PASAA Journal
Volume 69, July – December 2024, 533–560

PASAA

Computerized Adaptive Receptive Vocabulary Assessment Tool: An Experimental Study with Vietnamese EFL Learners

Bui Thi Kim Phuong[a], Nguyen Quy Thanh[b], Le Thai Hung[b]* and Nguyen Thai Ha[b]

[a] Hanoi University of Science and Technology, Hanoi, Vietnam

[b] University of Education – Vietnam National University, Hanoi, Vietnam

**Corresponding author: hunglethai82@gmail.com*

Article information

Abstract

Digital transformation has revolutionized education all over the world, leading to an innovative testing approach—computerized adaptive testing (CAT). In the Vietnamese educational context, despite being a relatively new area of research, CAT has steadily gained increasing interest from researchers and educators over the past few decades. While the number of studies conducted on the application of CAT remains limited in Vietnamese educational practices, the pioneering UEd-CAT system has reported initial results in assessing mathematical modeling and reading comprehension competencies among 10th graders. This study focused on the CARVAT - Computerized Adaptive Receptive Vocabulary Assessment Tool - developed within the UEd-CAT system to evaluate Vietnamese EFL learners' receptive vocabulary knowledge, and to expand the system's applicability beyond its current domains. With a standardized item bank imported into the system, the researchers conduct an experimental study with 98 Vietnamese EFL learners. The data of their performance and results were collected and analyzed to provide additional validation evidence of the CARVAT and the

	UEd-CAT system in educational assessment practices. It is also expected to provide practical insights into the application of computerized adaptive tests in EFL assessment and education in Vietnam.
Keywords	computerized adaptive testing, receptive vocabulary, EFL learners, experimental research, language assessment
APA citation:	Phuong, B. T. K., Thanh, N. Q., Hung, L. T., & Ha, N. T. (2024). Computerized adaptive receptive vocabulary assessment tool: An experimental study with Vietnamese EFL learners. *PASAA, 69*, 533–560.

1. Introduction

The era of digital transformation has impacted every aspect of education and promoted innovation in testing and assessment methods in education. In the field of language assessment, the application of computer technology has become more widespread in every home and school, thereby facilitating a more effective testing initiative—computerized adaptive language testing (CALT). In recent years, more computerized adaptive language tests have been developed (Pathan, 2012). Many scholars have also drawn their attention to discussing both advantages and disadvantages of CAT-integrated language assessment options in the past decades (Khoshsima & Toroujeni, 2017; Larson & Madsen, 1985; Okhotnikova et al., 2019; Pathan, 2012).

In the context of Vietnamese education and training, CAT remains a relatively new research field despite the country’s ever-growing interest in technology applications in education in recent years. Only a modest number of studies have been conducted on the development and validation of the UEd-CAT system of the University of Education - Hanoi National University (Le & Nguyen, 2021; Le et al., 2019; Nguyen & Le, 2018). Accordingly, the system has now developed and provided initial validation evidence for adaptive tests used to measure 10th grade students’ mathematical modeling and Vietnamese reading comprehension competencies. To further validate the UEd-CAT system, this study investigated the performance of CARVAT—the Computerized Adaptive Receptive Vocabulary Assessment Tool developed within the UEd-CAT system, addressing the following research questions:

1. How does the CARVAT assess Vietnamese EFL learners' receptive vocabulary knowledge?
2. What is the correlation between the CARVAT and the traditional fixed-format test?

This study aimed to offer additional validation evidence for the UEd-CAT system by investigating the adaptive system's potential for English vocabulary assessment in addition to its extant assessment capabilities, and to support the application of CAT in the field of language assessment in EFL contexts.

2. Literature Review

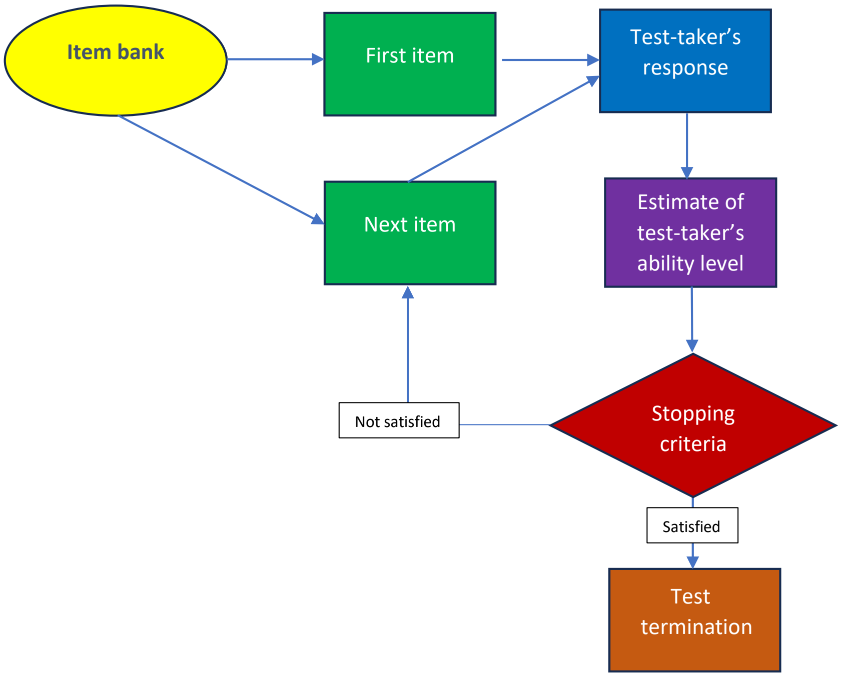
2.1 Computerized Adaptive Testing

CAT is a modern and effective approach to computer-based language testing that adapts the difficulty of test questions to the individual test-taker's language level. Instead of offering the same set of questions to all test-takers, as is the case in traditional paper-and-pencil tests or fixed-length computer-based tests, CAT customizes the test for each individual based on their responses to previous questions.

A CAT system comprises several key elements that work together to deliver adaptive tests tailored to each test-taker's ability level. These elements include a standardized item bank, algorithms for item selection, test-takers' ability estimates and test termination (Thompson & Weiss, 2019). The item bank is a collection of test items that are calibrated based on Item Response Theory (IRT) and organized based on their difficulty levels and content domains. Questions are selected from the item bank for each step of the CAT procedure. In a complete testing process as shown in Figure 1, the administration of the test begins with a question selected from a calibrated item bank. This starting question can be chosen randomly or from a group of medium difficulty items in the item bank (Choi & McClenen, 2020; Oppl et al., 2017). After the test-taker responds to the initial item, their response is scored, and their ability level is estimated based on this response and any previous responses when applicable. Next, the CAT algorithm uses the estimated ability level to select the next test item. The item chosen is intended to be at the appropriate difficulty level to refine the estimate of the test-taker's ability. The test ends once the predefined stopping criterion, such as a minimum number of questions or a desired level of measurement precision, is satisfied. It is noteworthy that the process of presenting an item, scoring the response, and selecting the next item continues iteratively, with each item selection being based on the test-taker's previous responses. The CAT aims to determine as accurately as possible the test-taker's ability level.

Figure 1

CAT process



2.2 CAT Potential for Language Assessment

CALT has enormous advantages over conventional fixed-format language testing, including paper-based and computer-based language tests.

Ensuring Precision

CAT provides highly precise estimates of a test-taker's language proficiency. The algorithm selects questions strategically to maximize the information gained from each response, resulting in a reliable proficiency score (Choi et al., 2003; Gawliczek et al., 2021; Giouroglou & Economides, 2003; He & Min, 2017; Khoshsima & Toroujeni, 2017; Larson & Madsen, 1985; Meunier, 1994; Mizumoto et al., 2019; Pathan, 2012).

Improving Efficiency

Computerized adaptive tests are typically shorter in length compared to fixed-length tests because they only administer questions that are relevant to the test-taker's ability level, which reduces the number of items used and the test-taking time (Giouroglou & Economides, 2003; Khoshsima & Toroujeni, 2017; Meunier, 1994; Mizumoto et al., 2019; Okhotnikova et al., 2019; Pathan, 2012; Tseng, 2016). Moreover, random selection of questions from a large item bank reduces the likelihood of test-takers sharing questions and correct answers, which helps strengthen test security (Meunier, 1994; Okhotnikova et al., 2019; Pathan, 2012; Rasskazova et al., 2017; Vie et al., 2017).

Optimizing Test Experiences

Test-takers may experience less test anxiety with adaptive language tests because they are not exposed to questions that are significantly above or below their proficiency level (Gawliczek et al., 2021; Giouroglou & Economides, 2003; Meunier, 1994; Rasskazova et al., 2017; Wise, 2014). Moreover, the fact that test-takers often receive immediate results and feedback after completing an adaptive test allows them to quickly identify their language proficiency level and areas for improvement (Burston & Neophytou, 2014; Gawliczek et al., 2021; Meunier, 1994).

Offering Versatile Applications

CAT can be used in a variety of purposes of language assessment, including diagnostic assessment (Mizumoto et al., 2019), formative assessment (Choi & McClenen, 2020; Giouroglou & Economides, 2005), high-stake testing (He & Min, 2017), and large-scale assessment (Khoshsima & Toroujeni, 2017; Pathan, 2012).

However, the development and implementation of a CAT system can be complicated, as it requires a robust item bank, a well-designed algorithm, and thorough validation to ensure the reliability and validity of the results. In addition, ongoing research and development are necessary to keep CAT systems updated and aligned with evolving language standards and teaching methodologies. Overall, CAT has revolutionized language assessment by providing a more efficient, precise, and personalized means of measuring language proficiency.

2.3 Receptive Vocabulary Assessment

Since vocabulary is fundamental to all language use, vocabulary assessment plays an important role, bringing great value to both research and practice in language teaching and learning. First, the assessment of English vocabulary knowledge is clearly related to determining the language proficiency of the test-taker (Schmitt et al., 2017) and is therefore an important part of language learning. After taking an English vocabulary test, learners can self-determine their level, promote their learning, and raise their awareness of different aspects of vocabulary thanks to the test results (Yanagisawa & Webb, 2019). English vocabulary tests can also be used for other purposes of the teaching and learning process such as pointing out learners' difficulties to find solutions, placing them in appropriate-level classes, tracking learning progress, and evaluating the success of a course (Gyllstad, 2019; Kremmel, 2019; Nation, 2013). Moreover, vocabulary tests can influence whether vocabulary learning continues or not (Read, 2019) and provide insight into the impact of learning experiences on vocabulary development (Stoeckel & Bennett, 2015).

English vocabulary knowledge assessment can target three main aspects, each of which includes three sub-aspects: (1) word form (sub-aspects: spoken form, written form, and word components), (2) word meaning (sub-aspects: word form and meaning, concepts and references, and links), and (3) word use (sub-aspects: grammatical function, word combination, and usage constraints) (Nation, 2013). Of these, the receptive vocabulary knowledge, which is assessed through one's recognition of the relationship between form and meaning of each word, is the foundation for learning and acquiring other aspects of vocabulary and has therefore been the focus of many studies on vocabulary assessment over the years (Webb & Chang, 2012).

In the realm of English language education in Vietnam, a limited number of studies have examined Vietnamese EFL learners' vocabulary knowledge, using various written receptive vocabulary tests and different vocabulary lists. Nguyen and Nation (2011) employed a bilingual version of the Vocabulary Size Test (VST) by Nation and Belgar (2007) using the BNC/COCA word list (Nation, 2012). Vu and Nguyen (2019) utilized the Vocabulary Levels Test (VLT) by Schmitt et al. (2001) with the General Service List (West, 1953). Nguyen and Webb (2017) employed a test of English collocations at the first three word-frequency levels (1,000, 2,000, 3,000) from the BNC/COCA list by Nation (2012). Dang (2020) applied the updated VLT by Webb et al. (2017) with 5000 words from the BNC/COCA word list (Nation, 2012). More recently, Bui et al. (2024) used the bilingual version of the New General Service List Test with Browne's New General Service List (NGSL) (2013). These studies collectively contribute to reflecting the vocabulary knowledge of Vietnamese EFL learners and offer recommendations for the development of vocabulary assessment alternatives that are appropriate for EFL learners, aiming to enhance language education and assessment practices in this context.

2.4 Previous Computerized Adaptive Tests

Given that numerous studies have embraced the use of CAT to assess language skills, particularly receptive, but focused mostly on English-as-the-first-language, with limited attention given to English as a Second Language (ESL) or English as a Foreign Language (EFL) contexts (Mizumoto et al., 2019; Tseng, 2016), it is crucial to address the scarcity of studies that explore CAT in ESL and EFL contexts. Research in this area could provide valuable insights into how CAT can offer more precise and adaptive means of language assessment across various teaching and learning contexts, thereby potentially opening new avenues for effective language assessment and education and benefiting learners, educators, and researchers.

Notable studies applying CAT in assessing English receptive vocabulary include CATPRO, developed as a totally novel initiative to evaluate students, vocabulary size (Tseng, 2016); CATSS, catering to both receptive and productive vocabulary assessment (Aviad-Levitzky at al., 2019); and CAT-WPLT, focused on a specific aspect, which is knowledge about affixes (Mizumoto et al., 2019). Despite these differences, it is worth underscoring that CAT offers exceptional efficiency and features for assessing English receptive vocabulary, especially when supplemented with a substantial bank of discrete-point items that have been calibrated using Item Response Theory (IRT) models. Also, in these studies, the evaluation of CAT's validity and reliability in measurement often relied on test-takers' performance within the system. Additionally, comparisons were made between the effectiveness of CAT and traditional fixed tests. This study aimed to follow this direction to highlight the CARVAT's features and evaluate the CARVAT's effectiveness.

3. Methodology

3.1 Research Participants

The experimental study involved 98 students from a Vietnamese university, selected through simple random sampling. All participants were enrolled in A2-level English courses. The group comprised 63 males and 35 females from various academic disciplines, though none were English majors. All the students were interested in finding an effective tool to assess their vocabulary knowledge and volunteered for the study. Each student made between one and seven attempts to test their vocabulary knowledge in the system and completed a fixed 100-item test, providing researchers with valuable data to address the two research questions.

3.2 Data Collection Instruments

3.2.1 The CARVAT

The CARVAT consists of two key components: adaptive algorithms and a standardized item bank for English vocabulary assessment.

The adaptive algorithms of the system have been summarized in previous publications about the UEd-CAT system (Le & Nguyen, 2021; Nguyen & Le, 2018). As a scientific product of the research team from the Faculty of Quality Management - VNU University of Education, the adaptive system is developed based on the Maximum a Posteriori estimation, Gradient Descent algorithm, IRT one-parameter model and programmed in PHP and JavaScript languages. It is reported that the UEd-CAT 1.0 system ensures outstanding features of CAT in assessing mathematical modeling and Vietnamese reading comprehension competencies of 10th-grade students with a smaller number of items used in each test, thereby shortening test-taking time. Moreover, assessment accuracy is still guaranteed, along with test-takers' readiness for adaptive testing.

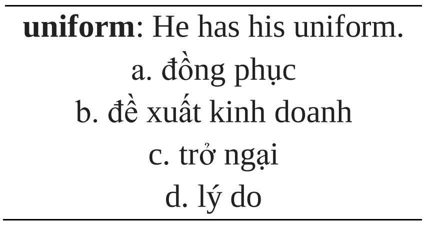
The second component of the CARVAT is the newly designed item bank of English receptive vocabulary. The construction of a question bank of 552 items followed a predefined specification of the bilingual New General Service List (NGSLT) conducted by Bui et al (2022). After expert appraisal with the participation of experienced and certified reviewers, the raw item bank was divided into seven tests for an experiment with 1619 students. As a result of the test analysis with Conquest software, 30 questions were omitted because they did not match the model, while 84 questions were edited to increase the quality of the distractors. The difficulty indexes of all items in the bank were then calibrated on the same scale thanks to the equation using R statistical software with the equation method of Loyd and Hoover (1980). Finally, a standardized bank of 522 items was determined to meet the requirements and imported into the UEd-CAT system. It should be emphasized that the research team succeeded in standardizing a bank with a large number of more than 500 items accompanied with difficulty indexes for English receptive vocabulary, which is similar to the studies conducted for math and reading comprehension within the UEd-CAT system (Nguyen et al., 2023).

3.2.2 Fixed Test

The fixed test used in this study is the bilingual version of the New General Service List Test (NGSLT), a diagnostic test of English receptive vocabulary knowledge using the NGSL (Browne, 2013). Targeting Vietnamese EFL learners, the bilingual version of the NGSLT consists of 100 target words in English while options are constructed using Vietnamese, which is the test-takers' native language (Bui et al., 2022) (See one example item in Figure 2). It should be noted that this bilingual version is not unprecedented as reliable evidence has been presented in previous studies to affirm the suitability and benefits of the bilingual version, especially when target test-takers may possess limited levels of grammar knowledge and reading skills (Stoeckel et al., 2018).

Figure 2

Example item of the bilingual NGSLT (Bui et al., 2022)



uniform: He has his uniform.
a. đồng phục
b. đề xuất kinh doanh
c. trở ngại
d. lý do

3.3 Data Collection Procedures

The experimental study received ethical approval from a Vietnamese university. Prior to confirming their voluntary participation and submitting consent forms, all participating students were informed about the study's purpose and given participation instructions. In the first phase, they were provided with accounts and detailed instructions to make multiple attempts in the adaptive testing system. This phase saw 98 students voluntarily participating, resulting in 209 test attempts, as detailed in Table 1. In the second phase, the students completed a fixed test consisting of 100 questions in a traditional testing format to assess their vocabulary knowledge.

Research ethics were taken into thoughtful consideration in all the stages, from sampling to sending invitations to participants, from conducting the experiment to collecting data. All personal information and answers were subject to confidentiality and only used for research purposes. The results were reported accurately and honestly, and the researcher anonymized all participants to ensure that their personal information, dignity, and rights would not be affected in any way.

Table 1

Summary of test administrations on the system

	Number of participants	Number of test attempts in the system
	40	1
	26	2
	20	3
	8	4
	3	6
	1	7
Total	**98**	**209**

3.4 Data Analysis

Test-takers' performance and results were analyzed to provide evidence for validating the adaptive test. The analysis steps included the use of SPSS and R statistical software to conduct descriptive statistics of test-takers' performance and results in the adaptive system to describe the adaptive features of the

CARVAT. Then, the correlation between candidates' ability scores on the adaptive testing system and test scores on the fixed test was analyzed and interpreted. Based on the results of the analysis, the researcher provided evidence about the validity and reliability of the CARVAT in assessing test-takers' English receptive vocabulary.

4. Results

4.1 The CARVAT's Operation for Receptive Vocabulary Assessment

In this section, the participants' performance in the system will be reported to illustrate the adaptive system's operation with emphasis on its outstanding features related to efficiency and accuracy.

Test-taking Time

As can be seen from Table 2, the CARVAT administered different tests of six to 20 questions to test-takers. With the study sample, a test of 20 questions was the most common. With much fewer items in the test sets delivered by the CARVAT, the test-taking time ranged from 12 seconds with six questions to 12 minutes three seconds (723 seconds) with twenty questions. For the longest tests with 20 items, test-taking time varied between 110 seconds and 723 seconds. The time efficiency of the CARVAT is clearly evident compared to the 100-item test, which takes about ten to 20 minutes to finish.

Table 2

Tests provided in the CARVAT

	Number of items in each test													**Total**
	6	**8**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	
Tests	2	2	3	4	20	5	2	4	23	10	2	17	115	209

Adaptive Path of Difficulty in the CARVAT

Figure 3 illustrates the adaptive path of difficulty taken on May 27th by one test-taker, whose ID is HONG. The total number of questions in the test was20.

The test-taker provided ten correct answers and got a score of 59.93. It can be seen that when HONG answered one question correctly, the difficulty of the next item was higher, and when she answered the question incorrectly, the system responded in two directions to adapt to the test-taker's ability. With the incorrect answer to item 8, the system provided the test-taker with a next item with a lower level of difficulty. With the incorrect answer to item 6, however, the system increased the difficulty for the next item but the difference in the difficulty level was reduced in comparison to the difference in the difficulty between Item 4 and Item 5.

Figure 3

Items' difficulty in HONG's test

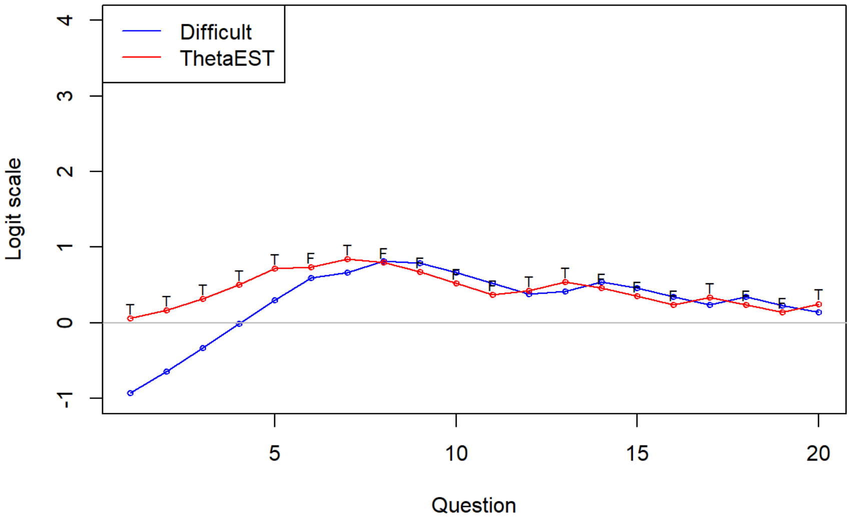
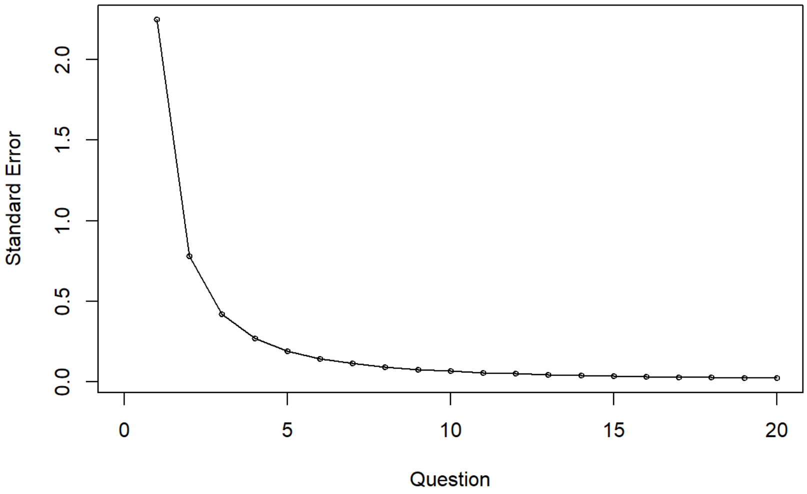


Figure 4

Standard error of the estimate after each question in HONG's test

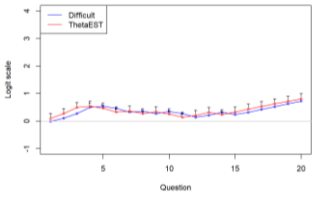
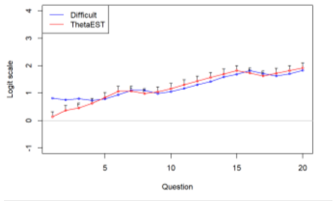
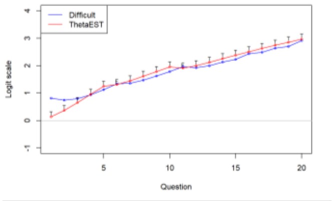


The ability-estimating process is iteratively repeated after each item on the basis of the test-taker's answer and the difficulty level of the item. The standard error of estimation gradually decreases in the process until it satisfies the stopping criteria of standard error that is predetermined at 0.02 in the system. In this case, the test ended after the test-taker answered the 20th question, and the system finished estimating the test-taker's ability with the standard error of 0.0227 (Figure 4).

Variety of Items Offered to Test-takers

Table 3

DAN's three testing attempts in the CARVAT

Test-taker's ID: DAN								
Test date: 22 May			Test date: 27 May			Test date: 6 June		
No	Item	Difficulty	No	Item	Difficulty	No	Item	Difficulty
1	D5:25L2	-0.0082	1	D7:88L5	0.82	1	D7:88L5	0.82
2	D1:81L4	0.1089	2	D1:78L4	0.7519	2	D3:30L2	0.7498
3	D5:34L3	0.2725	3	D1:46L3	0.7988	3	D1:46L3	0.7988
4	D7:63L4	0.5062	4	D3:62L3	0.7409	4	D1:82L5	0.9431
5	D3:91L5	0.5458	5	D3:69L4	0.7881	5	D2:78L4	1.1326
6	D4:36L3	0.4619	6	D1:82L5	0.9431	6	D7:93L5	1.344
7	D5:55L3	0.3249	7	D1:23L2	1.1019	7	D4:31L2	1.3587
8	D1:94L5	0.3541	8	D6:57L3	1.1016	8	D7:87L5	1.4732
9	D7:76L4	0.2739	9	D2:89L5	0.9874	9	D4:85L5	1.63
10	D2:13L1	0.3453	10	D5:71L4	1.0546	10	D2:83L5	1.7891
11	D5:32L3	0.2735	11	D7:86L5	1.17	11	D2:95L5	1.9612
12	D4:60L3	0.1414	12	D6:74L4	1.3071	12	D4:74L4	1.9286
13	D6:98L5	0.2154	13	D5:96L5	1.4183	13	D7:55L3	2.0014
14	D1:52L3	0.3178	14	D4:99L5	1.5889	14	D7:95L5	2.1368
15	D1:37L3	0.2361	15	D7:73L4	1.6915	15	D4:47L3	2.2349
16	D3:56L3	0.3203	16	D4:52L3	1.8215	16	D2:72L4	2.4374
17	D5:29L2	0.4287	17	D5:78L4	1.7157	17	D2:100L5	2.488
18	D7:54L3	0.5278	18	D4:85L5	1.63	18	D5:95L5	2.6408
19	D3:75L4	0.6265	19	D2:73L4	1.7065	19	D1:35L3	2.7066
20	D4:67L4	0.7274	20	D2:71L4	1.8304	20	D6:38L3	2.9185
								

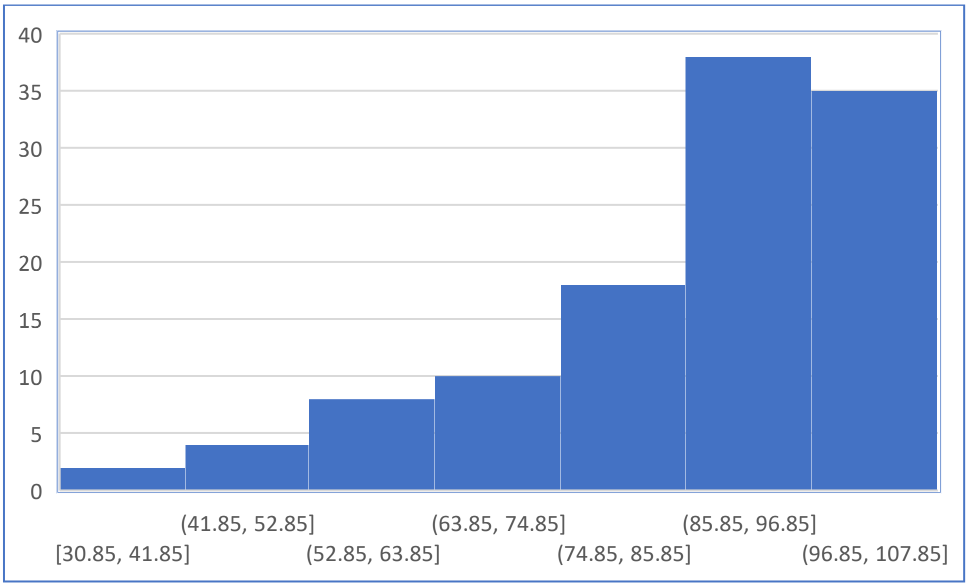
It is noteworthy that even though one test-taker could have the same number of questions in his or her tests, he or she encountered a different adaptive path of items in each attempt with different items. The rate of overlapping items among different tests was minimized thanks to the substantial size of the item bank. An example, as shown in Table 3, was the case of one test-taker who performed three testing attempts in the CARVAT. This example illustrates the varied jumps in item difficulty in the 20-item tests and the rate at which the test-taker encountered items.

It can also be seen from Table 3 that of the 60 items he encountered in the CARVAT, only four were repeated in the second and third attempts. These overlapping items are highlighted with the same colors in the table. It can be indicated that the use of different test items can eliminate the possibility of test-takers' remembering specific items, thus ensuring that earlier test-taking attempts have little effect on the results of later testing attempts, ensuring test security and supporting regular assessment for learning.

Discrimination Possibilities with Tests of the Same Number of Questions

Figure 5

Test-takers' scores in 20-item tests



With the same number of items in the tests, the test results were different, reflecting distinct levels of receptive vocabulary among the test-takers. As can be seen from Figure 5, the different tests with the same number of items (20 items) could divide the test-takers into diverse groups of ability: "low level" with scores below 52.85, "middle level" with scores up to 85.85, and "prominent level" with scores above 85.85. This is an indication that the CARVAT can classify diverse groups of test-takers by their ability.

Consistency in the Scores of One Test-taker in Different Attempts

Table 4 shows the CARVAT results of DANH and TDUC. While DANH completed the CARVAT with three 20-item tests, TDUC completed three tests with different numbers of items. Despite this difference, the results they received were consistent enough to show their prominent level of receptive vocabulary.

Table 4

CARVAT results in different attempts

Test-taker's ID	TDUC			DANH		
Test	1	2	3	1	2	3
Number of questions	20	17	10	20	20	20
Number of correct answers	19	15	8	15	13	18
CARVAT score	99.91	99.97	99.63	97.24	96.24	96.85

Test-retest Reliability

To take a further step to evaluate the reliability of the CARVAT test scores, the test-retest reliability coefficient was calculated for the test-takers who took the test multiple times in the system. Following Ye's (2014) suggestion, the reliability coefficient was determined by correlating the scores from the first and second attempts in the CARVAT, despite its adaptive nature where the same test-taker may encounter different items in each attempt. Based on the data obtained from 58 test-takers who retook the test in the CARVAT, the reliability coefficient for the CARVAT was found to be .88, demonstrating high reliability of the CARVAT scores.

4.2 Correlation Between CARVAT and Fixed Test Scores

The test scores of 98 participants in the fixed test and CARVAT are synthesized in Table 5 to provide an overview of the participants' vocabulary level. The test-takers' scores ranged from 27 to 99 in the fixed test, and from 7.61 to 99.98 in the CARVAT. The average score of the CARVAT was higher than that of the fixed test.

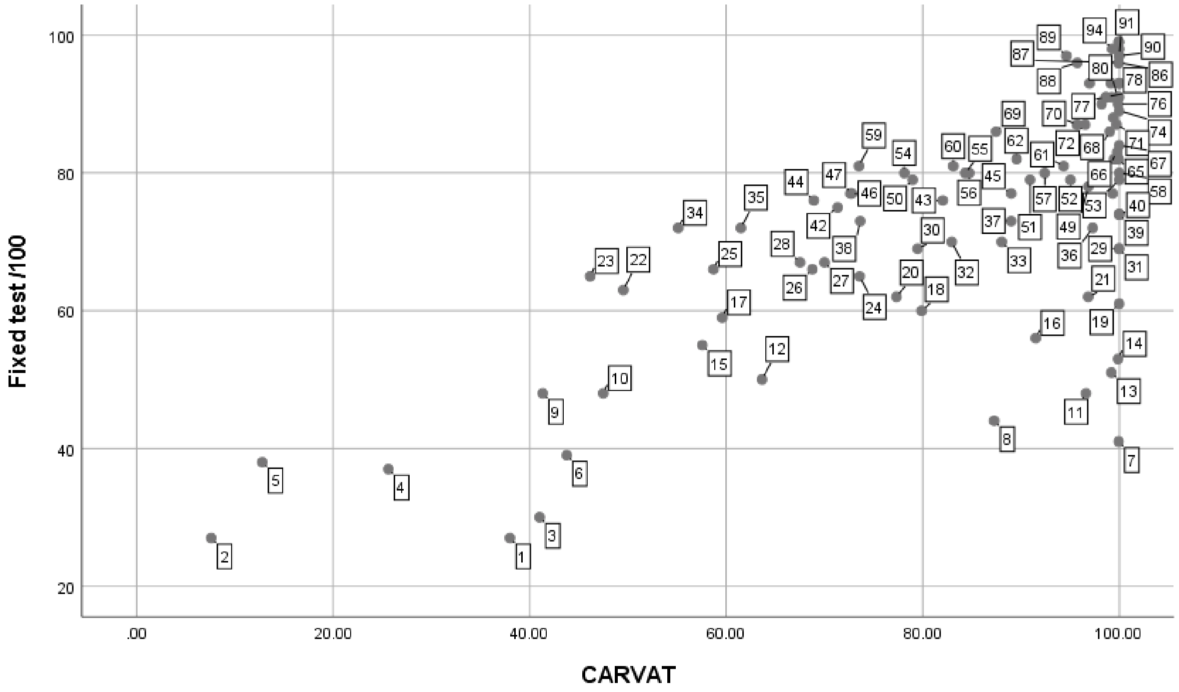
Table 5

Descriptive Statistics of CARVAT and fixed test scores

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Fixed test	98	72	27	99	75.5	17.8
CARVAT	98	92.37	7.61	99.98	84.5	20.2

Figure 6

Scatter plot of scores in fixed test and CARVAT



As can be seen from Figure 6, the participants covered varied levels of vocabulary knowledge, from low to high, with the majority belonging to the higher level. Moreover, there is a potential relationship between the scores of fixed tests and those of the CARVAT.

Table 6

Pearson correlation analysis

		Fixed test
CARVAT	Pearson Correlation	.717**
	Sig. (2-tailed)	.000

N	98

A correlation analysis was performed between fixed test and CARVAT scores to provide more validation evidence for the validity and reliability of the CARVAT. As can be seen from Table 6, the Sig. (2-Tailed) value is .000, which is less than .05. meaning that there is a statistically significant correlation between the two studied variables. The coefficient of 0.717 in Pearson correlation analysis indicates a strong positive relationship between the scores of the fixed test and those of CARVAT. It can be interpreted that the CARVAT can also reliably generate reliable scores to indicate the test-takers' receptive vocabulary knowledge levels, just like the fixed test.

5. Discussion

This study aimed to investigate the possibility of creating an assessment tool of English receptive vocabulary in the UEd-CAT system for Vietnamese EFL learners. Using available adaptive algorithms of the UEd-CAT system, the researchers imported a standardized bank of bilingual items so that the CARVAT can be used as an alternative to assess English receptive vocabulary in the Vietnamese EFL teaching and learning context.

The findings have revealed that the CARVAT possesses remarkably striking features. First of all, with the same objectives as the NGSLT, the CARVAT can determine the test-taker's receptive vocabulary knowledge with much fewer items, from six to 20 instead of 100 items. Only about one-fifth of the test items are needed in the CARVAT in comparison to the fixed test. As a result, both the time spent, as well as effort made by the test-takers, could be reduced. This efficiency of the CARVAT provides more evidence to support economy of time as a strength of CAT in educational assessment practices along with previous studies (Giouroglou & Economides, 2003; Khoshsima & Toroujeni, 2017; Meunier, 1994; Mizumoto et al., 2019; Okhotnikova et al., 2019; Pathan, 2012; Tseng, 2016). Secondly, adaptive algorithms of the CARVAT can help select relevant items tailored to the test-takers' ability from the item bank and provide varied items and tests with a low rate of overlapping among one test-taker's different attempts. This feature can enhance the test-takers' experiences as it may relieve them from test stress and boredom and reduce some test security threats (Gawliczek et al., 2021; Giouroglou & Economides, 2003; Meunier, 1994; Rasskazova et al., 2017; Wise, 2014; Vie et al. 2017).

Furthermore, similar to the fixed version of the NGSLT, the CARVAT generates scores that can be interpreted to determine a test-taker's receptive vocabulary, a fundamental aspect of language mastery. The test-retest reliability and the correlation analysis of scores in the fixed test and in the CARVAT showed the high reliability of the CARVAT in comparison to the fixed test. In other words, the CARVAT can be used as a trustworthy alternative to the 100-item fixed test with guaranteed precision, which could help in directing Vietnamese EFL learners towards mastering the NSGL of 2801 high-frequency words, thereby providing a solid foundation for effective language teaching and learning (Webb & Chang, 2012). This is also consistent with the findings and discussions in earlier publications on the accuracy offered by CALT (Choi, Kim & Boo, 2003; Gawliczek et al., 2021; Giouroglou & Economides, 2003; He & Min, 2017; Khoshsima & Toroujeni, 2017; Larson & Madsen, 1985; Meunier, 1994; Mizumoto et al., 2019; Pathan, 2012).

The usefulness of the CARVAT was also explored in this study. Along with other adaptive tests like CATSS, CARPRO, and CAT-WPLT, the CARVAT proves that CAT offers exceptional and efficient features for assessing English receptive vocabulary (Aviad-Levitzky at al., 2019; Mizumoto et al., 2019; Tseng, 2016). Looking ahead, thanks to the positive findings from the CALT-based studies, the development of adaptive language assessments can be diversified to encompass various linguistic domains. Initially, thematic vocabulary assessments could be introduced, catering to specific subjects or different practical purposes. Subsequently, assessments might venture into the assessment of EFL learners' grammar and pronunciation, as well as receptive skills like reading and listening. The ultimate goal is to provide EFL learners with free access to effective tools and contribute to the literature on language assessment to promote the potential of CALT in EFL contexts.

Similar to other assessments in the UEd-CAT system, the CARVAT can serve as a means for self-assessment and regular practice (Le et al., 2019; Le & Nguyen, 2021; Nguyen & Le, 2018). It should also prove to be valuable when employed to evaluate and expand one's knowledge and ability. It is also expected to contribute to students' adaptive learning for better performance and results. Additionally, offering options for students to test and learn vocabulary independently may nurture essential self-directed learning skills, which are invaluable for lifelong learning.

It is believed that CAT-integrated assessment tools not only empower learners to take control of their learning process but also provide educators with valuable insights for more effective teaching strategies. It is also implied for researchers and system developers that CAT requires ongoing validation, refinement, and innovation efforts to serve the ever-changing needs of education practices so that it can remain effective and accessible for all, regardless of their circumstances or challenges. Only with rigorous development methodologies can valid and reliable assessment tools be yielded for language assessment and enhance educational practices.

6. Conclusion

The researchers conducted the trialing of the CARVAT to assess Vietnamese EFL learners' receptive vocabulary. The CARVAT is embedded in the UEd-CAT system developed by the University of Education - Hanoi National University, with 522 IRT-calibrated items, and is designed to dynamically make adjustment to the difficulty of the questions based on test-takers' responses, resulting in an efficient and precise assessment tool.

The findings have indicated that the CARVAT is effective in assessing the test-takers' receptive vocabulary while using fewer items compared to fixed-format tests. The adaptive nature of the tool also allows for test-takers' positively personalized experience, where they encounter items relevant to their ability level. In general, the CARVAT succeeds in broadening the UEd-CAT system's utility beyond its existing scope. Moreover, the CARVAT's outstanding features make it a promising alternative for Vietnamese EFL learners' receptive vocabulary assessment and a good contribution to the field of language assessment in EFL contexts. Finally, the positive results of this study pave the way for more widespread application of CAT in educational practices with different contents and purposes.

However, there are certain limitations in the study. Firstly, the number of test-takers participating in the study remains limited. Future studies may consider expanding the sample size and investigating with test-takers of different language levels and learning contexts. Additionally, they can examine test-takers' viewpoints after their direct exposure to the test in order to obtain more useful and multidimensional results. Finally, future scholarly works could be directed towards other contents for different assessment purposes to maximize the potential of CAT in language assessment and in education in general.

7. About the Author

Bui Thi Kim Phuong is a senior lecturer at the Faculty of Foreign Languages, Hanoi University of Science and Technology. She holds a Ph.D. degree in Educational Measurement and Assessment from the VNU University of Education, Vietnam. Her research interests include TESOL, translator training, educational assessment, and measurement.

Nguyen Quy Thanh is a professor at the VNU University of Education, Vietnam. His research interests include social sciences, educational accreditation

and quality assurance, testing, university ranking, and competency assessment. He has had more than eighty scientific articles and book chapters published in prestigious national and international scientific journals and publishers. He is currently the Rector of the VNU University of Education, Vietnam.

Le Thai Hung is an associate professor at the VNU University of Education, Vietnam. His research interests include quality management, educational assessment, and measurement. He has published more than fifty books and articles related to physics and educational assessment. He is now the Vice Rector of the VNU University of Education, Vietnam.

Nguyen Thai Ha is a lecturer of the Quality Management Faculty at the VNU University of Education, Vietnam. He holds a master's degree in educational measurement and assessment at the VNU University of Education. His main research direction is the application of mathematics in the science of educational measurement and evaluation.

8. References

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.

Browne, C. (2013). The new general service list: Celebrating 60 years of vocabulary learning. *The Language Teacher*, *37*(4), 13–16.

Bui, T. K. P., Le, T. H., Tran, T. T. A., & Tran, T. T. A. (2024). Kiểm tra, đánh giá kiến thức từ vựng của sinh viên đại học với danh sách từ vựng tiếng Anh thông dụng mới [Assessing university students' vocabulary knowledge of the New General Service List]. *Journal of Education*, *24*(12), 24–28.

Bui, T. K. P., Nguyen, Q. T., & Le, T. H. (2022). *The development of a Vietnamese-English bilingual version of the New General Service List Test*. 2nd Hanoi Forum on Pedagogical and Educational Sciences, Hanoi, Vietnam.

Burston, J., & Neophytou, M. (2014). Lessons learned in designing and implementing a computer-adaptive test for English. *The EuroCALL Review*, *22*(2), 19–25. https://doi.org/10.4995/eurocall.2014.3632

Choi, I.-C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*, 295–320. https://doi.org/10.1191/0265532203lt258oa

Choi, Y., & McClenen, C. (2020). Development of adaptive formative assessment system using computerized adaptive testing and dynamic bayesian networks. *Applied Sciences*, *10*(22), Article 8196. https://doi.org/10.3390/app10228196

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), 1–11. https://doi.org/10.25073/2525-2445/vnufs.4553

Gawliczek, P., Krykun, V., Tarasenko, N., Tyshchenko, M., & Shapran, O. (2021). Computer adaptive language testing according to NATO STANAG 6001 requirements. *Advanced Education*, *8*(17), 19–26. https://doi.org/10.20535/2410-8286.225018

Giouroglou, H., & Economides, A. (2003). *Cognitive CAT in foreign language assessment*. Eleventh International PEG Conference, Powerful ICT Tools for Learning and Teaching, St. Petersburg, Russia.

Gyllstad H. (2020). Measuring knowledge of multi-word items. In Webb S. (Ed.), *The Routledge handbook of vocabulary studies* (pp. 387–405). Routledge. https://doi.org/10.4324/9780429291586-25

He, L., & Min, S. (2017). Development and validation of a computer adaptive EFL test. *Language Assessment Quarterly*, *14*(2), 160–176. https://doi.org/10.1080/15434303.2016.1162793

Khoshsima, H., & Toroujeni, S. M. H. (2017). Computer adaptive testing (Cat) design; testing algorithm and administration mode investigation. *European Journal of Education Studies*, *3*(5), 764–795.

Kremmel, B. (2019). Measuring vocabulary learning progress. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 406-418). Routledge.

Larson, J. W. & Madsen. H. S. (1985). Computer-adaptive language testing: Moving beyond computer-assisted testing. *CALICO Journal*, *2*(3), 32–36. https://doi.org/10.1558/cj.v2i3.32-37

Laurier, M. (2000). Can computerized testing be authentic? *ReCALL, 12*(1), 93–104. https://doi.org/10.1017/s0958344000001014

Le, T. H. & Nguyen T. H. (2021). *Experimental research and application of computerized adaptive tests to assess learners' competencies*. 2021 3rd International Conference on Computer Science and Technologies in Education (CSTE), Beijing, China*.* https://doi.org/10.1109/cste53634.2021.00021

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), 49–63. https://doi.org/10.25073/2588-1159/vnuer.4301

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*(3), 179–193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x

Meunier, L. E. (1994). Computer adaptive language tests (CALT) offer a great potential for functional testing. Yet, why don't they?. *CALICO journal*, *11*(4), 23–39. https://doi.org/10.1558/cj.v11i4.23-39

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

Nation, I. S. P. (2012). The BNC/COCA word family lists. http://www.victoria.ac.nz/lals/about/staff/paulnation

Nation, I. S. P. (2013). *Learning vocabulary in another language*. Cambridge University.

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

Nguyen, T. H., Bui, T. K. P., & Le, T. H. (2023). Phát triển ngân hàng câu hỏi trắc nghiệm thích ứng đánh giá từ vựng tiếng Anh thông dụng: áp dụng IRT và phương pháp cân bằng đề [The development of CAT item bank to assess the knowledge of high frequency English words: Applying IRT and equating methods]. *Journal of Education, 23*(19), 8–14.

Nguyen, T. G. & Le, T. H. (2018). Mô phỏng một bài kiểm tra thích nghi trên máy tính thông qua phần mềm R [Simulating an adaptive test on a computer through R software]. *Vietnam Journal of Educational Sciences, 11*, 6–11.

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

Okhotnikova, A., Daminova, J., Muzafarova, A., & Rasskazova, T. (2019). *Challenges of designing and administering computer-adaptive tests*. 13th International Technology, Education and Development Conference (INTED), Valencia, Spain. https://doi.org/10.21125/inted.2019.1383

Oppl, S., Reisinger, F., Eckmaier, A., & Helm, C. (2017). A flexible online platform for computerized adaptive testing. *International Journal of Educational Technology in Higher Education*, *14*(1), 1–21. https://doi.org/10.1186/s41239-017-0039-0

Pathan, M. M. (2012). Computer Assisted Language Testing [CALT]: Advantages, implications and limitations. *Research Vistas*, *1*(4), 30-45.

Rasskazova, T., Muzafarova, A., Daminova, J., & Okhotnikova, A. (2017). Computerised language assessment: Limitations and opportunities. *eLearning & Software for Education*, *2*. https://doi.org/10.12753/2066-026x-17-110

Read, J. (2019). Key issues in measuring vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 545–560). Routledge.

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., 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.

Stoeckel, T., & Bennett, P. (2015). A test of the new General Service List. *Vocabulary Learning and Instruction*, *4*(1), 1–8. https://doi.org/10.7820/vli.v04.1.stoeckel.bennett

Stoeckel, T., Ishii, T., & Bennett, P. (2018). A Japanese-English bilingual version of the new general service list test. *JALT Journal*, *40*(1), 5–21. https://doi.org/10.37546/jaltjj40.1-1

Thompson, N. A., & Weiss, D. A. (2019). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation*, *16*(1), 1. https://doi.org/10.7275/wqzt-9427

Tseng, W. T. (2016). Measuring English vocabulary size via computerized adaptive testing. *Computers & Education*, *97*, 69–85. https://doi.org/10.1016/j.compedu.2016.02.018

Vie, J. J., Popineau, F., Bruillard, É., & Bourda, Y. (2017). A review of recent advances in adaptive assessment. In A. Peña-Ayala (Ed.), *Learning analytics: Fundaments, applications, and trends: A view of the current state of the art to enhance e-learning* (pp. 113–142). Springer.

Vu, D. V., & Nguyen, N. C. (2019). *An assessment of vocabulary knowledge of Vietnamese EFL learners*. The 20th English in Southeast Asia Conference, National Institute of Education, Nanyang Technological University, Singapore.

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

West, M. (1953). *A general service list of English words*. Longman, Green.

Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing*, *2*(3), 1–17.

Yanagisawa A., Webb S. (2020). Measuring depth of vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 371–386). Routledge.

Ye, F. (2014). *Validity, reliability, and concordance of the Duolingo English Test.* School of Education, University of Pittsburgh. https://docplayer.net/2636113-Validity-reliability-and-concordance-ofthe-duolingo-english-test.html