HNUE JOURNAL OF SCIENCE
Educational Sciences 2025, Volume 70, Issue 4, pp. 13-21
This paper is available online at https://hnuejs.edu.vn
DOI: 10.18173/2354-1075.2025-0064

ITEM RESPONSE THEORY ANALYSIS OF THE READING SECTION IN A MOCK VSTEP LEVEL 3-5 TEST

Dao Thi Bich Nguyen[*1], Nguyen Thi Hoai[1], Bui Tri Vu Nam[2], Nguyen Vinh Quang[3] and Hoang Nhat Linh[1]

[1]*Faculty of English, Hanoi National University of Education, Hanoi city, Vietnam*
[2]*Department of English, Ho Chi Minh University of Education, Ho Chi Minh city, Vietnam*
[3]*Quality Assurance Center, Hanoi National University of Education, Hanoi city, Vietnam*

[*]Corresponding author: Dao Thi Bich Nguyen, e-mail: nguyendb@hnue.edu.vn

Received February 6, 2025. Revised March 21, 2025. Accepted April 16, 2025.

Abstract. This study investigates the use of Item Response Theory (IRT) to evaluate and validate the items in the reading section of a mock Vietnamese Standardized Test of English Proficiency (VSTEP) level 3-5 according to the Vietnamese version of the Common European Framework of Reference (CEFR). The primary focus is to evaluate how well the test items align with the abilities of test-takers, comparing the question difficulty as designed (R-level) with the estimated difficulty generated through IRT analysis. Data were collected from responses of 100 test-takers to the Reading section of the mock VSTEP Level 3-5. By applying IRT, the study offers a detailed analysis of item characteristics, including item difficulty and discrimination parameters, which are crucial for validating whether the test accurately reflects the proficiency levels it is intended to measure. The results provide insights into the validity and reliability of the test items, highlighting potential discrepancies between the original design of question difficulty and the actual performance data. This analysis contributes to the overall improvement of the test, ensuring that it better serves as an accurate tool for assessing language proficiency at the specified levels. The findings also offer valuable implications for future test development and refinement in similar contexts.

Keywords: Item Response Theory, VSTEP, Reading, reading proficiency.

1. Introduction

Language testing is a critical tool for assessing language proficiency, particularly within educational and professional settings, where the outcomes of such assessments often carry significant consequences. Over the past few decades, research in the field of language testing has increasingly emphasized the importance of rigorous validation processes to ensure that tests accurately measure their intended constructs and that their results are applied ethically and appropriately (Kane, 2013; Brunfaut, 2023) [1], [2]. Validation entails a systematic evaluation of the inferences drawn from test scores, focusing on their alignment with the test's intended purposes and mitigating potential negative consequences arising from improper use (Kane, 2013) [1].

Item Response Theory (IRT) has emerged as an essential analytical framework for enhancing the reliability and validity of language assessments. IRT evaluates each test item individually through parameters like difficulty, discrimination, and guessing. These parameters remain

consistent across diverse test-taker populations, ensuring more stable and valid assessments. By modeling the probability of a correct response based on both item characteristics and test-taker ability, IRT provides a deeper and more nuanced understanding of test quality and performance (An & Yung, 2014; Min & Aryadoust, 2021) [3], [4].

This study aims to investigate the use of IRT to assess the quality of items in the Reading section of a VSTEP mock test, using data collected from a mock examination conducted in October 2023 at the Faculty of English, Hanoi National University of Education. From the large pool of participants, a sample of 100 test papers was selected for detailed analysis. Specifically, the study seeks to answer the question: *How effectively do the items in the reading section measure students' reading competence?*

2. Content

2.1. Literature review

2.1.1. Item Response Theory in Language Assessment

Item Response Theory (IRT), also referred to as Latent Trait Theory (LTT), the Item Characteristic Curve (ICC), or the Item Characteristic Function (ICF), gained prominence among measurement experts in the 1970s (e.g., Crocker & Algia, 1986; Henard, 1998) [5], [6]. Despite its rapid advancement and popularity among psychometric researchers, the True Score theory revealed significant limitations (Lee et al., 2012) [7], including the lack of invariance in item parameters across groups and the inability of classical testing methods to identify item bias. These issues led to renewed interest in IRT.

**** The IRT two-parameter logistic model (IRT 2PL)***

The two-parameter logistic model (2PL) within IRT is designed to analyze data by emphasizing two key aspects: the difficulty level of questions (Moore & Gordon, 2015) [8] and their discriminatory power. The 2PL model allows for more precise measurement by taking both item difficulty and item discrimination into account. Unlike Classical Test Theory (CTT), IRT ensures that the same item characteristics hold across different populations, which makes it particularly useful for comparing test-taker abilities across multiple proficiency levels.

2.1.2. Previous operationalizations of IRT Model in language testing

The application of IRT in language testing, especially in English as a Foreign Language (EFL) context, has significantly advanced over the past decades. This evolution is marked by the development of IRT models, particularly the one-parameter (Rasch), two-parameter (2PL), and three-parameter (3PL) models, that provide precise measurement at the item level, which is critical for analyzing individual test items and their functionality across diverse populations (Hambleton & Slater, 1997) [9].

Previous studies have utilized IRT in exploring diverse areas within EFL research. Specifically, Yamakawa et al. (2008) [10] focus on grammar acquisition, using IRT to equate data from different grammaticality judgment tests and explore the interrelationships among grammatical features. In terms of vocabulary research using IRT, Lee et al. (2012) [7] concentrate on vocabulary size, comparing learners' vocabulary levels across different test forms. Another study (Nix & Tseng 2014) [11] investigates EFL listening beliefs, employing multidimensional item response theory (MIRT) to create a measurement model and identify the underlying belief structure.

From these studies, a key strength of IRT is its ability to provide a deeper understanding of EFL constructs. For example, Yamakawa et al. (2008) [10] demonstrate how IRT can illuminate the acquisition process of different grammatical features, while Lee et al. (2012) [7] and Nix & Tseng (2014) [11] show its effectiveness in revealing the two-dimensional structure of EFL

listening beliefs and in comparing vocabulary levels more accurately. Additionally, Lee et al. (2012) [7] and He & Min (2016) [12] discuss the use of IRT in a computer adaptive test, highlighting its contribution to effective and efficient measurement.

However, previous studies also acknowledge limitations associated with IRT applications. Yamakawa et al. (2008) [10] stress the importance of having a large enough sample size and a sufficient number of items for each linguistic category to ensure accurate IRT analysis. This echoes the concerns raised in McCarron & Kuperman's (2021) study [13], where the limited sample size, particularly for ESL students, was identified as a major limitation. Nix & Tseng (2014) [11] relate to the potential cultural bias inherent in some assessment instruments. They argue that the Author Recognition Test, while commonly used, might not be suitable for all second-language learners due to its reliance on Western literary traditions. Therefore, the need for future research is highlighted, suggesting exploring the influence of the native language cue dependency on grammaticality judgment tasks and calling for investigating gender effects and testing the model on diverse EFL learner populations.

2.1.3. The VSTEP test

**** Overview***

The Vietnamese Standardized Test of English Proficiency (VSTEP) is the first standardized English proficiency test in Vietnam, organized by the Ministry of Education and Training since 2015 (Hung, 2023) [14]. It is an English proficiency test aligned with the 6-level Foreign Language Proficiency Framework for Vietnam. The test evaluates proficiency in all four language skills: Listening, Speaking, Reading, and Writing, and is widely used to measure English ability across different levels. [15]

**** VSTEP Reading Section***

The Reading Section of the VSTEP consists of four reading passages, with multiple-choice questions (MCQs) designed to assess the ability to read and comprehend various types of texts. The section contains four reading passages, each ranging from 400 to 550 words, totaling between 1,700 and 2,050 words across the entire section. The difficulty of the reading passages corresponds to levels 4 and 5 of the proficiency framework, while the questions range from levels 3 to 5, offering a well-rounded test of reading comprehension skills. [15]

2.2. Methodology

2.2.1. Research design

This research utilizes IRT as the primary analytical framework to investigate the reading section of a mock VSTEP Level 3-5 test. The decision to use IRT over CTT was based on the more sophisticated insights IRT provides regarding item-level analysis.

**** Participants***

This study involved 100 undergraduate students from the Hanoi National University of Education (HNUE), with ages ranging from 18 to 22. All participants were in their second or third year of university, having completed at least two years of formal English language instruction. This educational background ensures that participants' proficiency levels span across the target levels (VSTEP Levels 3, 4, and 5, equivalent to B1, B2, and C1 on the CEFR scale).

**** Data collection procedure***

The test was conducted in a controlled environment at HNUE, ensuring that all participants were subject to the same procedure. Each participant was given 60 minutes to complete the reading section. To maintain consistency and fairness, a set of standardized instructions was provided to all participants. Participants were first informed that the test was part of a research study designed to analyze the reading section of the mock VSTEP by the use of IRT, and they were reassured that their individual performance on the test would not impact their academic

records. It was then emphasized that once the test started, no further clarifications about the content would be provided. In addition, participants were instructed that no external materials, such as dictionaries, phones, or notes, were allowed during the test.

2.2.2. Data analysis and IRT model application

**** Item calibration and parameter estimation***

To prepare the data for analysis using IRT models, student responses were entered into a spreadsheet. Once data entry was complete, the file was saved in .csv format for import into the R statistical program for further analysis.

All test items were calibrated using the 2PL model within the Conquest software (Wu et al., 1988) [16]. This model estimates two parameters:

- Item Difficulty (b): This measures how challenging each item is for the test-takers, with higher values indicating more difficult items. The difficulty levels calculated using IRT were compared to the designed difficulty levels (Rlevel) to identify discrepancies.
- Item Discrimination (a): This parameter assesses how well each item differentiates between test-takers of varying proficiency levels. Items with higher discrimination values are more effective in distinguishing between high and low-proficiency test-takers.

MMLE (Marginal Maximum Likelihood Estimation), a statistical method used in data analysis, especially popular in IRT, was also employed. It is a way to estimate the parameters of the question (such as difficulty and discrimination) and the ability of the test taker based on the test results. Bock & Aitkin (1981) [17] argued that the MMLE is a reliable technique that incorporates both test takers' responses and their unknown ability levels, treating ability levels as latent variables. This method allows for precise parameter estimates, especially in large data sets, making it suitable for this analysis [17]. Given a test with many questions of different difficulty levels and many test takers with various abilities, MMLE helps to find out the difficulty level of each question and the ability of each test taker.

To begin the analysis, a Wright Map or test taker map visually displayed the relationship between question difficulty and test taker ability on the same scale. This map provides an effective way to assess the overall fit of questions to test taker ability. By plotting the distribution of question difficulty and test taker ability, it is easy to identify any large mismatches or gaps in question difficulty relative to the test taker's ability range, such as visualizing where questions fall within the test taker's ability range, which questions are too difficult or too easy. Questions that appear to be significantly mismatched will be flagged at this stage for further investigation.

Following this broad analysis with the Wright Map, ICCs were employed to perform a more detailed examination of individual items. The ICCs offered insights into both item difficulty and discrimination, showing the probability of a correct response across different ability levels. Specifically, the steepness of each curve provided a measure of discrimination, revealing how well each item distinguished between high- and low-ability test-takers. This step confirmed whether the items flagged on the Wright Map indeed had alignment or discrimination issues, allowing for a nuanced understanding of each item's effectiveness.

**** Interpretation of results and item revision***

Once the IRT analysis was complete, results were compared to the designed difficulty levels for each item. Items demonstrating significant discrepancies between the R-level and the IRT-estimated difficulty levels, as identified on the Wright Map and confirmed with the ICCs, were flagged for potential revision.

A Discrepancy Summary Table was created to capture the differences between the designed difficulty levels (R-level) of test items and their difficulty levels as estimated through the IRT model. This table provided a clear, organized overview of any mismatches between the intended

and actual performance of each item, offering insight into how well the items aligned with the target proficiency levels.

In addition to the Discrepancy Summary Table, the study also utilized an Item Analysis Report to examine the performance of each test item in greater detail. This report provided valuable diagnostic statistics for each item, including discrimination, item threshold, and weighted mean square error, which offered insights into the quality and effectiveness of the items. For each response option, the table displayed point-biserial correlations, counts, and response percentages, allowing for an in-depth evaluation of item functionality and response patterns. By examining these metrics, researchers could identify items with poor discrimination, misfitting responses, or options that were not functioning as intended. This item analysis report thus served as a complementary tool alongside the Wright Map and ICCs, offering a granular view of item characteristics that informed potential revisions to enhance the validity and reliability of the assessment instrument.

2.3. Findings

2.3.1. Difficulty level of the Mock VSTEP Level 3-5 Test's Reading Section

Results show that the exam effectively assesses and classifies candidates across reading proficiency levels, ranging from low to very high. However, certain anomalies exist. First, some items were too easy for all candidates, such as items 1, 6, 11, and 17, where all test-takers answered correctly. Conversely, certain items, like item 8, were too difficult, with no candidates able to answer them correctly. These findings suggest the need for a review of the content, answer options, and distractors in these specific items to enhance their alignment with the test's intended difficulty range.

A comparison between the intended difficulty levels (as per the test matrix) and the actual difficulty levels (estimated using Item Response Theory, IRT) further highlights discrepancies that need attention as follows:

Table 1. Items' intended difficulty levels and actual difficulty levels

Item	Intended difficulty levels	Actual difficulty levels
1	3	1
2	3	3
3	4	4
5	3	5
7	3	3
9	4	3
10	4	3
14	3	3
17	3	1
20	3	5
23	3	4

Findings reveal that in the test, 14/40 items exactly match their intended difficulty levels (e.g., items 2, 3, 7, 14, as shown in Table 1), and 11/40 items nearly match their intended difficulty levels with the discrepancy of one level only (e.g., 9, 10, 23, as shown in Table 1). This reflects the effectiveness of the test design.

However, 15/40 items do not align with their intended difficulty levels. For example, items 1 and 17 have the intended difficulty level 3 while their actual difficulty level is 1, meaning that

they are too easy compared to their intended difficulty levels. Some items such as items 5 and 20 (with intended level 3 and actual level 5) are, on the other hand, more difficult than their intended levels. These items need revising to ensure the compatibility between the intended level in the matrix and the actual level.

2.3.2. Item discrimination

Of the 40 test items, 28 items, accounting for 70%, have the discrimination index over 0.30, such as items 7 and 10 illustrated in Figure 1 below. Items with such a high discrimination index can effectively distinguish between test-takers' reading proficiency levels.

Item 7

Cases for this item 97 Discrimination 0.34
Item Threshold(s): -0.05 Weighted MNSQ 1.03
Item Delta(s): -0.06

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
A	1.00	67	69.07	0.34	3.58(.001)	1.09	0.82
B	0.00	8	8.25	-0.25	-2.55(.012)	0.09	0.78
C	0.00	18	18.56	-0.12	-1.13(.260)	0.57	0.53
D	0.00	4	4.12	-0.23	-2.26(.026)	-0.05	1.02

Item 10

Cases for this item 97 Discrimination 0.48
Item Threshold(s): -0.60 Weighted MNSQ 0.93
Item Delta(s): -0.60

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
A	1.00	76	78.35	0.48	5.41(.000)	1.06	0.78
B	0.00	6	6.19	-0.27	-2.68(.009)	0.17	0.91
C	0.00	14	14.43	-0.38	-4.06(.000)	0.10	0.64
D	0.00	1	1.03	-0.01	-0.07(.943)	0.62	0.00

Figure 1. Item discrimination index of items 7 and 10

Nevertheless, there are 2 items with near 0 and 1 item with negative discrimination index (items 1, 35, and item 8, respectively) (see Figure 2). These items with poor discrimination index generally fail to distinguish between examinees' proficiency and need thorough revision to improve their functionality.

Item 1

Cases for this item 97 Discrimination 0.02
Item Threshold(s): -5.13 Weighted MNSQ 2.85
Item Delta(s): -5.12

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
B	0.00	1	1.03	-0.02	-0.24(.810)	0.74	0.00
D	1.00	96	98.97	0.02	0.24(.810)	0.86	0.85

Item 35

Cases for this item 97 Discrimination 0.06
Item Threshold(s): 1.78 Weighted MNSQ 1.18
Item Delta(s): 1.79

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
A	0.00	2	2.06	-0.21	-2.06(.042)	-0.25	0.08
B	0.00	52	53.61	0.09	0.87(.389)	0.96	0.97
C	1.00	30	30.93	0.06	0.60(.553)	0.83	0.67
D	0.00	13	13.40	-0.13	-1.24(.219)	0.73	0.64

Item 8

Cases for this item 97 Discrimination -0.02
Item Threshold(s): 5.66 Weighted MNSQ 1.00
Item Delta(s): 5.66

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
9	0.00	1	1.03	-0.02	-0.24(.810)	0.74	0.00
A	0.00	1	1.03	0.03	0.26(.792)	0.50	0.00
B	1.00	1	1.03	-0.02	-0.24(.810)	0.58	0.00
C	0.00	9	9.28	-0.31	-3.23(.002)	0.17	1.12
D	0.00	85	87.63	0.28	2.89(.005)	0.94	0.80

Figure 2. Items with poor discrimination index

2.3.3. Distractor effectiveness

35/40 items of the test contained well-functioning distractors. Some of these items can be seen in Figure 3, which reveals that no option was neglected by the test-takers. For instance, of the four options of item 2, about 14% of the test-takers chose option A, and over 75% chose option B, leaving the other two options C and D with around 5% each (see Figure 3). Accordingly, these 35 items successfully challenged students and contributed positively to assessment quality.

Item 2

Cases for this item 97 Discrimination 0.30
Item Threshold(s): -0.50 Weighted MNSQ 1.09
Item Delta(s): -0.50

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
A	0.00	14	14.43	-0.24	-2.36(.020)	0.39	0.76
B	1.00	74	76.29	0.30	3.03(.003)	1.00	0.82
C	0.00	4	4.12	-0.13	-1.27(.207)	0.42	0.86
D	0.00	5	5.15	-0.08	-0.78(.436)	0.51	1.11

Item 29

Cases for this item 97 Discrimination 0.50
Item Threshold(s): -1.56 Weighted MNSQ 0.82
Item Delta(s): -1.56

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
A	0.00	3	3.09	-0.24	-2.46(.016)	-0.21	0.37
B	0.00	2	2.06	-0.31	-3.13(.002)	-0.76	0.84
C	1.00	87	89.69	0.50	5.59(.000)	0.98	0.74
D	0.00	5	5.15	-0.30	-3.03(.003)	0.02	1.34

Figure 3. Examples of items with good distractors

Nevertheless, some items contained ineffective distractors that failed to attract any responses. For example, in item 6, distractors B and D were entirely ignored; in item 22, distractor D was not selected by any candidate (see Figure 4). These non-functioning distractors likely appear too implausible, reducing the overall quality of these test items. Revisions are necessary to enhance their attractiveness and functionality.

Item 6

Cases for this item 97 Discrimination 0.27
Item Threshold(s): -4.13 Weighted MNSQ 1.01
Item Delta(s): -4.13

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
A	1.00	96	98.97	0.27	2.70(.008)	0.88	0.83
C	0.00	1	1.03	-0.27	-2.70(.008)	-1.20	0.00

Item 22

Cases for this item 96 Discrimination 0.21
Item Threshold(s): -1.10 Weighted MNSQ 1.02
Item Delta(s): -1.10

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
A	1.00	82	85.42	0.21	2.08(.040)	0.93	0.82
B	0.00	8	8.33	-0.10	-0.99(.323)	0.74	0.80
C	0.00	6	6.25	-0.19	-1.88(.064)	0.46	1.01

Figure 4. Examples of items with ineffective distractors

2.4. Discussion

The findings of this study highlight both the strengths and areas for improvement in the Mock VSTEP Level 3-5 Test's Reading Section. The results demonstrate that the test effectively differentiates candidates across reading proficiency levels, aligning well with the intended design. This supports the applicability of IRT in evaluating test quality, as emphasized by Nix and Tseng

(2014) [11]. However, discrepancies between intended and actual item difficulty levels, along with items exhibiting poor discrimination indices or ineffective distractors, underline challenges in constructing universally reliable test items. For example, items like 1, 6, and 17 were found to be too easy, while items like 8 and 20 were overly difficult, which aligns with findings from Bachman (1990) and Sireci (2009) that stress the complexity of designing balanced test items [18]-[19].

These findings have several implications for test design, administration, and future research. First, for test developers, the analysis reinforces the need to rigorously align item difficulty with test specifications during the design phase. Items with extreme difficulty levels or poor discrimination indices should be systematically reviewed to improve their quality. IRT provides a robust framework for such iterative improvements, ensuring that each item contributes meaningfully to the overall test's reliability and validity. Secodn, for educators and policymakers, the effectiveness of distractors in differentiating proficiency levels demonstrates their critical role in assessment design. Training for test developers should include the design of plausible distractors that reflect realistic candidate errors. Finally, for researchers, this study highlights the utility of IRT in identifying test strengths and weaknesses. Future studies could apply similar methods to other sections of the VSTEP or comparable standardized tests. Additionally, the integration of IRT with cognitive diagnostic approaches may provide deeper insights into how specific test items measure reading sub-skills.

3. Conclusions

The application of IRT has provided valuable insights into the Reading Section of the Mock VSTEP Level 3-5 Test, allowing for a thorough evaluation of its strengths and areas for enhancement. Overall, the analysis demonstrates that the test performs well in assessing and classifying candidates across a wide range of reading proficiency levels. Most test items exhibit strong alignment with their intended difficulty and discrimination parameters, while well-functioning distractors further contribute to the test's effectiveness. These strengths reflect the robustness of the exam's design and its capacity to measure candidates' reading abilities reliably.

However, the findings also reveal certain shortcomings that warrant attention. Specifically, some items, such as overly easy or overly difficult ones, require revisions to better align with their intended difficulty levels. Moreover, a small number of items with low or negative discrimination indices need to be improved to ensure they effectively differentiate between test-takers. Finally, ineffective distractors in some items should be redesigned to maintain the overall quality and challenge of the assessment. These areas for improvement underscore the importance of continual refinement to enhance the test's reliability and validity.

While this study provides valuable insights into the reading section of the Mock VSTEP Level 3-5 Test, it is limited in scope. Only the reading section was analyzed, and the dataset is restricted to a single test administration. Future research could expand to other test sections, such as listening, speaking, and writing, to provide a comprehensive evaluation of the Mock VSTEP. Additionally, longitudinal studies could explore the impact of item revisions based on IRT analysis on test validity and reliability over time.

REFERENCES

[1] Kane MT, (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. DOI: 10.1111/jedm.12000.

[2] Brunfaut T, (2023). Future challenges and opportunities in language testing and assessment: Basic questions and principles at the forefront. *Language Testing*, 40(1), 15–23. DOI: 10.1177/02655322221127896.

[3] An X & Yung YF, (2014). *Item response theory: What it is and how you can use the IRT procedure to apply it*. SAS Institute.

[4] Min S & Aryadoust V, (2021). A systematic review of item response theory in language assessment: Implications for the dimensionality of language ability. *Studies in Educational Evaluation*, 68, 100963. DOI: 10.1016/j.stueduc.2021.100963.

[5] Crocker L & Algia J, (1986). *Introduction to classical & modern test theory*. Belmont: Wadsworth.

[6] Henard DH, (1998). *Using spreadsheets to implement the one-parameter item response theory (IRT) model*. Paper presented at the annual meeting of the Southwester Psychological Association, New Orleans.

[7] Lee Y, Chon YV & Shin D, (2012). Vocabulary size of Korean EFL university learners: Using an item response theory model. *English Language & Literature Teaching*, 18(1), 171-195.

[8] Moore M & Gordon PC, (2015). Reading ability and print exposure: Item response theory analysis of the author recognition test. *Behavior research methods, 47*(4), 1095–1109.

[9] Hambleton RK & Slater SC, (1997). Item response theory models and testing practices: Current international status and future directions. *European Journal of Psychological Assessment*, 13(1), 21-28.

[10] Yamakawa K, Sugino N, Ohba H, Nakano M & Shimizu Y, (2008). Acquisition of English grammatical features by adult Japanese EFL learners: The application of item response theory in SLA research. *Electronic Journal of Foreign Language Teaching*, 5(1), 13-40. Centre for Language Studies, National University of Singapore.

[11] Nix J-ML & Tseng W-T, (2014). Towards the measurement of EFL listening beliefs with item response theory methods. *International Journal of Listening*, 28(2), 112-130. DOI: 10.1080/10904018.2013.872990.

[12] He L & Min S, (2016). Development and validation of a computer adaptive EFL test. *Language Assessment Quarterly*. DOI: 10.1080/15434303.2016.1162793.

[13] McCarron SP & Kuperman V, (2021). Is the author recognition test a useful metric for native and non-native English speakers? An item response theory analysis. *Behavior Research Methods*, 53, 2226–2237. DOI: 10.3758/s13428-021-01556-y.

[14] Hung TN, (2023). The predictive validity of standardized tests of English proficiency in Vietnam (VSTEP). *Tạp chí Khoa học Đại học Tân Trào*, 9(2). DOI: 10.51453/2354-1431/2023/917.

[15] Ministry of Education and Training, (2015). *Document Application Guidelines Format of test questions to assess English proficiency from level 3 to level 5 according to the 6-level Foreign Language Competency Framework for Vietnam in developing exam questions and marking exams* - Approved attached Decision No: 730/QD-BGDDT dated March 11, 2015 of the Minister of Education and Training. Hanoi.

[16] Wu ML, Adams RJ & Wilson M, (1988). *ACERConquest [computer program]*. Hawthorn, Australia: ACER.

[17] Bock RD & Aitkin M, (1981). Marginal maximum likelihood estimation of item parameters: *Application of an EM algorithm. Psychometrika*, 46(4), 443–459. DOI: 10.1007/BF02293801.

[18] Bachman L, (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

[19] Sireci SG, (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R.W.Lissitz (Ed.), *The concept of validity* (pp. 19-39). Charlotte, NC: Information Age Publishing, Inc.