439367

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL-International Journal of Applied Linguistics*, *168*(1), 33-69. https://doi.org/10.1075/itl.168.1.02web

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition*, *27*(4), 567-595. https://doi.org/10.1017/S0272263105050254