IJOLAE Indonesian Journal on Learning and Advanced Education

http://journals.ums.ac.id/index.php/ijolae

Development of a Computerized-Adaptive Test to Measure English Vocabulary Size with IRT

Andhita Dessy Wulansari[1]✉, Dhinuk Puspita Kirana[2], Restu Mufanti[3]

[1,2]Faculty of Tarbiyah and Teacher Training, Institut Agama Islam Negeri Ponorogo, Indonesia

[3]Faculty of Teacher Training and Education, Universitas Muhammadiyah Ponorogo, Indonesia

[3]Faculty of Education, University of Technology Sydney, New South Wales, Australia

DOI: 10.23917/ijolae.v5i3.22953

Received: July 28th, 2023. Revised: September 8th, 2023. Accepted: September 19th, 2023.
Available Online: September 29th, 2023. Published Regularly: September, 2023.

Abstract

To speak English fluently like a native speaker, having a vocabulary of 20,000 words are required. In many universities in Indonesia, Vocabulary Size Test (VST) is still held with Paper-Based Test (PBT), so it is less effective and efficient. The aims of this study were 1) to develop Computerized-Adaptive Testing (CAT), which helps get students with a standardized English vocabulary size, and 2) to implement it in measuring vocabulary size. The approach in this study is Research and Development (R&D) with product development procedures following the steps of Borg and Gall. The findings of this study are, 1) the CAT program has Item Response Theory (IRT) parameters that can change according to the test takers' answers, and the test will stop if the estimated ability of the test takers is known (Standard Error/SE close to 0.01), and 2) with the application of the CAT program it can be seen that the total vocabulary of English Language and Teaching Department students is still <20,000. Based on these findings, this study implies that for students' vocabulary mastery to be on par with native speakers' abilities, the English Language and Teaching Department needs to evaluate the learning process and determine a particular program to increase student vocabulary.

Keywords: computerized-adaptive testing, item response theory, learning innovation process, vocabulary size test

Corresponding Author:
Andhita Dessy Wulansari, Faculty of Tarbiyah and Teacher Training, Institut Agama Islam Negeri Ponorogo, Indonesia
Email: andhita@iainponorogo.ac.id

1. Introduction

English is an international language that is widely used by people all over the world. Therefore, if one masters English, he will be easy to adapt (converse) with many people worldwide. With good mastery of English, people will quickly get a career at the international level (Ammigan, 2019; Krishnan et al., 2020; Sari & Aminatun, 2021; Torres & Alieto, 2019, Hikmat et al., 2023).

Learning is a dynamic activity and influenced by contextual factors and development (Ratih et al., 2021). In English, just like in other languages learning, four things must be mastered: listening, speaking, reading, and writing. The critical factors that significantly influence the 4 skills are mastery of vocabulary and grammar (Abera, 2020; Castillo-Cuesta, 2020). For English learners, the main thing that must be mastered first is vocabulary. The definition of *vocabulary* is a collection of words that are known and used by certain people in a language (Schmitt & Schmitt, 2020). Mastering much vocabulary is the primary way to compose sentences and speak fluently.

The large number of English learners who need more vocabulary will make it easier to express their ideas in English. Vocabulary is the building block of language (Arjmandi & Aladini, 2020; Fadi, 2019; Godfroid, 2019; Kohnke et al., 2019). Therefore, when someone wants to speak English well, they must have sufficient vocabulary to communicate fluently. English learners may need more time to study English in class. Therefore, they must increase their vocabulary by listening and reading as much as possible.

Every English learner may have thought about how many English words he or she needs to know to be proficient in English when carrying a dictionary during English class (Schmitt, 2013). A vocabulary size test is needed to find out the vocabulary that students have learned. The Vocabulary Size Test (VST) is designed to provide an estimate of vocabulary size for second and foreign language learners (Nation, 2012). Using the snowball throwing method can also increase mastery of english vocabulary (Kusumaningrum, et al., 2020).

To learn effectively, English learners need to know their goals about how much vocabulary they need to know or learn. According to Nation, there are 6,000-7,000 word families needed to listen to English and 8,000-9,000 word families needed to be able to read without a dictionary (Nation, 2012). In order to attain high standards and achieve native speaker proficiency, students need to know about 20,000-word families (not including proper names and transparently derived forms) (Goulden et al., 1990).

Many arguments state that knowing 2,000-word families is a sufficient measure for English learners to be able to listen and speak in daily activities (Hsu, 2020; Husnanissa, 2020; Noreillie, 2019). If foreign language learners have less than 2,000 words, they will have difficulty communicating in their daily activities. By mastering a minimum amount of vocabulary of 2,000 words, language learners can express their ideas in English in everyday conversations.

However, recent studies argue that more than 2000-word families are needed. Recently, vocabulary experts suggested at least 3000-word families. Recent research found that highly educated English speakers studying for a postgraduate degree indicated their English vocabulary size was around 6,000 to 7,000 for spoken text and 8,000 to 9,000-word families for unaided comprehension (Nation, 2006). Foreign language students at the university level must master a minimum of 3000 words to support their success in academic reading (Hartono & Prima, 2021; Setiawan & Wiedarti, 2020; Susanto et al., 2019).

Several theories have been proposed to show the relevance of English vocabulary in second language acquisition. There are many reasons to pay attention to terminology. First, learners often use dictionaries rather than grammar books, suggesting that vocabulary is a good measure of language proficiency (Krashen, 1989). Wilkins said that, without grammar, very little can be conveyed; without vocabulary, nothing can be conveyed (Wilkins, 1972). Vocabulary is one of the most essential aspects of language competence, both mother tongue and second language or foreign language. You need to know the words of a language to be able to communicate in it.

Having a large vocabulary can help students speak more and positively impact others. However, recent studies argue that a vocabulary size of 2000 needs to be considered sufficient. Recently, vocabulary experts proposed acquiring a minimum of 3000 words. Recent research has found that non-native speakers of highly educated, native

English speakers studying advanced degrees use the medium of English, indicating that their receptive English is about 6,000 to 7,000 for spoken text and 8,000 to 9,000-word families for unaided comprehension (Laufer & Nation, 1999; Nation, 2006).

The world of education in Indonesia also understands that English is positively correlated with increasing the competitiveness of Human Resources (HR). English is essential for Indonesians to master (Yosintha, 2020). Therefore, English language subject becomes integral part in the education curriculum in Indonesia. English has been studied as a foreign language from elementary school level to university. However, even though English has been prioritized since elementary school, Indonesian students still have difficulty mastering this language.

Kirana and Basthomi have conducted field research at the State Islamic Institute of Ponorogo (IAIN Ponorogo). Their study aims to analyze the vocabulary size and mastery level of students majoring in English at IAIN Ponorogo. This study uses a descriptive-quantitative research design. The research subjects were the first, third, and fifth-semester students majoring in English Language and Teaching at the English Language and Teaching Department, State Islamic Institute of Ponorogo. Three hundred and nineteen students participated in this study and were given a Paper-Based Test-based VST to measure their English vocabulary size. The findings from this study revealed that students only knew about 1,366-word families. The results are still below the threshold, as suggested by vocabulary experts. The findings also show that participants need a higher level of vocabulary mastery. The results of this study suggest that future research is needed to investigate vocabulary learning and instructional strategies that are effective in developing words that have a high frequency used in daily communication and academic words in lectures (Kirana & Basthomi, 2020).

Thus, it is crucial to conduct VST to determine whether they have enough vocabulary to appear in English without external support, especially for academics. So far, VST has been held in various places manually/classically, commonly known as the Paper Based Test. The same is true of what is happening at IAIN Ponorogo, one of the State Islamic Religious Colleges in Indonesia, which is currently organizing itself to become a research university whose main product is research. To achieve this, IAIN Ponorogo must start improving to increase the competitiveness of its human resources, one of which is by increasing the foreign language skills of lecturers and students, especially English.

The Paper Based Test (PBT) held at IAIN Ponorogo has been considered ineffective and inefficient. There are many problems/obstacles, from the preparation to the distribution of the test questions and the need for LJK scanning and scoring, so it requires extra cost, effort, and time. However, the current technological developments are so rapid that it is possible to carry out tests modernly using what is from now on called a computerized test.

However, in its application, there are not a few computerized tests that have a working principle similar to PBT, known as the Computer Based Test (CBT) (Pramono & Retnawati, 2020). The working principle of CBT is that the computer provides several questions that have been determined by the examiner. The test takers answer the questions, and the computer calculates the results of the test takers' work; the computer calculates the results of the test takers' work as to how many are true/false. From the correct/false answer information, an estimate of

the test taker's ability can be estimated. Of course, CBT has advantages over PBT in terms of time and effort, but from CBT, information is still not optimal because the test design is still the same for all test takers (you cannot adjust the test questions to the test taker's abilities) (Suhardi, 2020).

To overcome this, Computerized-Adaptive Testing was born. Computerized-Adaptive Testing is a development of CBT (Mizumoto et al., 2019; Ridwan et al., 2020; Ridwan et al., 2021). Computerized-Adaptive Testing (CAT) has the advantage of adapting the test items given to the test takers' abilities (Brunn et al., 2022). In addition, the advantages of CAT are that the test will stop if the estimated ability of the test takers can be known so that time is more efficient and more effective, in contrast to CBT, where the amount of time to complete the test and the number of questions have been predetermined (Bagus et al.; Egbe et al., 2023; Khoshsima et al., 2019; van Groen & Eggen, 2019). Test takers' ability can be estimated using Item Response Theory (IRT), likewise, with the estimation of item parameters. (Hermita et al., 2021) The magnitude of the estimated results of the item parameters/items will later be compared with the test takers' ability (Wulansari et al., 2019).

With the help of CAT, the examiner can find out the absolute vocabulary mastery of the test takers in a relatively short time. The estimated ability of the test takers is measured through the IRT-based CAT based on the pattern of their answers. This research was conducted as a form of support for efforts to increase the competitiveness of IAIN Ponorogo towards a research university. The development of a computerized test to measure English vocabulary size with a logistic model in IRT is expected to provide an alternative picture as a modern evaluation system at IAIN Ponorogo.

The development of computerized tests in IAIN Ponorogo can provide several contributions, such as:

a. For students, it can be used as a benchmark for students' English proficiency (vocabulary mastery) so that students know their initial abilities now so they can increase awareness to improve their English skills
b. For lecturers, it can provide information related to the results of data analysis and evaluation in detail regarding the English vocabulary size of students so that lecturers can plan language learning programs that are suitable for students.
c. For the IAIN Ponorogo, especially in the Department of English Language Education, the results of this study can provide benefits as a test standard for prospective IAIN Ponorogo students so that student input is standardized according to the English vocabulary size students have since the entrance exam test. Besides that, it can provide an alternative for institutions, employees, lecturers, and students as a modern evaluation system at IAIN Ponorogo.

Based on the description above, the development of CAT to measure English vocabulary size with the logistic model in this IRT at IAIN Ponorogo is essential. Because to speak English fluently, one must have sufficient vocabulary. For students of the Department of English Language and Teaching, being fluent in English is a skill that must be mastered. To find out how much vocabulary has been learned, VST is needed. So far, at IAIN Ponorogo, the measurement of English vocabulary testing is still done manually (PBT) with all its drawbacks and has yet to be touched by modern evaluation systems such as CAT, which can select items according to test takers' abilities. In addition, the test results will be known immediately

after the test stops, thereby saving more time for implementation.

The focus of the problems that will be discussed in this study is:

a. To produce a Computerized-Adaptive Test (CAT) program that can measure English vocabulary size with a logistic model in Item Response Theory (IRT);
b. To find out how the implementation of the CAT program developed in measuring English vocabulary size with the logistic model in IRT. The product development procedure follows the steps proposed by Borg and Gall, including the Pressman and Rolston model. The Borg and Gall model uses a waterfall flow at its development stage.

2. Method

As previously stated, the main objective of this research is to develop a computerized test to measure English vocabulary size with a logistic model in IRT. This development is expected to provide an alternative picture of a modern evaluation system at IAIN Ponorogo. The development is carried out by carrying out several stages of the procedure. The product development procedure follows the steps proposed by Borg and Gall, including the Pressman and Rolston model. The Borg and Gall model uses a waterfall flow at its development stage (Gall et al., 1996).

a. Stage I, Research, and Information Collection. The researcher conducted a literature review, made observations related to identifying the need for carrying out English vocabulary testing, what kind of test model was used, and what the problems encountered during the implementation of the test so that it was necessary to develop a test that was considered appropriate to quickly find out the test takers' English vocabulary mastery, accurate and safe.
b. Stage II is Planning. The researcher formulates an initial test model to solve the problems found in Stage I and explores the ease of administering the test.
c. Stage III, Develop a Preliminary Form of Product. Researchers design and develop test models that are fast, accurate, and secure. The test model developed is Computerized-Adaptive Testing (CAT) and prepares equipment to measure product success. At this stage, functional testing of the program is also carried out, or whether the program is functioning as desired by the researcher, revising and analyzing it until a product is produced that is ready to be tested. Functional testing is carried out through 2 stages: internal and external. The developers/researchers themselves carry out internal and external testing by IT experts who also understand the science of measurement/assessment.
d. Stage IV is Preliminary field testing. The researcher applies the test model to the actual situation. After the test model in Stage III is ready for use, the next activity is to try out the test model. The trial was carried out according to the researcher's scenario. The product was randomly applied to lecturers and English Language Education students at IAIN Ponorogo. In addition, observations were made to obtain information related to the results/impact of using the product. Product testing from the user side, namely:
    1) The first user, namely a lecturer at the Department of English Language and Teaching at IAIN Ponorogo with alpha testing, and

2) The last user, namely students at the Department of English Language and Teaching at IAIN Ponorogo with beta testing. The test is in the form of product verification and validation.

e. Stage V is the Main Product Revision. Researchers make product improvements after the product is applied or tested in the field. This improvement is done if results show that it could be more optimal during implementation.

f. Stage VI, Main Field Testing. Researchers use products that have been repaired in class. The product is applied in a class by re-involving English Language Education students at IAIN Ponorogo. In this case, data collection was again carried out with various instruments and analysis of the collected data.

g. Stage VII is Operational Product Revision. The researcher again made product improvements based on the trial results in Stage VI concerning the results of the analysis of the data collected.

h. Stage VIII is Operational Field Testing. Researchers reuse products that have undergone improvements in Stage VII in class. The product is applied in the classroom by involving the Department of English Language and Teaching students at IAIN Ponorogo. Furthermore, data were collected with various instruments and analyses of the collected data.

i. Stage IX is the Final Product Revision. Researchers make final improvements or product improvements.

j. Stage X is Dissemination and Implementation. Researchers report products refined in Stage IX from several previous trials. Once reported, the product is ready to be implemented on a broader scale, which is in this case.

a. Subject

The subjects in the trial in this study were lecturers and students of the Department of English Language and Teaching, IAIN Ponorogo, Academic Year 2022/2023. The selection of trial subjects was conducted using the Purposive Cluster Sampling Technique, namely the selection of trial samples based on specific considerations or objectives. Given that the products developed in this study are general and intended for all students, both those with high and low abilities, this trial was conducted on students of the English Language Education Department, IAIN Ponorogo, in the final semester to represent high-ability students and early semester students to represent students low ability.

b. Data Collection Technique

Data collection techniques for product development use:

1) Observations regarding the identification of needs;
2) A questionnaire regarding the completeness and accuracy of the CAT function; and
3) Documentation regarding the test's material, form, and model.

c. Data Analysis Technique

The data analysis technique used in this study is descriptive quantitative, and evaluative. This data analysis technique is applied because it is optional to test the hypothesis. To answer the formulation of the problem, the researcher tested the feasibility of the CAT program, which was used to evaluate the ability to master vocabulary size with a computerized test.

1) Conduct a qualitative analysis to determine the validity of the content analyzed by experts.
2) Analysis of the characteristics of the items using the Logistics Model 1 Parameter to determine the difficulty level of the items to be entered into the CAT question bank.
3) Descriptive analysis related to respondents' responses to the effectiveness of the CAT program performance in evaluating vocabulary size mastery skills with computerized tests, with the criteria:
   a) Operational or usage performance;
   b) Display performance;
   c) Relevance of test material, and usability.

3. Result and Discussion

The product developed in this study is a test program to measure English vocabulary size using computer assistance through the Item Response Theory (IRT) algorithm, logic, and statistics to select items/items based on the responses of the test takers according to the abilities of the test takers. Various calculations related to person parameters and item parameters here are carried out through the mechanism of a computer program. The development of a computerized test to measure English vocabulary size with the logistic model in the Item Response Theory (IRT) is expected to:

a. be able to be used as a benchmark for students' English proficiency (vocabulary mastery) so that students know their initial abilities now that they can increase awareness to improve their English skills
b. be able to provide information related to the results of data analysis and evaluation in detail regarding the English vocabulary size of students so that lecturers can plan language learning programs that are suitable for students.
c. be able to provide benefits as a test standard for prospective students of IAIN Ponorogo so that student input is standardized according to the English vocabulary size students have since the entrance exam test. Besides that, it can provide an alternative for institutions, employees, lecturers, and students as a modern evaluation system at IAIN Ponorogo.

To realize these expectations, in this study, an IRT-based CAT was developed based on the procedure described in the research methods section.

a. CAT Program Development to Measure English Vocabulary Size

The CAT Program Development Stage for Measuring English Vocabulary Size in **Table 1**, can be described as follows.

Table 1. CAT Program Development Stage

Stage	Description	Objective
I	*literature review*	Identification of needs for the implementation of English vocabulary testing.
II	*Planning*	Formulate an initial test model to solve the problems in the first stage and explore the ease of the test.
III	*Develop a Preliminary Form of Product*	Design systems and develop test models that are fast, accurate, and secure.
IV	*Preliminary field testing*	Carry out the application of the test model in real situations.

Stage	Description	Objective
V	*Main Produt Revision*	Make product improvements after the product is implemented or tested in the field.
VI	*Main Field Testing*	Test the improved product in class.
VII	*Operational Product Revision*	Again, to make product improvements based on the results of the trial in the sixth stage regarding the results of the data analysis.
VII	*Operational Field Testing*	Reuse products that have undergone improvement in the seventh stage in the classroom.
IX	*Final Product Revision*	Perform final repairs or product enhancements.
X	*Dissemination and Implementation*	Reports a product refined in the ninth stage from several previous trials.

1) Stage 1

In the first stage, the researcher conducted a literature review, made observations related to identifying the need for carrying out English vocabulary testing, what kind of test model was used, and what the problems encountered during the implementation of the test so that it was necessary to develop a test that was considered appropriate to find out the test takers' English vocabulary mastery quickly, accurately and safely. In the first stage, a needs analysis is carried out to identify the need for the implementation of English vocabulary testing, what kind of test model is used, and what are the problems encountered during the implementation of the test so that it is necessary to develop a test that is considered appropriate to find out the test takers' English vocabulary mastery quickly. Accurate and safe. Identification of needs is carried out by researchers using observation sheets.

2) Stage 2

Then in the second stage, Planning is carried out, where the researcher formulates an initial test model to solve the problems found in the first stage and explores the ease of carrying out the test. The product developed in this study is a test program (CAT) to measure English vocabulary size using computer assistance through the Item Response Theory (IRT) algorithm, logic, and statistics to select items/items based on the responses of test takers according to the participants' abilities. The tests of various calculations related to person parameters and item parameters here are carried out through the mechanism of a computer program. This IRT-based computer-assisted measurement system (CAT program) was developed by involving 3 interrelated parties. The first is that the lecturers developing this system are interested in measuring students' (test takers') vocabulary skills/mastery. The second is the students (test takers) who are interested in measuring their vocabulary skills/mastery in the development of this system. The third is the developer/researcher who makes the system so that it can be used by lecturers and students (test takers). In this case, the following third party in the CAT program is called admin.

3) Stage 3

The third stage is to Develop a Preliminary Form of Product. Researchers design systems and develop test models that are fast, accurate, and secure. The test model developed is Computerized-Adaptive Testing (CAT) and prepares equipment to measure

product success. Functional testing of the program is also carried out at this stage or whether the program is functioning as desired by the researcher, revising and analyzing it until a product is produced that is ready to be tested. Functional testing is carried out through 2 stages: internal and external. The developers/researchers carry out internal and external testing by IT experts who also understand measurement/assessment.

System design is applying various techniques and system principles (CAT program). The goal is to clearly define the components and system development process to make it easy to implement. The CAT system determines the level of difficulty of the items to be given to the test takers based on the responses of the test takers' answers to the items previously given. If the answer to the previous item is wrong, then the difficulty level of the question in the next item will be lower (down) than the difficulty level of the previous item. However, the CAT system will increase the difficulty level of the following item if the test taker's response to the previous item is correct. The data from the test results can then be used to determine the category of ability/vocabulary mastery of the test takers. Mastery of English vocabulary is used to classify the test takers' abilities. Mastery of English vocabulary is good if the value of $\Theta \geq 2$. Mastery of English vocabulary is sufficient if the value of $\Theta \geq 1$ to $\Theta < 2$. Mastery of English vocabulary is poor if the value of $\Theta < 1$.

After the system design is carried out, then the system development (CAT program) is carried out. At this stage, the program commands are written using a programming language and program commands for database access. The language used to write the program is PHP, and the database uses MySQL.

Implementation into program commands is written using standard rules. These commands are, from now on, referred to as scripts. The script is based on the design described in the design section of data flow diagrams, flowcharts, and program display designs. There are two things in the implementation section:

a) The stage of writing program scripts and
b) The display of program results.

A program script is a sequence of commands written using a specific language (computer language) arranged so a computer can understand, implement, and understand it. Based on the script or program code that has been made, the program is compiled into a single unit, run, and adjusted so that the display of the program results as expected.

The IRT-based CAT program was tested through 2 stages, namely, first testing by internal parties. After passing the internal test (with various revisions), external parties carried out testing. The developers/ researchers themselves carry out internal and external testing by IT experts who also understand psychometrics. In this study, external testing was carried out by Information Technology experts with expertise in the Assessment and Evaluation of Vocational Learning. This test is used to test the functional program or whether the program is functioning as desired by the researcher. This test intends to trace errors (debugging) of programs made. The sequence of internal tests is as follows:

a) Testing the accuracy of instructions or command sentences (syntax errors),
b) Testing the accuracy of the process when the program is run or executed (run time error/ running error),
c) Testing the accuracy of program results when executed (logic error), and

d) Product verification and validation (black box testing).

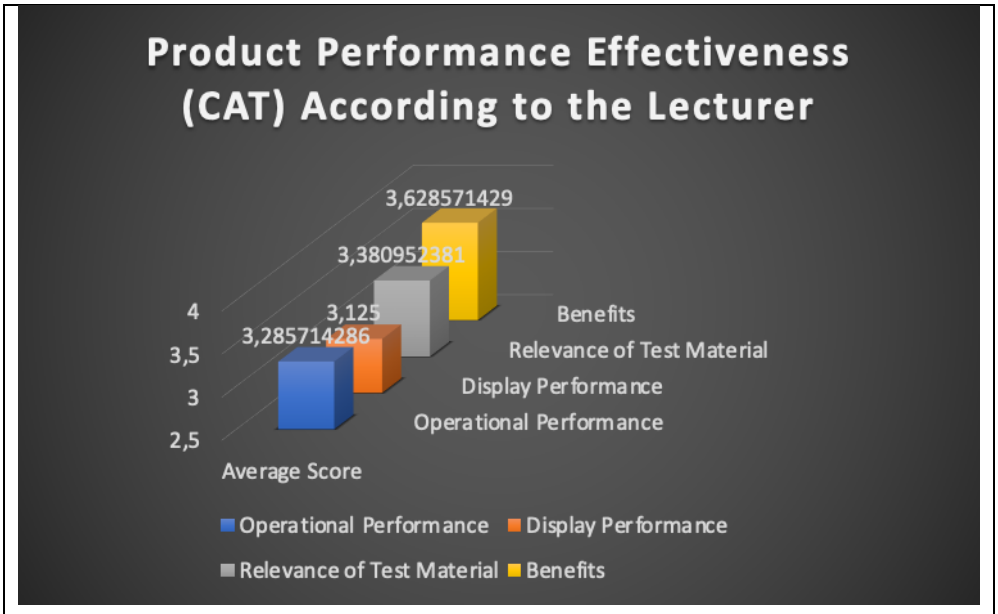
The examiner carries out the four sequences continuously until the program is complete.

4) Stage 4

The fourth stage is Preliminary field testing. The researcher applies the test model to the actual situation. After the test model in the third stage is ready for use, the next activity is to try out the test model. The trial was carried out according to the researcher's scenario. The product was randomly applied to lecturers and English Language Education students at IAIN Ponorogo. In addition, observations were made to obtain information related to the results/impact of using the product. Product testing from the user side, namely:

a) The first user, namely the lecturer at the Department of English Language and Teaching at IAIN Ponorogo with alpha testing;
b) The last user, namely students at the Department of English Language and Teaching at IAIN Ponorogo with beta testing. The test is in the form of product verification and validation.

Alpha testing is used to test the effectiveness of product performance (IRT-based CAT program) at an early stage. Alpha testing was conducted using a questionnaire given to 7 lecturers of the Department of English Language and Teaching at IAIN Ponorogo. The data was taken after the lecturer tried the CAT program and used it according to his authority.



Information:

- A score of 0 to a score of 1 means inferior.
- A score of > 1 to score 2 means less.
- A score of > 2 to score 3 means good.
- A score of > 3 to 4 means excellent.

Figure 1. Lecturer's Response Regarding the Effectiveness of CAT

Based on **Figure 1**, the lecturer's response score related to the effectiveness of the product after the lecturer carried out the testing (Alpha testing) of the developed CAT

for each aspect was in the range of > 3 to 4, which means that, in general, the lecturer considered that the effectiveness of the developed CAT was already excellent. For aspects of operational performance or use of the program, the average score is 3.457, which means that the lecturer considers this program excellent. The average score for the display performance aspect is 3.125, which means that the lecturer considers the appearance of the CAT program to be very good, but the score for this aspect is the lowest compared to other aspects.

Therefore the developer/researcher still needs to make more improvements so that it looks more attractive and easy to use. For the relevance aspect of the test material, the average score is 3.524, which means that the lecturer considers the relevance of the test material to be very good. For the usefulness aspect, the average score is 3.8, which means that the lecturer considers the usefulness of this program to be very good. The score for this aspect is the highest compared to other aspects. Lecturers and institutes. At a minimum, it can provide information related to the results of data analysis and detailed evaluation of the student's English vocabulary size so that the lecturer can plan language learning programs that are suitable for students and, in the future, can provide information regarding the results of detailed data analysis and evaluation of the English vocabulary size they have. Students so that lecturers can plan appropriate language learning programs for students.

Beta testing is carried out to determine the performance of the CAT program in predicting student abilities through the items worked. Previously, in the CAT program, item items (question bank) had been entered with an estimated difficulty level using Program R 4.2.1. Beta testing shows that the CAT program can correctly predict students' abilities (final theta) and can adequately convert the final theta into vocabulary size (vocabulary size).

5) Stage 5

Next, the fifth stage is the Main Product Revision. Researchers make product improvements after the product is applied or tested in the field. This improvement is done if the results show that it could be more optimal during implementation. Based on the lecturer's response scores related to product effectiveness after the lecturer tested (Alpha testing) the developed CAT, for each aspect, it was in the range of > 3 to 4, which means that, in general, the lecturer considered that the developed CAT had excellent effectiveness. The average score for the display performance aspect is 3.125, which means that the lecturer thinks that the appearance of the CAT program is excellent, but the score for this aspect is the lowest compared to other aspects. One suggestion related to the performance aspect of the lecturer representatives is:

> *"When you open it for the first time, the font size looks small, so you have to pinch the zoom to increase the font size, or was CAT originally for computer/laptop sizes? So the main screen is huge, and the text looks small."*

Besides that, some provide suggestions regarding the display menu, such as:

> "*The complete operational support menu might be able to add back, next, or number selection buttons/question numbers so that when someone presses, and a question is missed, it can still be done and not lost, and maybe the participant can add a display feature for numbers that have been done or haven't been done."*

Therefore the developer/researcher still needs to make further improvements according to the suggestions of the lecturers so that it looks more attractive and easier to use.

6) Stage 6

The sixth stage is Main Field Testing. Researchers use products that have been repaired in class. The product is implemented in the classroom by randomly re-involving the Department of English Language and Teaching students at IAIN Ponorogo. In this case, data collection was again carried out with various instruments and analysis of the collected data.

At this stage, where students evaluate the effectiveness of the CAT program, improvements still need to be made and require special attention because students are the primary users of the CAT program. With this improvement, the ability score produced by the CAT program is expected to be the actual student ability score without being influenced by the error factors of the CAT program being developed.

7) Stage 7

The seventh stage is the Operational Product Revision. The researcher again made product improvements based on the trial results in the sixth stage concerning the data analysis.

8) Stage 8

In the eighth stage, namely Operational Field Testing, researchers reuse products that have undergone improvements in the seventh stage in the classroom. The product is applied in a class by randomly involving the Department of English Language and Teaching students at IAIN Ponorogo. Furthermore, data were collected with various instruments and analyses of the collected data. Here the recap of the test results shows that the CAT program can perform its functions properly. Students again evaluate the effectiveness of the CAT program; improvements still need to be made so that the developed CAT program is more accessible for students. With this continuous improvement, the developed CAT program will be even better and more valuable.

9) Stage 9

The ninth stage is the Final Product Revision, where the researcher makes final or product improvements.

10) Stage 10

Next, the tenth stage is Dissemination and Implementation. At this last stage, the researcher reports a product that has been perfected in the ninth stage from several previous trials. Once reported, the product is ready to be implemented on a broader scale.

The last product produced is a Computerized-Adaptive Testing (CAT) program to measure English vocabulary size with a logistic model in Item Response Theory (IRT). With the following specifications:

a) The implementation of the adaptive system can be seen in the system's ability to change the item questions in each vocabulary proficiency/mastery test. Based on this, it can also be seen that the questions have IRT parameters changing according to the test takers' answers. In the resulting CAT program, the test will stop if the estimated ability of the test takers is known.
b) The calculation of the accuracy level of the CAT program developed here is based on the Standard Error (SE) estimate. The SE value will decrease as the number of question items submitted

to the test takers. The test can stop when the Standard Error (SE) is near 0.01.

c) The developed CAT program stores answer data and information on selecting questions based on parameter calculations. An analysis table of answers can be compiled, showing changes in test takers' abilities. Based on the table, it can be seen whether the item questions are better than the participants or vice versa, and the indicator is the test takers' final ability (theta).

d) The CAT program for measuring English vocabulary size with the logistic model in the resulting Item Response Theory (IRT) can provide an alternative picture as a modern evaluation system at IAIN Ponorogo.

b. Implementation of the CAT Program to Measure English Vocabulary Size

After the last repair or improvement of the product with several previous trials, the product is ready to be implemented on a broader scale. Semester 7 students of the Department of English Language and Teaching at IAIN Ponorogo are the subjects who will be involved as test takers, to be subjected to VST using the CAT produced in this study. For the final semester of the Department of English Language and Teaching students in semester 7, fluency in English is a skill that must be mastered. By being fluent in English, the competitiveness of human resources will increase so that opportunities to reach higher career levels will also increase, even reaching international career levels. To speak English fluently, one must have sufficient vocabulary. Therefore, here we will see how the ability/ mastery of the vocabulary of 7th-semester students majoring in English Language and Teaching at IAIN Ponorogo with the CAT program was produced in this study. The following is a summary report on the number of English vocabulary owned by Semester 7 students majoring in English Language and Teaching at IAIN Ponorogo in **Figure 2**.

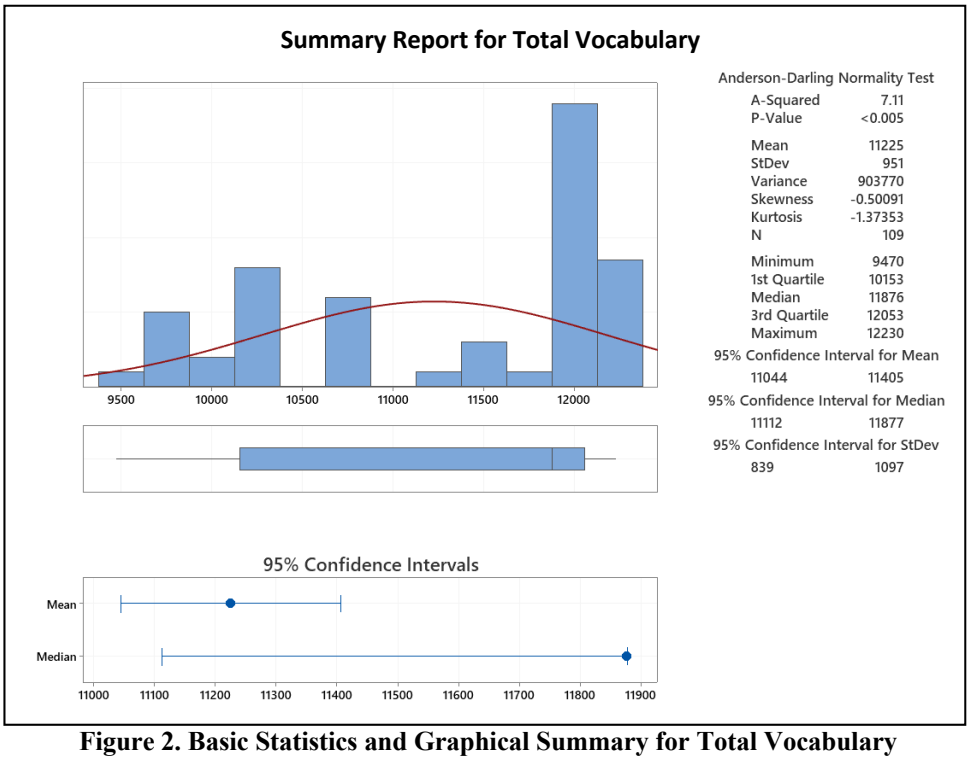


Figure 2. Basic Statistics and Graphical Summary for Total Vocabulary

The results of exploratory data on the number of English vocabulary owned by Semester 7 students majoring in English Language and Teaching at IAIN Ponorogo based on the Basic Statistics and Graphical Summary Figure for the Total Vocabulary below show that the average is 11,225 with the lower limit of the confidence interval being 11,044 and the upper limit being 11,405. However, of course, this average value cannot be used to conclude that all Semester 7 students majoring in English Language and Teaching at IAIN Ponorogo have moderate vocabulary mastery. The lowest number of vocabulary owned by Semester 7 students majoring in English Language and Teaching IAIN Ponorogo was 9470, while the highest number of vocabulary owned by Semester 7 students majoring in English Language and Teaching IAIN Ponorogo was 12,230.

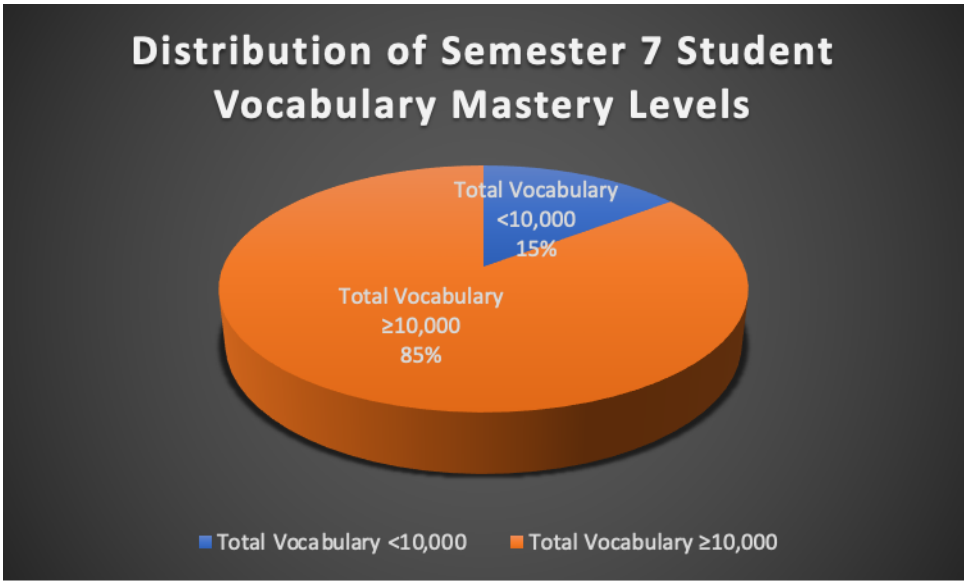


Figure 3. Pie Chart on Distribution of Semester 7th Semester Student Vocabulary Mastery Levels

The distribution of data on the number of vocabulary students in semester 7 of the Department of English Language and Teaching at IAIN Ponorogo in **Figure 3** has a left-slant of -0.50091. There are 7th-semester students of the Department of English Language and Teaching at IAIN Ponorogo with a low vocabulary count of < 10,000, but the number is less than those with a total vocabulary of ≥ 10,000.

Total vocabulary of semester 7 of the Department of English Language and Teaching students at IAIN Ponorogo (in percentage) based on the category of total vocabulary. Sixteen students (15%) had a total vocabulary of <10,000, and 93 students (85%) had a total vocabulary of ≤ 10,000. This phenomenon shows that there are still quite several 7th-semester students who need special attention from the Department of English Language and Teaching at IAIN Ponorogo because the average native speaker of English has a vocabulary of 20,000 words, so to be on par with native speaker abilities, semester 7 students of the Department of English Language and Teaching at IAIN Ponorogo still really need to improve their vocabulary mastery.

As shown in the Pie Chart for the Percentage of Vocabulary Amount above, the

distribution of the total vocabulary of the 7th-semester of the Department of English Language and Teaching students at IAIN Ponorogo has a reasonably high variance, namely 903,770. Therefore, in the future, it is necessary to carry out an analysis to find out the characteristics of the 7th-semester students of the Department of English Language and Teaching at IAIN Ponorogo that cause this high diversity so that in the future, it can be included as a policy to increase the vocabulary size of the Department of English Language and Teaching students at IAIN Ponorogo so that the competitiveness of human resources increases and can even compete internationally.

Based on the description above, the findings of this study are, 1) the CAT program has Item Response Theory (IRT) parameters that can change according to the test takers' answers, and the test will stop if the estimated ability of the test takers is known (Standard Error/SE close to 0.01), and 2) with the application of the CAT program it can be seen that the total vocabulary of English Language and Teaching Department students is still <20,000.

Based on these findings, this study implies that for students' vocabulary mastery to be on par with native speakers' abilities, the English Language and Teaching Department needs to evaluate the learning process and determine a particular program to increase student vocabulary.

Based on the development of the CAT program that has been carried out, several suggestions can be written down for further research.

1) System application can be done using more VST question packages, not limited to 2 packages as used in this study, so the tested questions are even more prosperous.
2) In the question bank for the CAT program, items on vocabulary questions have been entered with varying difficulty levels. Therefore, the developed CAT program needs to be tested on test takers whose abilities vary (high to low) to obtain more accurate person parameter results.
3) Parameter items in this study still use only 1 parameter (level of item difficulty). It is better for further research to use more parameters such as discriminant (2PL) or guessing (3PL) so that the data analysis process and discussion can be even more detailed.
4) In further research, it is recommended to keep using the most modern devices/platforms because the application on a newer system will make it easier for users to operate it on devices commonly owned by many people.

4. Conclusion

After developing the CAT program, several conclusions can be drawn The final product produced is a Computerized-Adaptive Testing (CAT) program to measure English vocabulary size with a logistic model in Item Response Theory (IRT). The implementation of the adaptive system can be seen in the system's ability to change the item questions in each vocabulary proficiency/mastery test. Based on this, it can also be seen that the questions have IRT parameters changing according to the test takers' answers. In the resulting CAT program, the test will stop if the estimated ability of the test takers is known. The calculation of the accuracy level of the CAT program developed here is based on the Standard Error (SE) estimate. The SE value will decrease as the number of question items submitted to the test takers. The test can stop when the Standard Error (SE) is near 0.01. The developed

CAT program stores answer data and information on selecting questions based on parameter calculations. An analysis table of answers can be compiled, showing changes in test takers' abilities. Based on the table, it can be seen whether the item questions are better than the participants or vice versa, and the indicator is the test takers' final ability (theta). The CAT program for measuring English vocabulary size with the logistic model in the resulting Item Response Theory (IRT) can provide an alternative picture as a modern evaluation system at IAIN Ponorogo.

Based on the application of the CAT program to semester 7 students of the Department of English Language and Teaching at IAIN Ponorogo, it can be seen that there are students who have a low vocabulary count of <10,000. However, the number is less than those with a total vocabulary of ≥10,000. This phenomenon shows that some 7th-semester students need special attention from the Department of English Language and Teaching at IAIN Ponorogo because actually, the average native English speaker has a vocabulary of 20,000 words, so to be on par with native speaker abilities, 7th-semester students of the Department of English Language and Teaching at IAIN Ponorogo still really needs to improve his vocabulary mastery.

5. Reference

Abera, M. (2020). The Impact of Cooperative Learning on Students' Paragraph Writing Skills: The Case of Third Year Health Informatics Stu-dents at University of Gondar. *Ethiopian Renaissance Journal of Social Sciences and the Humanities, 7*(2).

Ammigan, R. (2019). Institutional satisfaction and recommendation: What really matters to international students. *Journal of International Students, 9*(1), 262-281.

Arjmandi, M., & Aladini, F. (2020). Improving EFL Learners' Vocabulary Learning Through Short Story Oriented Strategy (SSOS). *Theory and Practice in Language Studies, 10*(7), 833-841.

Bagus, H. C., Tola, B., & Tjalla, A. Achievement Tests Administration using Computerized Adaptive Testing (CAT) with Constrain Response Time Item.

Brunn, G., Freise, F., & Doebler, P. (2022). Modeling a smooth course of learning and testing individual deviations from a global course. *Journal for educational research online, 14*(1), 89-121.

Castillo-Cuesta, L. (2020). Using digital games for enhancing EFL grammar and vocabulary in higher education. *International Journal of Emerging Technologies in Learning (iJET), 15*(20), 116-129.

Egbe, C. I., Agbo, P. A., Okwo, F. A., & Agbo, G. C. (2023). Students' Perception of Computer-Based Tests in the Use of English Programme in Nigerian Universities. *TechTrends*, 1-12.

Fadi, A.-K. (2019). The impact of vocabulary knowledge on the reading comprehension of Saudi EFL learners. *Journal of Language and Education, 5*(3 (19)), 24-34.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction*. Longman Publishing.

Godfroid, A. (2019). Sensitive Measures of Vocabulary Knowledge and Processing: Expanding Nation's Framework 1. In *The Routledge handbook of vocabulary studies* (pp. 433-453). Routledge.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied linguistics, 11*(4), 341-363.

Hartono, D. A., & Prima, S. A. B. (2021). The correlation between Indonesian university students' receptive vocabulary knowledge and their reading comprehension level. *Indonesian Journal of Applied Linguistics, 11*(1), 21-29.

Hermita, N., Putra, Z. H., Alim, J. A., Wijaya, T. T., Anggoro, S., & Diniya, D. (2021). Elementary Teachers' Perceptions on Genially Learning Media Using Item Response Theory (IRT). *Indonesian Journal on Learning and Advanced Education (IJOLAE), 4*(1), 1-20.

Hikmat, M. H., Santos, R. F., Suharyanto, S., Maudy, A. G., & Phommavongsa, K. (2022). Toward Continuous Innovation in Teaching: Reflective Practice on English Teaching of Indonesian and the Philippine Teachers. Indonesian Journal on Learning and Advanced Education (IJOLAE), 5(1), 45-60.

Hsu, W. (2020). Can TED talk transcripts serve as extensive reading material for mid-frequency vocabulary learning? *TEFLIN Journal, 31*(2), 181-203.

Husnanissa, A. (2020). *Measuring English Students'vocabulary Size At The First Semester Of The Eighth Grade Of Smpn 5 Bandar Lampung In The Academic Year Of 2017/2018* UIN Raden Intan Lampung].

Khoshsima, H., Toroujeni, S. M. H., Thompson, N., & Ebrahimi, M. R. (2019). Computer-based (CBT) vs. paper-based (PBT) testing: Mode effect, relationship between computer familiarity, attitudes, aversion and mode preference with CBT test scores in an asian private EFL context. *Teaching English with Technology, 19*(1), 86-101.

Kirana, D. P., & Basthomi, Y. (2020). Vocabulary size among different levels of university students. *Universal Journal of Educational Research, 8*(10), 4357-4364.

Kohnke, L., Zhang, R., & Zou, D. (2019). Using mobile vocabulary learning apps as aids to knowledge retention: Business vocabulary acquisition. *Journal of Asia TEFL, 16*(2), 683.

Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The modern language journal, 73*(4), 440-464.

Krishnan, I. A., Ching, H. S., Ramalingam, S., Maruthai, E., Kandasamy, P., De Mello, G., Munian, S., & Ling, W. W. (2020). Challenges of learning English in 21st century: Online vs. traditional during Covid-19. *Malaysian Journal of Social Sciences and Humanities (MJSSH), 5*(9), 1-15.

Kusumaningrum, S., Setyawati, I. G., Sutopo, A., Ratih, K., Badri, T. I. A., & Tawandorloh, K. A. (2019). Snowball Throwing: An English Learning Method to Improve Vocabulary Mastery and Psychomotor Ability. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, *2*(1), 10-19.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language testing, 16*(1), 33-51.

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.

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian modern language review, 63*(1), 59-82.

Nation, I. (2012). Measuring vocabulary size in an uncommonly taught language. International Conference on Language Proficiency Testing in the Less Commonly Taught Languages,

Noreillie, A.-S. (2019). It's all about words. Three empirical studies into the role of lexical knowledge and use in French listening and speaking tasks.

Pramono, A. J. B., & Retnawati, H. (2020). Implementation of CAT in Indonesia school: Current challenges & strategies. *Universal Journal of Educational Research, 8*(11), 5599-5609.

Ratih, K., Syah, M. F. J., Nurhidayat, N., Jarin, S., & Buckworth, J. (2021). Learning patterns during the disruptive situation in informal education: Parents' efforts and challenges in the adjustment of progressive learning. *Indonesian*

Journal on Learning and Advanced Education (IJOLAE), *3*(3), 180-193.

Ridwan, W., Wiranto, I., & Dako, R. (2020). Ability estimation in computerized adaptive test using mamdani fuzzy inference system. IOP Conference Series: Materials Science and Engineering,

Ridwan, W., Wiranto, I., & Dako, R. (2021). Computerized adaptive test based on sugeno fuzzy inference system. IOP Conference Series: Materials Science and Engineering,

Sari, S. N., & Aminatun, D. (2021). students'perception on the use of English movies to improve vocabulary mastery. *Journal of English language teaching and learning, 2*(1), 16-22.

Schmitt, N. (2013). *An introduction to applied linguistics*. Routledge.

Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching*. Cambridge university press.

Setiawan, M. R., & Wiedarti, P. (2020). The effectiveness of Quizlet application towards students' motivation in learning vocabulary. *Studies in English Language and Education, 7*(1), 83-95.

Suhardi, I. (2020). Comparison of Test Scores Using Paper and Computer Media as an Indicator of Computer Self-Efficacy and Test Anxiety in Facing Computer-Based-Testing. *Nusantara Journal of Social Sciences and Humanities*.

Susanto, A., Halim, F. A., & Thasimmim, S. N. (2019). Vocabulary learning strategies, vocabulary skills, and integrative motivation levels among university students. *Vocabulary Learning Strategies, Vocabulary Skills, and Integrative Motivation Levels among University Students, 8*(5C), 323-334.

Torres, J., & Alieto, E. (2019). English learning motivation and self-efficacy of Filipino senior high school students. *Asian EFL Journal, 22*(1), 51-72.

van Groen, M. M., & Eggen, T. J. (2019). Educational test approaches: The suitability of computer-based test types for assessment and evaluation in formative and summative contexts. *Journal of Applied Testing Technology, 21*(1), 12-24.

Wilkins, D. A. (1972). *Linguistics in language teaching* (Vol. 111). Edward Arnold London.

Wulansari, A. D., Kumaidi, K., & Hadi, S. (2019). Two Parameter Logistics Model with Lognormal Response Time for Computer-based Testing. *International Journal of Emerging Technologies in Learning (iJET), 14*(15). https://doi.org/10.3991/ijet.v14i15.10580

Yosintha, R. (2020). Indonesian students' attitudes towards EFL learning in response to industry 5.0. *Metathesis: Journal of English Language, Literature, and Teaching, 4*(2), 163-177.