

CHAPTER III

RESEARCH METHODOLOGY

A. Research Design

In examining the effect of using the busuu application in teaching to improve vocabulary skills, the method used is quantitative research method. The quantitative approach use in the research is the quasi-experimental method, to determine the impact of using the busuu application in teaching to improve vocabulary skills. Quantitative research is a research method that is widely required to use numbers, starting from data collection, interpretation of the data, and appearance of the results (Arikunto 2019).

Research design is a strategy for organizing research settings in order to obtain valid data. In conducting this research, the researcher use the quasy experiment method. This research is one of the most powerful research methodologies that can be used by researchers.

This research is an approach rooted in the positivist paradigm, which focuses on collecting numerical data to analyze cause-and-effect relationships between variables. This method uses structured designs, such as surveys and experiments, to ensure objectivity and reliability of research results. The resulting data is analyzed using statistical techniques with the aim of obtaining conclusions that can be

generalized to a wider population. This approach is suitable for answering specific and measurable research questions, and is often used to identify certain patterns, relationships or influences in various fields of study (Pilcher, N., & Cortazzi, M. 2024).

Quantitative approach is a research method based on the philosophy of positivism, used to research on certain populations or samples. Sampling techniques are generally carried out randomly, data collection uses research instruments, and data analysis is quantitative / statistical with the aim of testing predetermined hypotheses (Sugiyono 2019).

This quantitative approach was used to examine the effect of using Busuu app on students' vocabulary acquisition. Using a Non-Equivalent Control Group design, the study compared two groups of students, the experimental group using the Busuu app and the control group using traditional teaching methods. This approach allowed the analysis of differences in results between the two groups through pre-test and post-test.

This research design uses a Non-Equivalent Control Group Design scheme, which is part of a quasi-experiment. In this design, there are two groups with different conditions: the experimental group receives treatment using the Busuu app,

while the control group learns with traditional methods. The research design scheme is described in the following table:

Table 3. 1 Scheme non equivalent control group design

Class	Pretest	Treatment	Posttest
Experiment	O ₁	X	O ₂
Control	O ₃	-	O ₄

Note:

O₁ = Experimental group pretest result before given treatment

O₂ = Experimental group pretest result after given treatment

O₃ = Experimental group post test result before given treatment

O₄ = Experimental group post test result after given treatment

X = Treatment

- = No treatment

B. Location and Time of Research

This research was conducted at SMP Negeri 20 Kota Bengkulu, which was located at Jalan Depati Payung Negara, Suka Rami, Kecamatan Selebar, Kota Bengkulu, Bengkulu Province. The school had an excellent reputation, known as an educational institution that had produced many

outstanding students in various fields, making it the first choice for many parents who wanted their children to obtain quality education. Many students were interested in attending SMP Negeri 20 because of the outstanding academic and non-academic achievements produced by its students.

The school had five classes for grade VII which were spread across several spacious and comfortable classrooms. The facilities owned by SMP Negeri 20 Bengkulu City supported optimal teaching and learning activities. Some of the main facilities included a library that provided a variety of reference books, a laboratory equipped with equipment to support science learning, an administrative room that organized school administration, a UKS for student health, and a mosque that was used for religious activities. In addition, there was also a comfortable teachers' room and a principal's room used for coordination and decision-making related to academic and non-academic activities in the school.

C. Population and Sample

1. Population

Population is a generalized area consisting of objects or subjects that possess certain qualities and characteristics defined by the researcher for study and subsequent conclusions. This means that the population includes all individuals or research objects that are of concern to the

researcher, regardless of whether their numbers are large or small.

The group of people or entities that were the focus of this research was referred to as the "population of interest." This included individuals, pairs, groups, organizations, or similar entities (Hossan et al.,2023). The population can be people, events, or other things that can provide information relevant to the research objectives.

From some of the above opinions, it can be concluded that population is an object or subject that is in an area and fulfills certain conditions related to the problem or object of research. Researcher took the population in class VII students of SMP Negeri 20 Bengkulu City in the 2024/2025 school year.

The population of this research was the 7th Grade of SMP Negeri 20 Bengkulu City, totaling 165 students. In purposive sampling, researchers chose class VII D as the control class totaling 25 students and class VII F as the experimental class totaling 30 students.

Table 3. 2 Population of the Research

No	CLASS	FEMALE
1.	VII.A	30
2.	VII.B	30
3.	VII.C	25
4.	VII.D	31

5.	VII.E	25
6.	VII.F	31
Total		165

Source: SMPN 20 Bengkulu City

2. Sample

Sample is a part of the population selected using purposive sampling techniques, either probability or non-probability. The selection of the sample must align with the research objectives; probability sampling is used for generalization, while purposive sampling can be used for exploration (Campbell 2020). According to Upadhyay and Pradhan (2023), a sample is a representation of the target population, and its quantity must be adequate for the research results to be valid and generalizable to the population. This definition emphasizes not only the meaning of a sample as a part of the population but also the importance of sample size as a critical requirement in both quantitative and medical research. In this study, researcher used purposive sampling. Purposive sampling is a non-probability method in which units are selected intentionally because they have specific characteristics required for the research (Nikolopoulou, K.2022).

In purposive sampling, the two class groups must have the same or almost the same ability. Researcher took two classes which were divided into two groups. The first class will be the experimental class (VII.D) consisting of 31 students and

the second class will be the control class (VII.F) consisting of 31 students. The experimental class will treat using the busuu application as a learning medium and the control class will treat traditionally (traditional learning methods). The same things given to both classes are the same amount of time, and the same teacher. From this opinion, it can be concluded that the sample is part of the population that has the characteristics or conditions to be studied.

Table 3. 3 The sample of the Research

No	Group	Students	Classes
1.	Experiment	31	VII.D
2.	Control	31	VII.F
	Total	62	

Source: Managed Data, SMPN 7 Bengkulu City

D. Operational definition of Variables

1. Busuu Aplication

Busuu was founded in 2008 by Bernhard Niesner and Adrian Hilti during their MBA studies at IE Business School, Spain. The app offers language learning materials focused on practical vocabulary and everyday phrases, aligned with the CEFR. Its features include contextual vocabulary, repetition, multiple-choice and fill-in-the-blank tasks, plus writing and speaking exercises reviewed by native speakers. These features help users develop both receptive and productive

vocabulary (Busuu, 2025; McGraw-Hill Education, 2020). Vocabulary

Vocabulary as a set of words that a person knows or uses in a particular language. In the context of language learning, vocabulary includes the meaning, pronunciation, spelling, and usage of words in various communicative situations. Nation also highlights the importance of understanding aspects such as synonyms, antonyms, and collocations to enhance language proficiency (Paul Nation 2022).

E. The Technique of Collectinng Data

The data collection technique in this study was designed to measure the effect of using the Busuu application on students' vocabulary mastery at SMP Negeri 20 Bengkulu City. Data collection was conducted through quantitative methods to obtain objective results. This study conducted a pre-test and a post-test for all samples. The pre-test was given to students before they used the Busuu app to determine the initial level of vocabulary acquisition. After students completed the learning period using the app, a post-test was conducted to measure the improvement of their vocabulary skills. The test consisted of several types of questions, such as multiple choice to test vocabulary recognition, sentence completion to test vocabulary application, and matching synonyms or antonyms to measure understanding of word meaning. The results of the pre-test and post-test were used to analyze the

difference in students' vocabulary skills before and after using the app.

1. Test

a. Pre-test

Pre-test is an initial test given to participants before any treatment or research intervention is conducted. In this study, the pre-test is administered to students before they use the Busuu application. The format of the pre-test questions includes multiple choice to measure vocabulary recognition, sentence completion to test the application of vocabulary in context, and matching synonyms or antonyms to assess understanding of word meanings. The main objective of the pre-test is to determine the initial ability of students in vocabulary mastery, which can serve as a basis for comparison with the results obtained after the treatment (post-test).

Arikunto (2010) states that a pre-test is a test administered before the treatment to determine the initial ability or starting condition of the research subjects. The pre-test is used to measure the extent of participants' understanding before they receive the treatment. The pre-test is used to obtain initial data that will be compared with the post-test results after the treatment is given.

b. Post-test

Post-test is a test administered after the learning treatment or intervention is completed, which in this study involves the use of the Busuu application by students. The purpose of the post-test is to measure the improvement in students' vocabulary mastery after they have participated in learning using the application. By comparing the results of the post-test with those of the pre-test, the researcher can determine the effectiveness of the Busuu application in enhancing students' vocabulary skills.

According to Arikunto (2010), A post-test is a test given after a treatment has been conducted to see the extent of the results or changes that occur as a result of that intervention or treatment. In other words, the purpose of the post-test is to determine whether there has been an improvement in learning outcomes or skills of the students after undergoing a specific learning process or treatment.. This test aims to evaluate the effectiveness of the treatment that has been given, as well as compare the results with the pre-test. The function of the post-test is to assess the extent to which students master vocabulary after receiving treatment. The questions in the post-test have the same weight as the pre-test, making it easier to compare the results of the two tests.

The purpose of the post-test is to measure students' ability to expand their academic vocabulary after receiving the

treatment. Through analyzing the post-test scores, the researcher can evaluate whether there is a significant difference in student achievement between the experimental and control groups. Since the items in the pre-test and post-test were the same, the analysis of the post-test results provided a basis for measuring significant changes in students' vocabulary ability. After administering both tests, the researcher score the students' pre-test and post-test results and then arranged the scores in rank order for further analysis. The students would get a score of 100 if all questions will be answered correctly.

c. Treatment

Treatment in experimental research refers to the specific treatment or intervention given by researchers to study subjects to determine its effect on the variables being investigated. According to Sugiyono (2018), treatment is an action deliberately given to a research group so that its impact on learning outcomes can be observed. In this study, the treatment involves the use of the Busuu application to improve vocabulary mastery among seventh-grade students at SMP Negeri 20, Bengkulu City. The implementation of the treatment occurs after students take a pre-test and before they complete a post-test. The treatment content includes introducing the Busuu application to students, conducting application-based learning activities over a specific period, and practicing vocabulary through various available features,

such as word selection, sentence completion, listening to pronunciation, and matching word meanings. During the treatment, the researcher assists students to ensure the application is used according to the research objectives. After the treatment is completed, students are given a post-test to evaluate how much their vocabulary has improved after using the Busuu application.

d. Documentation

A document is any written or printed material that can serve as evidence, a source of information, or a reference in an activity. In this research, the documents include archives, grade lists, curricula, textbooks, student learning records, photos of activities, and teaching materials. Through document analysis, researchers can gain additional insights that further support the research findings, making the obtained data more comprehensive, credible, and accountable.

F. Research Instruments

Research instruments are tools used to collect, measure, and analyze data related to the research topic. The selection of appropriate research instruments is very important because it affects the validity and reliability of the data collected (Lasmana et al., 2024). The instrument of this research is a vocabulary test for students. This test is conducted to measure the level of vocabulary mastery of students before and after using the Mondly application. The researcher use the same

test questions in both classes (research subjects). The test is administer through multiple-choice questions. This test is design to determine the changes that occur in students' vocabulary mastery after they use the application, by comparing the results of the pre-test and post-test. The pre-test will be conducted before the students start using the application, while the post-test will take place after the learning period using the Mondly application to assess their vocabulary skill improvement.

The table below shows the assessment of vocabulary skills (Multiple Choice Test). The assessment includes five components: word meaning, synonyms and antonyms, usage in context, collocation, and word category used to determine classification for the assessment.

Table 3. 4 The Scoring Criteria

Number of Correct Answers	Scale Score	Vocabulary Assessment Description
18–20 correct	(Very Good)	Uses vocabulary and idioms almost like a native speaker.
15–17 correct	(Good)	Occasionally uses inappropriate terms or needs to repeat ideas due to limited lexical resources.
12–14 correct	(Fair)	Frequently uses incorrect words; conversation is somewhat limited due to insufficient vocabulary.

9–11 correct	(Poor)	Errors in word and vocabulary usage make comprehension fairly difficult.
6–8 correct	(Very Poor)	Extreme limitations in vocabulary, making communication almost impossible.
0–5 correct	(Not Proficient)	Long pauses while searching for the desired word meanings.

Source : (J.B Heaton 1998)

G. Data Analysis Technique

1. Validity of the Test

The validity test in this study will use to test the validity of the instrument. Validity means that the researcher checks for the accuracy of the findings by employing certain procedures (Creswell, 2012). The validity test is carried out to measure whether the statements contained in the research instrument are valid or not. Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. (Gay et al., 2012). To find the validity of an item, the researcher analyzes the test results using computer statistics, namely SPSS. In addition, the test results consist of 100 questions containing four vocabulary indicators such as nouns, verbs, adjectives, and adverbs. This test was conduct to determine whether there will a significant effect of using the busuu application on students' vocabulary.

Table 3. 5 Test Instrumen Validity Criteria

No	Cronbach's Alpha	Validity Level
1.	$\alpha \geq 0.90$	Excellent
2.	$0.80 \leq \alpha < 0.90$	Good
3.	$0.70 \leq \alpha < 0.80$	Acceptable
4.	$0.60 \leq \alpha < 0.70$	Questionable
5.	$0.50 \leq \alpha < 0.60$	Poor
6.	$\alpha < 0.50$	Unacceptable

Source: Olsen, Espen, et al , 2024

Furthermore, the author also analyzed the test results using SPSS version 29. Below are the results of the validity test instrument data analysis:

Table 3. 6 Instrument validity

Item	R table	R count	Result
Item 1	0,367	0,412	valid
Item 2	0,367	0,518	Valid
Item 3	0,367	0,482	Valid
Item 4	0,367	0,358	Unvalid
Item 5	0,367	0,441	Valid
Item 6	0,367	0,120	Unvalid
Item 7	0,367	0,463	Valid
Item 8	0,367	0,440	Valid
Item 9	0,367	0,-409	Unvalid
Item 10	0,367	0,190	Unvalid
Item 11	0,367	0,514	Valid
Item 12	0,367	0,-140	Unvalid
Item 13	0,367	0,221	Unvalid
Item 14	0,367	0,448	Valid
Item 15	0,367	0,-117	Unvalid
Item 16	0,367	0,421	Valid
Item 17	0,367	0,472	Valid
Item 18	0,367	0,586	Valid

Item19	0,367	0,474	Valid
Item 20	0,367	0,292	Unvalid
Item 21	0,367	0,111	Unvalid
Item 22	0,367	0,394	Valid
Item 23	0,367	0,664	Valid
Item 24	0,367	0,429	Valid
Item 25	0,367	0,030	Unvalid
Item 26	0,367	0,400	Valid
Item 27	0,367	0,460	Valid
Item 28	0,367	0,098	Unvalid
Item 29	0,367	0,582	Valid
Item 30	0,367	0,417	Valid
Item 31	0,367	0,495	Valid
Item 32	0,367	0,-019	Unvalid
Item 33	0,367	0,482	Valid
Item 34	0,367	0,-011	Unvalid
Item 35	0,367	0,427	Valid
Item 36	0,367	0,-213	Unvalid
Item 37	0,367	0,429	Valid
Item 38	0,367	0,-300	Unvalid
Item 39	0,367	0,-242	Unvalid
Item 40	0,367	0,405	Valid
Item 41	0,367	0,117	Unvalid
Item 42	0,367	0,405	Valid
Item 43	0,367	0,397	Valid
Item 44	0,367	0,410	Valid
Item 45	0,367	0,469	Valid
Item 46	0,367	0,483	Valid
Item 47	0,367	0,418	Valid
Item 48	0,367	0,478	Valid
Item 49	0,367	0,420	Valid
Item 50	0,367	0,418	Valid
Item 51	0,367	0,432	Valid
Item 52	0,367	0,-342	Unvalid
Item 53	0,367	0,470	Valid
Item 54	0,367	0,-156	Unvalid
Item 55	0,367	0,417	Valid

Item 56	0,367	0,519	Valid
Item 57	0,367	0,443	Valid
Item 58	0,367	0,402	Valid
Item 59	0,367	0,393	Valid
Item 60	0,367	-0,221	Invalid

The validity test was conducted at SMPN 16 Kota Bengkulu using 60 multiple-choice questions, where items were declared valid if the calculated r-value exceeded the r-table value of 0.364; ultimately, 40 items were valid and 20 were invalid, indicating that the instrument was overall of good quality and suitable for use in the research.

Table 3. 7 Instrument taraf signifikan

Item	Taraf SIG	SIG 2-Tailed	Result
Item 1	0,05%	0,021	valid
Item 2	0,05%	0,003	Valid
Item 3	0,05%	0,006	Valid
Item 4	0,05%	0,048	Invalid
Item 5	0,05%	0,013	Valid
Item 6	0,05%	0,520	Invalid
Item 7	0,05%	0,009	Valid
Item 8	0,05%	0,013	Valid
Item 9	0,05%	0,022	Invalid
Item 10	0,05%	0,305	Invalid
Item 11	0,05%	0,003	Valid
Item 12	0,05%	0,451	Invalid
Item 13	0,05%	0,232	Invalid
Item 14	0,05%	0,012	Valid
Item 15	0,05%	0,530	Invalid
Item 16	0,05%	0,018	Valid
Item 17	0,05%	0,007	Valid
Item 18	0,05%	0,001	Valid

Item19	0,05%	0,007	Valid
Item 20	0,05%	0,112	Invalid
Item 21	0,05%	0,554	Invalid
Item 22	0,05%	0,028	Valid
Item 23	0,05%	0,000	Valid
Item 24	0,05%	0,016	Valid
Item 25	0,05%	0,873	Invalid
Item 26	0,05%	0,026	Valid
Item 27	0,05%	0,009	Valid
Item 28	0,05%	0,600	Invalid
Item 29	0,05%	0,001	Valid
Item 30	0,05%	0,019	Valid
Item 31	0,05%	0,005	Valid
Item 32	0,05%	0,918	Invalid
Item 33	0,05%	0,006	Valid
Item 34	0,05%	0,952	Invalid
Item 35	0,05%	0,017	Valid
Item 36	0,05%	0,250	Invalid
Item 37	0,05%	0,016	Valid
Item 38	0,05%	0,101	Invalid
Item 39	0,05%	0,189	Invalid
Item 40	0,05%	0,024	Valid
Item 41	0,05%	0,532	Invalid
Item 42	0,05%	0,024	Valid
Item 43	0,05%	0,027	Valid
Item 44	0,05%	0,022	Valid
Item 45	0,05%	0,008	Valid
Item 46	0,05%	0,006	Valid
Item 47	0,05%	0,019	Valid
Item 48	0,05%	0,006	Valid
Item 49	0,05%	0,019	Valid
Item 50	0,05%	0,019	Valid
Item 51	0,05%	0,015	Valid
Item 52	0,05%	0,059	Invalid
Item 53	0,05%	0,008	Valid
Item 54	0,05%	0,402	Invalid
Item 55	0,05%	0,020	Valid

Item 56	0,05%	0,003	Valid
Item 57	0,05%	0,012	Valid
Item 58	0,05%	0,025	Valid
Item 59	0,05%	0,029	Valid
Item 60	0,05%	0,231	Invalid

Based on the results of the validity test conducted on 60 research instrument items using a significance level of 5% ($\alpha=0.05$), it was concluded that 40 items were declared valid as they had Sig. (2-tailed) values below 0.05, while the remaining 20 items were declared invalid because their significance values exceeded the threshold. Overall, the instrument achieved a success rate of 63.3% and was deemed feasible for use in field research, provided that only the valid items are included.

Table 3. 8 Indikator Of Vocabulary Test

No	Question	Indicator
1	My bedroom is ____ the second floor.	Students are able to use the correct preposition of place to show location.
2	The kitchen is ____ the dining room.	Students are able to choose the correct preposition of place based on room position.
3	We watch TV in the ____.	Students are able to identify the correct room vocabulary based on its function.

4	The car is parked in the ____.	Students are able to identify the correct place vocabulary related to parking a car.
5	The flowers are in the ____.	Students are able to identify the correct place for plants.
6	The dining room is used for ____.	Students are able to identify the function of a room.
7	The lamp is ____ the table.	Students are able to use the correct preposition of place for object position.
8	My mother keeps vegetables in the ____.	Students are able to identify the correct storage place for food.
9	We eat using a fork and ____.	Students are able to identify eating utensils vocabulary.
10	The clothes are in the ____.	Students are able to identify the correct place to keep clothes.
11	I sleep on the ____.	Students are able to identify bedroom object vocabulary.
12	A broom is used for ____.	Students are able to identify the function of cleaning tools.
13	The window is ____ the wall.	Students are able to use the correct preposition of place for building parts.
14	The mirror is ____ the sink.	Students are able to use the correct preposition of place in bathroom context.

15	The broom is ____ the door.	Students are able to choose the correct preposition of place for object location.
16	I usually ____ my bed every morning.	Students are able to use the correct daily activity verb.
17	My sister ____ the dishes after dinner.	Students are able to identify the correct household chore verb.
18	I ____ the plants every morning.	Students are able to identify the correct verb for plant care activities.
19	My brother ____ the floor with a mop.	Students are able to identify the correct cleaning activity verb.
20	My mother ____ the laundry every Saturday.	Students are able to use the correct verb related to washing clothes.
21	She ____ her room every weekend.	Students are able to identify the correct verb for cleaning a room.
22	I ____ the table before dinner.	Students are able to identify the correct verb for preparing meals.
23	I ____ my toys after playing.	Students are able to use the correct verb for tidying up.

24	She uses a cloth to ____ the table.	Students are able to identify the correct verb for cleaning furniture.
25	My sister ____ the books neatly on the shelf.	Students are able to identify the correct verb for arranging objects.
26	I ____ the dishes after lunch.	Students are able to identify the correct verb for washing dishes.
27	A trash bin is used to ____ garbage.	Students are able to identify the function of a trash bin.
28	My mom ____ all the dirty clothes today.	Students are able to use the correct verb related to laundry activities.
29	I ____ breakfast at 6 a.m. every day.	Students are able to identify the correct verb for daily eating activities.
30	My sister ____ to school at 7 a.m.	Students are able to identify the correct verb for going to school.
31	My father ____ coffee every morning.	Students are able to identify the correct verb for drinking activities.
32	They ____ homework in the evening.	Students are able to identify the correct verb for studying activities.

33	She ____ TV after cleaning her room.	Students are able to identify the correct leisure activity verb.
34	The light is ____. Please turn it on.	Students are able to identify the correct adjective describing condition.
35	Please ____ the door before you sleep.	Students are able to use the correct imperative verb.
36	I ____ the pillow on my bed.	Students are able to use the correct verb for placing objects.
37	My sister ____ the bathroom every Friday.	Students are able to identify the correct verb for cleaning the bathroom.
38	The floor is dirty. Please ____ it.	Students are able to identify the correct verb for cleaning the floor.
39	I ____ my shoes near the door.	Students are able to identify the correct verb for putting objects.
40	There is dust on the shelf. Please ____ it.	Students are able to identify the correct verb for cleaning dust.

2. Reliability of the test

Reliability test is an examination conducted to determine the extent to which a research instrument can provide consistent results when used repeatedly on the same subjects

or respondents. Sugiyono (2016), reliability is the level of trust in a research instrument that shows the consistency of measurement results over time or between different measurements. A reliable instrument will produce the same or similar data when used under the same conditions. Testing is done using the Cronbach Alpha technique. Where an instrument can be said to be reliable if it has a specification coefficient or alpha of: (a) 0.8, very good.

A questionnaire is said to be reliable or reliable if a person's answer to a statement is consistent or stable over time. Researcher took the reliability score using the Statistical Package for Social Science (SPSS) program. After that the researcher collect the answer sheets and gave scores based on the correct answers. Then the researchers analyz all items using SPSS to find valid items that could be used in the research sample at SMP Negeri 20 Bengkulu City.

Table 3. 9 Test Instrumen Reliability Criteria

Cronbach's Alpha	Reliability Level
0,00-0,20	Lack of Realible
0,20-0,40	Somewhat Realible
0,40-0,60	Fairly Realible
0,60-0,80	Realible
0,80-1,00	Highly Realible

Source: Menon, Vikas, et al, 2025.

Table 3. 10 Instrumen Reability

Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
0,828	0,820	60

Based on the results of the validity test conducted on 60 research instrument items using a significance level of 5% ($\alpha=0.05$), it was concluded that 40 items were declared valid as they had Sig. (2-tailed) values below 0.05, while the remaining 22 items were declared invalid because their significance values exceeded the threshold. Overall, the instrument achieved a success rate of 63.3% and was deemed feasible for use in field research, provided that only the valid items were included in the final data collection.

Furthermore, the reliability test results for the instrument items showed a Cronbach's Alpha value of 0.803, indicating that the instrument fell into the high reliability category. Given that the obtained alpha coefficient exceeded the minimum standard threshold of 0.60, the instrument was declared highly reliable and possessed strong internal consistency in measuring the research variables. The combination of these validity and reliability results confirmed that the measurement tool met the scientific requirements to produce accurate and consistent data.

3. Difficulty Level Test

Item difficulty indicates how “easy” or “difficult” a question is based on the proportion of participants who answered the question correctly. The higher the proportion of students who answered correctly, the easier the question is conversely, the lower the proportion of correct answers, the more difficult the question (Rashwan, N. I., Aref, S. R, et al, (2024).

$$P = \frac{Rh+Ri}{Nh+Ni} \times 100$$

Note

P = Level of difficulty in percent

Nh = Number of test participants in the high score group

Rh = Number of correct answers in the high score group

Ni = Number of test participants in the low score group

Ri = Number of correct answers in the low score group

$$P = \frac{263}{8} + \frac{92}{8} \times 100\%$$

$$P = 44,3\%$$

table 3. 11
Interpretation of difficult question categories

Question category	Interpretation of results
Relatively easy	71%-----100%
Moderate	41%-----70%
Classified as difficult	21%-----40%

Source : Haladyna 2004

Based on the results of the analysis, the difficulty level was found to be 44,3%, which falls into the moderate category. This means that, in general, the test items have an appropriate and balanced level of difficulty. The questions are neither too easy nor too difficult for the students, making them suitable for objectively measuring students' abilities.

Thus, it can be concluded that the test items meet the criteria of good quality in terms of difficulty level and can be used as an effective evaluation tool to measure students' understanding of the material taught.

4. Discriminating Power Test

The discrimination power evaluates an item's ability to distinguish between high-performing and low-performing students. Good test items will yield a higher proportion of correct answers among high-performing students compared to low-performing students (Makhija, Hirsh, et al 2024). This analysis is crucial in a study examining the impact of the

Busuu application on students' vocabulary improvement, as it helps determine which tasks effectively differentiate skill levels.

$$DP = \frac{U-L}{N}$$

Note

Dp = Discrimination index of the question

U = Number of correct answers from the high-scoring group

L = Number of correct answers from the low-scoring group

N = Number of students in each group

$$DP = \frac{263-92}{8}$$

$$DP = 21,3\%$$

Table 3. 12
Intrepretasi of Discrimination test

<i>Discriminating Power</i>	<i>Interpretation of Result</i>
0,71 ≤ DP ≤ 1,00	Very good
0,41 ≤ DP ≤ 0,70	Good
If +0,21 < DP < 0,40	Sufficiently qualified
If 0 < DP ≤ 0,20	weak

Source: (Gronlund,1985)

Based on the calculation results, the discriminating power (DP) was found to be 21.3%. When compared to the criteria in Table 3.13, this value falls within the range of 0.21–0.40, which is categorized as sufficiently qualified.

This result indicates that the test item analyzed has a sufficiently qualified of discriminating power, meaning it can reasonably differentiate between high-achieving and

low-achieving students. Although the level of discrimination is not yet strong, the test item is still considered effective enough to measure differences in students' understanding and performance.

5. Normality Test

Normality testing is an important assumption in many statistical analyses, because many parametric tests depend on data being normally distributed. The first step in conducting a normality test is to collect data on English vocabulary mastery, which can take the form of pre-test and post-test scores, assessments, or vocabulary measurements taken before and after the implementation of the Mondly application. After the researcher collects those data, they can proceed with statistical tests to assess normality. The normal distribution is the most important distribution in statistical data analysis. Most statistical tests were developed based on the assumption that the outcome variable/model residuals are normally distributed. (Yang & Berdine, 2021)

There are several common normality tests, including the Shapiro–Wilk test, the Kolmogorov–Smirnov test, and the Anderson–Darling test. According to Biu, Nwakuya & Wonu (2019), the Kolmogorov–Smirnov test is more appropriate for large samples or samples larger than 40. Meanwhile, the Shapiro–Wilk test is more accurate when used for small samples. Furthermore, according to Razali & Wah (2011), the Anderson–Darling test is used as a normality or

goodness-of-fit test for quantitative variables and can be used to assess the normality of various types of data distributions such as normal, lognormal, gamma, Weibull, and logistic distributions. These tests are designed to compare the study sample data with the normal distribution. The testing criteria for the normality test are:

- a. Sig (P value) < 0,05 indicates that the data are not normally distributed.
- b. Sig (P value) > 0,05 indicates that the data are normally distributed.

6. Homogeneity Test

Homogeneity test is a statistical technique used to determine whether several populations have the same variance or not. This test is important in research because it is a prerequisite for performing parametric statistical analysis, such as Independent Sample t-Test and ANOVA (Analysis of Variance). One of the main assumptions in variance analysis is that the variances of the populations being compared must be equal (homogeneous).

The equality test of two variances is used to evaluate whether the distribution of data from two or more groups is homogeneous. This test is done by comparing variances between groups. If the two or more groups have the same variance, then the data is considered homogeneous and the homogeneity test does not need to be done again. However,

if the variances are different, then further steps need to be taken to overcome the inhomogeneity.

It is important to note that the homogeneity test can only be performed if the data being compared is in a normal distribution. This test is also conducted to ensure that differences found in parametric statistical analyses, such as t-test, ANOVA, or ANCOVA, truly reflect differences between the groups being compared, rather than coming from differences in the variances of the groups themselves. In other words, the homogeneity of variance test helps ensure that the results of statistical tests are valid and not biased due to differences in the underlying data structure.

As a first step, the homogeneity test is essential in comparing two or more groups to ensure that the statistical analysis is conducted fairly. By conducting this test, researchers can verify that differences in results found come from differences in variables between groups, rather than from inhomogeneity within the groups being compared. This makes the homogeneity test an essential component in ensuring the validity of research results involving parametric statistical analysis.

7. T-test

The T test in SPSS will be used to analyze data and evaluate whether there is a significant difference between two or more groups related to the dependent variable. In this analysis, the formula used aims to measure the extent to

which independent factors affect the dependent variable, assuming other independent variables are held constant.

T-Test Procedure:

- a. If the significance level $< \alpha$ (0.05)

In this case, the test results show that the independent variables individually have a significant effect on the dependent variable. This means that the differences found in the tested groups are not due to chance, but due to the factors of the independent variables that actually affect the dependent variable.

- b. If the significance level $> \alpha$ (0.05)

If the test results show that the significance level is greater than 0.05, then the independent variable individually has no significant effect on the dependent variable. That is, there is no difference large enough or significant enough to conclude that the independent variable affects the dependent variable.

8. F-Test / Analysis of Variance (ANOVA / ANCOVA)

The F-test is used when comparing the means of more than two groups, or when testing several variables together, including in ANCOVA (Analysis of Covariance) models where there are control variables (covariates) that can affect the dependent variable. Assumptions that must be met include homogeneity of variance between groups, normality of

residuals, independence of observations, and in ANCOVA also the assumption that the relationship between covariates and dependent variables is uniform across all groups (Schwarz, Wolf 2025). After the analysis is complete, the F value and p-value are obtained. If $p < \alpha$ (e.g., 0.05) $\rightarrow$ it is concluded that there is a difference in the means between groups after controlling for the control variable. If $p > \alpha$, then there is no significant difference even though the means appear to be different. It is also important to report the related effect size (e.g., eta squared or partial eta squared) and confidence intervals, so that the results are more complete and the interpretation of the magnitude of the effect between groups is clear.

9. Effect Size

Effect size is a quantitative measure that shows how large the effect or strength of the relationship between two variables in a study is—not just whether or not there is an effect. For example, when comparing two groups or measuring the relationship between variables, effect size provides a clear picture of the extent to which the independent variable explains the variability of the dependent variable. In current research, it is highly recommended to report the effect size along with the p-value and confidence interval, so that readers can understand not only that the effect is significant, but also how large the effect is and whether it is practically

useful in a real-world context (Tagliaferri, Scott D., et al, 2024).

Table 3. 13 Value of the effect size

No.	Value	Effect size
1.	0,00-0,1,95	Very weak effect
2.	0,20-0,395	Weak effect
3.	0,40-0,595	Modest effect
4.	0,60-0,749	Strong effect
5.	0,80-1,00	Very strong effect

Source: Tagliaferri, Scott D., et al, 2024).

10. Statistical Hypotesis

A statistical hypothesis consists of two statements: H_0 (null hypothesis), which states that there is no effect or no difference, and H_1 (alternative hypothesis), which states that there is an effect or a difference (Emmert-Streib, Frank, 2024). After performing the analysis (t-test, ANOVA/ANCOVA, etc.), we observe the p-value and compare it with the α that was set before the analysis (often 0.05).

1. If $p < \alpha$, then the null hypothesis is rejected and the alternative hypothesis is accepted;
2. If $p > \alpha$, then the null hypothesis fails to be rejected (there is not enough evidence that an effect exists).