International Journal of Advanced Social Sciences and Education (IJASSE)
Vol. 3, No. 4, 2025: 205-216

Validity and Reliability Test of Test Instrument Development to Measure Students' Comprehension Ability in Class V of MIN 11 Banda Aceh

Intan Afriati[1]*, Qashdina Alya Thaha[2], Cut Nikmatul Ulya[3], Hikmah[4], Nurul Husna[5], Win Khamsyi Sirda[6]
Ar-Raniry State Islamic University, Banda Aceh
Corresponding Author: Intan Afriati intan.afriati@ar-raniry.ac.id

A R T I C L E I N F O

Keywords: Validity, Reliability, Evaluation

Received:19, June
Revised: 20, July
Accepted: 30, August

A B S T R A C T

This study aims to determine the validity and reliability of a developed test instrument. A test is considered effective and appropriate for use when it fulfills the essential criteria of validity and reliability. To achieve this objective, the research employed the ADDIE model, a well-known instructional design framework consisting of five stages: Analysis, Design, Development, Implementation, and Evaluation. Each stage played a critical role in ensuring that the test met the required standards for educational assessment. The instruments used in this research included validation sheets and a set of test items specifically designed for this study. The participants in the study were fifth-grade students from class V-1 at MIN 11 Banda Aceh. Based on the analysis of the data, the results indicated that the test developed through this process was generally categorized as valid and reliable. This suggests that the test is appropriate to be administered to students as a tool for assessment. However, a closer examination of the students' responses revealed that three of the test items did not meet the criteria for validity. These items were identified as invalid and may require revision or removal to enhance the overall quality and accuracy of the test instrument.

DOI: https://doi.org/10.59890/ijasse.v3i4.107
ISSN-E: 2988-2133
https://dmimultitechpublisher.my.id/index.php/ijasse

INTRODUCTION

Learning evaluation is a crucial element in the educational process. The purpose of learning evaluation is to establish the objectives to be achieved in the educational process. Therefore, evaluation serves as a benchmark for how well students understand the lessons delivered by teachers during the teaching and learning process. Therefore, evaluation significantly influences the success of the educational process. However, during field practice in educational activities, the accuracy and precision of measurement are less than optimal, partly due to the weakness of the testing instruments. One solution to this problem is the use of standardized tests. The idea of standardized tests is to ensure that the tests meet the requirements of validity and reliability. In Indonesia, this concept is implemented in national exams and various tests designed by teachers themselves. A test is a set of tasks that must be done or several questions that must be answered by students to measure their level of understanding and mastery of the required material and by certain teaching objectives. Miller, as quoted by Sukiman, describes it in more detail, namely that a test is a formal assessment instrument used to assess students' cognitive abilities in a subject as well as to collect quantitative information about students' psychomotor abilities (physical skills) and affective characteristics (such as attitudes, emotions, interests, and values). Tests generally include a series of questions, statements, or tasks arranged for a student or group of students. Tests in various subjects are very urgent in their position and function in seeing the extent of students' abilities and understanding of the material that has been taught. A test is a research instrument that presents a series of questions to be answered by students. Multiple-choice questions are a form of test where the teacher presents questions with several answer options. From these options, students choose the most appropriate answer that best fits the question. Facing this type of test encourages students to think critically in selecting the most appropriate answer. Therefore, the test must be designed to be of high quality, namely one that demonstrates validity and reliability in its items.

LITERATURE REVIEW

In education, validity and reliability testing are crucial steps that should not be overlooked. Testing the validity and reliability of a test ensures that the measurement instrument used is truly valid and reliable. This way, research results will be more reliable and can be used to support quality decision-making in the field of education. By analyzing various test items, weaknesses or areas for improvement can be identified. If any questions are not yet valid and reliable, adjustments can be made to achieve a higher level of quality. This allows for the development and refinement of test items to ensure more valid and reliable measurement of student achievement.

METHODOLOGY

This study used the research and development (R&D) method with the ADDIE development model. According to Sugiyono, the research and development method is a research method used to produce a specific product

and test its effectiveness. The research and development method with the ADDIE model consists of five stages: analysis, design, development, implementation, and evaluation. The ADDIE model is based on an effective and efficient systems approach and an interactive process between teachers, students, and the environment.

RESULT AND DISCUSSION

Validity Test

1. Understanding Validity Tests

Test validity, or the validity of a measuring instrument, is "the extent to which the test measures what it is intended to measure." Test validity essentially refers to the degree to which a test measures its function, or the degree of accuracy of a test's measurements. To assess the validity of a measuring instrument, namely the extent to which the measuring instrument measures what it is intended to measure. The following are several validity criteria according to experts, consisting of:

a) According to Gronlund and Linn, validity is the accuracy of the interpretation made from the results of measurement or evaluation.
b) According to Anastasi, validity is the accuracy of measuring a construct, concerning "What the test measures and how well it does".
c) According to Arikunto, validity is a condition that describes the level of the instrument in question that can measure what is to be measured.
d) According to Sukadji, validity is the degree to which a test measures what it should measure.
e) According to Azwar, validity is the extent to which a measuring instrument is accurate and precise in carrying out its function.

From the definitions above, it can be seen that validity is a measure that indicates the validity or authenticity of an instrument. Therefore, validity testing refers to the extent to which an instrument performs its function.

2. Use of Instrument Validity

The following are several uses of validity, consisting of:

1) Avoid unclear questions.
2) Eliminate words that are too foreign or words that arouse suspicion.
3) Correcting unclear questions.
4) Add necessary items or remove items deemed irrelevant.

3. Types of Validity

In general, there are three approaches to examining the validity of a measuring instrument, namely 1) content validity, 2) construct validity, and 3) criterion validity (Suryabrata, 2005).

a) Content validity: The content validity of an instrument is the extent to which the items in the instrument represent the components in the overall content area of the object to be measured (representation aspect) and the extent to which the items reflect the behavioral characteristics to be measured (relevance aspect). Content validity is validity that focuses on what elements are in the measurement (Coaley, 2010), so rational analysis is the main process carried out in content validity analysis. The content

validity of an instrument is determined by matching whether the items in the instrument have represented the components to be measured or not. This shows that the level of content validity of an instrument depends to some extent on the individual assessment of the assessor.

b) Criteria validity: Criterion validity is the linking of a measuring instrument with another measuring instrument as a criterion, whether the measuring instrument can be explained by its correlation results with its criteria based on existing theory (Devellis, 2003). Criterion-related validity shows the extent to which the developed test scores are related to independent external criteria that are believed to be able to describe the behavior or characteristics being investigated.

c) Construct validity: Construct validity is a description that shows the extent to which a measuring instrument produces results that are in accordance with theory (Azwar, 2005). Construct validity is validity that indicates the extent to which an instrument reveals a trait or theoretical construct that it is intended to measure. Testing construct validity is an ongoing process in line with the development of the concept of the trait to be measured. The concept of construct validity is very useful in tests that measure traits that do not have external criteria. Therefore, the construct validation procedure begins with an identification and definition of the variables to be measured and expressed in the form of a logical construct based on the theory regarding the variables. From this theory, a practical consequence is drawn regarding the measurement results under certain conditions, and this consequence will be tested. If the results are in accordance with expectations, the instrument is considered to have good construct validity.

Reliability Test

1. Understanding Reliability

Reliability is a series of measurements or a series of measuring instruments that have consistency when the measurements carried out with the measuring instrument are carried out repeatedly. Test reliability is the level of consistency of a test, namely the extent to which a test can be trusted to produce consistent scores, relatively unchanged even when tested in different situations.

Reliability is the degree to which a test consistently measures its intended target. Reliability is expressed numerically, usually as a coefficient. A high coefficient indicates high reliability. Reliability is also considered the consistency of measurement or observation results when a fact or reality is measured or observed repeatedly at different times. The instrument and the method of measurement or observation both play important roles simultaneously. Therefore, it can be concluded that reliability is a test for measuring or observing something that is the object of measurement.

Specifically, the concept of reliability refers to the consistency of the scores on the items in your questionnaire, so a reliability test essentially tests the accuracy of the measurement scales of a research instrument. Therefore, the primary objective of a research instrument reliability test is to measure the consistency of the measuring instrument used by quantitative researchers. In this context, researchers want to determine whether the measurement results

are accurate across the same sample at different times. In other words, a research instrument, such as a questionnaire, is considered reliable if it can provide consistent scores for each measurement. Therefore, the measurement tool (the statements/questions) continues to provide consistent measurement results at different times.

2. Implementation of Tests to Determine Reliability

In carrying out tests to determine reliability in order to estimate the reliability of an assessment tool (test and non-test), there are several methods that are most often used, namely single test, retest (test re-test), and parallel (equivalent).

a. Single Test: A single test is a test consisting of one instrument (one set) administered to a group of subjects in a single administration. Therefore, the results of this test contain only one set of data in the form of test scores.
b. Retest (test re-test): A retest is a test consisting of a set of tests administered to a group of subjects twice. Its reliability is calculated by correlating the results of the first test with the second test. (The retest method is the use of the same test twice on the same number of test participants.) The retest method is used to avoid compiling two series of tests. In using this technique or method, the tester only has one series of tests, but is administered twice. Because there is only one test and is administered twice, this method can be called the single-test double trial method. Then the results of the two tests are calculated for correlation.
c. Parallel (Equivalent): The parallel (equivalent) method involves two tests that have the same objective, difficulty level, and structure, but different items. The test is administered only once, but with two instruments, on the same respondents, at the same time, and with different instruments.

The stages of developing the ADDIE model in this research are as follows:

a) Analysis

- The purpose of testing and developing test instruments is to produce instruments that can be used to measure student abilities. Conducting tests requires valid and reliable instruments.
- Curriculum Analysis, based on the results of observations at MIN 11 Banda Aceh, the Arabic language curriculum used in the school is the 2013 KMA 183 curriculum in 2019.

Table 1. Curriculum Analysis

Basic competencies	Indicators of Competence Achievement	Learning objectives
Understanding the social function, text structure, linguistic elements (sounds, words, and meaning) of texts related to the theme of في المعمل	Able to answer general questions about في المعمل	Students are expected to be able to use vocabulary and expressions related to the theme of في المعملusing Arabic language rules.

- Analysis of Learning Materials, identifying learning materials that include the theme of المعمل with sub-materials of mufradat.
- Student Analysis, looking at student characteristics based on their abilities and understanding.

b) Design Stage

In this stage, what the researcher does includes:

a. Determining the Test Form: Before conducting validity and reliability tests on the test instruments, researchers designed the test format to be administered to students. The test format used was a multiple-choice test. This test consists of a series of questions with several answer options. Students select the most appropriate answer that best fits the question.

b. Compiling the Test Grid: Next, the researcher compiled the test outline in tabular form. The test outline can be seen in the following table:

Table 2. Compiling the Test Grid

Basic competencies	**Aspect**	**Question Indicator**	**Question Items**	**Number of Questions**
Understand the social function, text structure, linguistic elements (sounds, words, and meaning) of texts related to the theme of المعمل	Observing the picture	Given a picture, students choose the appropriate answer (vocabulary) based on the picture.	1-4	4
	Meaning of underlined words in Indonesian	Presented with simple Arabic sentences with underlined words, students are asked to choose the meaning that corresponds to the underlined words.	5-7	3
	Translate the vocabulary into Arabic	The meaning of the vocabulary (in Indonesian) is presented. Students are asked to translate the vocabulary into Arabic by choosing one of the available answers.	8-10	3

Development Stage

During the development phase, the test instrument was developed in the form of a multiple-choice test. After the questions were compiled, they were validated by an Arabic language teacher. Validators were given a grid containing multiple-choice questions, answer keys, and a questionnaire to assess each item on a scale of 1 to 5. The questionnaire assessed several aspects, including the suitability of the questions to the indicators, clarity of the scope of the questions and the expected answers, clear instructions for completing the questions, clarity of the images presented, and the use of easy-to-understand sentences.

Table 3. Development Stage

Item	1	2	3	4	5
Compliance of questions with indicators					
Clarity of the scope of the questions and the expected answers					
Clear instructions for working on questions					
Clarity of the images presented					
Use of easy-to-understand sentences					

Information:
1 = Very Poor 4 = Good
2 = Less 5 = Very Good
3 = Enough

The questions that will be distributed to students and validated by the Arabic language teacher are as follows:
Answer the questions by marking (x) the correct answer!
Look at the following picture! (For numbers 1-4)

1. مَا هَذَا ؟

أ. شَ اشَ ةٌ ج. حَ اسُ وْ بٌ
ب. طَ ابِ عَ ةٌ د. سَ اعَ ةٌ

2. مَا هَذِهِ ؟

أ. طَ ابِ عَ ةٌ ج. وَحْدَةُ النِّظَامِ
ب. بَابٌ د. لَوْحَةُ الْمَفَاتِيْحِ

3. مَا هَذِهِ ؟

أ. طَ ابِ عَ ةٌ ج. الفَأْرَةُ
ب. مَعْمَلٌ د. شَ اشَ ةٌ

4. مَا هَذِهِ ؟

أ. الفَأْرَةُ ج. سَمَّاعَةٌ
ب. مَعْمَلٌ د. وَحْدَةُ النِّظَامِ

What is the meaning of the underlined word in the following sentence? (For numbers 5-7)

5. هَذَا مَعْمَلُ الْ حَ اسُ وْ بِ ، فِيْهِ حَاسُوبٌ، وَلِكُلِّ حَاسُوبٍ شَاشَةٌ ، وَلَوْحَةُ مَفَاتِيْحِ ، وَفَ أ رَةٌ .

أ. Screen ج. Library

ب. Table Printer and

6. هَذَا مَعْمَلُ الْ حَ اسُ وْ بِ ، فِيْهِ حَاسُوبٌ، وَلِكُلِّ حَاسُوبٍ شَاشَةٌ، وَ لَوْحَةُ مَفَاتِيْحِ ، وَفَ أ رَةٌ .

أ. CPU ج. Computer

ب. Keyboard Screen and

7. هَذَا مَعْمَلُ الْ حَ اسُ وْ بِ ، فِيْهِ حَاسُوبٌ، وَلِكُلِّ حَاسُوبٍ شَاشَةٌ، وَلَوْحَةُ مَفَاتِيْحِ ، وَفَ أ رَةٌ .

أ. Mouse ج. CPU

ب. Keyboard Speaker and

Translate the following vocabulary into correct Arabic! (For numbers 8-10)

8. The Arabic word for **computer laboratory** is .....

أ. الفَأْرَةُ ج. مَعْمَلُ الْ حَ اسُ وْ بِ

ب. مَكْتَبَةٌ د. وَحْدَةُ النِّظَامِ

9. The Arabic word **for CPU** is...

أ. الفَأْرَةُ ج. وَحْدَةُ النِّظَامِ

ب. بَابٌ د. مَعْمَلٌ

10. The Arabic word for **internet** is...

أ. مَعْمَلٌ ج. God willing

ب. شَ اشَ ةٌ د. جَرِيْدَةٌ

After the validator assessed the aspects of the assessment questionnaire, the researcher then conducted a content validity analysis of each item assessed by the validator. The content validity test used in this study was Aiken's V test. Aiken's V test formula is as follows:

$$V=\frac{\sum s}{n\,(c-1)}$$

Information:
V = validity value
∑s = The sum of the scores given by the assessor for each item, after subtracting the lowest score
N = number of assessors
C = highest score given
S = score given by the assessor for each item

The V index ranges from 0 to 1. According to Aiken's expert validation criteria, if 0.80 < V ≤ 1.00 is considered very high. 0.60 < V ≤ 0.80 is considered high. 0.40 < V ≤ 0.60 is considered quite high. 0.20 < V ≤ 0.40 is considered low. 0.00 < V ≤ 0.20 is considered very low (Aiken, 1985). The table below is the result of 'Aiken's V content validity test using Excel:

Table 4. The Aiken's V Content Validity

Item	∑s	n(c-1)	V
Item 1	6	12	0.5
Point 2	9	12	0.75
Point 3	9	12	0.75
Item 4	12	12	1
Item 5	9	12	0.75

Item	∑s	n(c-1)	V	Note
Items 1-5	45	60	0.75	Tall

Based on Aiken's V index calculation above, a value of 0.75 can be interpreted as having a fairly good content validity index. Because V = 0.75 is in the range of $0.60 < V \leq 0.80$, it is in the high category. Therefore, overall, the test instrument is considered valid and can be administered to students.

Implementation Stage

After the test instrument was validated by the validator, a pilot test was conducted on a group of students. The pilot test was conducted on 19 students in grade V-1. At this stage, the researcher provided the test instrument to the students. They were asked to complete the test by selecting the most appropriate answer. Before the students began the test, the researcher provided instructions on how to complete the questions. Afterward, the students were able to complete the questions that had been distributed.

Evaluation Stage

The final stage involves analyzing the results of the student's trial work. This analysis aims to determine the quality of the test instrument or its suitability through validity and reliability tests.

a. Validity Test: Researchers conducted a test on a previously developed test in class V-1, consisting of 19 students. This validity test will determine the empirical validity of each item. The basis for decision-making in this test is by comparing the Sig. (2-tailed) value with a probability of 0.05. Therefore, if the Sig. (2-tailed) value is <0.05 and the Pearson Correlation is positive, then the questionnaire item is valid. The results of the validity test can be seen in the following table:

Table 5. The Results of the Validity Test

Question Items	Sig. (2-tailed)	Probability	Category
1	0.002	0.05	V
2	0.943	0.05	TV
3	0,000	0.05	V
4	0.022	0.05	V
5	0.396	0.05	TV
6	0.008	0.05	V
7	0,000	0.05	V
8	0.231	0.05	TV
9	0.021	0.05	V
10	0,000	0.05	V

The table of validity test results with SPSS is as follows:

		Total Score
Question 1	Pearson Correlation	.655 **
	Sig. (2-tailed)	.002
	N	19
Question 2	Pearson Correlation	-.018
	Sig. (2-tailed)	.943
	N	19
Question 3	Pearson Correlation	.835 **
	Sig. (2-tailed)	.000
	N	19
Question 4	Pearson Correlation	.520 *
	Sig. (2-tailed)	.022
	N	19
Question 5	Pearson Correlation	.207
	Sig. (2-tailed)	.396
	N	19
Question 6	Pearson Correlation	.587 **
	Sig. (2-tailed)	.008
	N	19
Question 7	Pearson Correlation	.835 **
	Sig. (2-tailed)	.000
	N	19
Question_8	Pearson Correlation	-.289
	Sig. (2-tailed)	.231
	N	19
Question 9	Pearson Correlation	.524 *
	Sig. (2-tailed)	.021
	N	19
Question 10	Pearson Correlation	.722 **
	Sig. (2-tailed)	.000
	N	19

Figure 1. Validity Test Results with SPSS

Based on the results of the analysis of 10 multiple-choice questions, it is known that 7 questions are declared valid, namely questions number 1, 3, 4, 6, 7, 9, and 10. The 7 questions are declared valid because, as the basis for decision making in this test, the Sig. (2-tailed) value is <0.05. Meanwhile, 3 questions are declared invalid, namely questions number 2, 5, and 8. The 3 questions are invalid because the Sig. (2-tailed) value is >0.05. As explained above, seven questions were found to be valid, demonstrating the questions' effectiveness in measuring students' understanding of the Al-Ma'malu material. Valid questions demonstrate that the questions accurately reflect students' understanding of the established learning objectives. However, it must be acknowledged that three

items were declared invalid. This highlights the importance of improving the evaluation instrument. In-depth analysis of the invalid items is crucial for evaluating the extent to which they meet validity criteria and for identifying the factors contributing to the invalidity. This process provides valuable insights for teachers in refining the item design, thereby continuously improving the quality of the evaluation instruments used in schools.

Reliability Test

In addition to examining the validity of the test items, reliability testing was also conducted to ensure good results. To determine the reliability of the research instrument, the researchers performed calculations using the Cronbach's Alpha formula. The following are the results of the reliability test:

Table 6. Reliability Statistics

Reliability Statistics	
Cronbach's Alpha	Number of Questions
0.696	10

The table of reliability test results with SPSS is as follows:

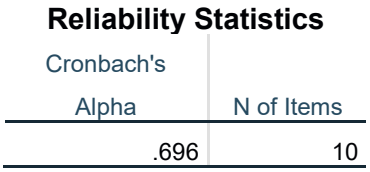

Figure 2. Reliability Test Results with SPSS

As for test reliability decision making according to Triton (2016), the alpha value measure which measures the level of reliability of an instrument, is as follows:

1) A Cronbach's Alpha value of 0.00 to 0.20 means less reliable.
2) Cronbach's Alpha value of 0.21 to 0.40 means slightly reliable.
3) Cronbach's Alpha value of 0.41 to 0.60 means it is quite reliable.
4) Cronbach's Alpha value of 0.61 to 0.80 means reliable.
5) Cronbach's Alpha value of 0.81 to 1.00 means it is very reliable.

An acceptable reliability score is >0.6 (Heale & Twycross, 2015). The Cronbach's Alpha reliability test results table above yields a result of 0.696. Therefore, 0.696 > 0.6. As a basis for decision-making based on reliability testing, it can be concluded that the test items are reliable or consistent. Therefore, in the context of learning evaluation, reliability can be used as a reference for future updates and adjustments. This information can be used by teachers in designing strategies and considering the use of the instrument in subsequent evaluations.

CONCLUSIONS AND RECOMMENDATIONS

Test validity, or the validity of a measuring instrument, is "the extent to which the test measures what it is intended to measure." There are three approaches in examining validity, namely 1) content validity, 2) construct validity, and 3) criterion validity. Meanwhile, reliability is a series of measurements that have consistency when measurements carried out with the measuring instrument are carried out repeated. There are the most widely used methods in determining test reliability, namely, single tests, retests, and

parallels. Based on the validity test results, the research results indicate that the developed test is categorized as valid, and the test can be used and tested on students. However, the validity test of the test items based on student answers found 7 valid items and 3 invalid items. Meanwhile, the reliability test of the test items can be concluded that the test items are reliable or consistent.

FURTHER STUDY

This research still has limitations, so further research is still needed on this topic, "Validity and Reliability Test of Test Instrument Development to Measure Students' Comprehension Ability in Class V of MIN 11 Banda Aceh".

REFERENCES

Anshari, Muhammad Isa, Rodiah Nasution, Muhammad Irsyad, Alifia Zuhriatul Alifa, dan Indah Aminatus Zuhriyah. "Analisis Validitas dan Reliabilitas Butir Soal Sumatif Akhir Semester Ganjil Mata Pelajaran PAI." Edukatif: Jurnal Ilmu Pendidikan 6, no. 1 (2024): 964–75.

Budiastuti, Dyah, dan Agustinus Bandur. Validitas dan Reliabilitas Penelitian. Metode Penelitian Pendidikan Matematika. Jakarta: Mitra Wacana Media, 2018.

Hamzah B. Uno, dan Satria Koni. Assessment Pembelajaran. Jakarta: Bumi Aksara, 2024.

Haryanto. Evaluasi Pembelajaran (Konsep dan Manajemen). Yogyakarta: UNY Press, 2020.

Hayati, Salma, dan Lailatussaadah Lailatussaadah. "Validitas dan Reliabilitas Instrumen Pengetahuan Pembelajaran Aktif, Kreatif dan Menyenangkan (Pakem) Menggunakan Model Rasch." Jurnal Ilmiah Didaktika: Media Ilmiah Pendidikan dan Pengajaran 16, no. 2 (2016): 169.

Safitri, Islamiani, Dewi Lestarani, Rahmah Dwi Nor Wita Imtikhanah, Nur Rahmi Akbarini, Meida Wulan Sari, Ilyas Muh. Fitrah, Taufik Rizki Sista, Ikhsan Dwi Setyono, dan Rizki Fitria Setyaningtyas. Teori Pengukuran dan Evaluasi. CV. Ruang Tentor, 2024.

Suseno, Endro, dan Purwo Susongko. Mengukur Validitas Tes. Jawa: Pemeral edukreatif, 2021.

Widodo, Slamet, Festy Ladyani, La Ode Asrianto, Rusdi, Khairunnisa, Sri Maria Puji Lestari, Dian Rachma Wijayanti, et al. Buku Ajar Metode Penelitian. Cv Science Techno Direct. Pangkal Pinang: CV Science Techno Direct, 2023.