Journal of Philosophy and Research in Islamic Mathematics Education
J-PRIMa
Vol. 1, No. 1, 2025, pp 29 – 36 e-ISSN:
https://doi.org/10.33474/
online at: https://riset.unisma.ac.id/index.php/j-prima/article/view/24565/

A Comprehensive Evaluation of Mathematical Communication: Validity and Reliability Evidence in Assessment Instruments

Sofiatul Ilmi[1*]

[1]MA Ahmad Yani, Jabung, Malang
*Correspondence: sofiatulilmi@gmail.com

Received: 22-09-2025; Revised: 28-10-2025; Accepted: 18-11-2025

Abstract. *This study aims to provide a comprehensive evaluation of a mathematical communication assessment instrument by examining its validity and reliability through both expert judgment and empirical testing. High-quality instruments are essential to ensure that the data collected accurately represent students' abilities and align with the intended constructs of mathematical communication. The validation process in this study involved mathematics education experts who reviewed the content relevance, clarity, and alignment of each item with the established indicators. In addition, a small-scale trial was conducted with students to obtain empirical evidence of item performance. The trial results indicated strong item–total correlations that exceeded the critical values in the validity table, demonstrating that each item adequately measured the intended aspects of mathematical communication. Reliability analysis was carried out using the test–retest method supported by Cronbach's Alpha to assess the internal consistency and stability of the instrument. The obtained coefficient of 0.812 reflects a high level of reliability, indicating that the instrument consistently yields dependable results across multiple administrations. These findings collectively confirm that the instrument possesses both strong validity and high reliability, providing solid evidence of its technical quality. Overall, this assessment instrument can be confidently used to evaluate students' mathematical communication skills in various instructional contexts. The validated instrument offers a robust foundation for further analysis in mathematics learning, supports the evaluation of communication-focused pedagogical strategies, and contributes to the development of evidence-based practices aimed at improving students' mathematical understanding and expression.*

Keywords: *Mathematical Communication, Reliability, Validity.*

How to cite: Ilmi, S. A Comprehensive Evaluation of Mathematical Communication: Validity and Reliability Evidence in Assessment Instruments. J-PRIMA, 1(1), 29 – 36. Retrieved from https://riset.unisma.ac.id/index.php/j-prima/article/view/24565/

Introduction

Mathematics learning is an educational process that requires the application of specific methods. Suherman (2003:56) emphasizes that mathematics has three main functions, namely as a tool, a way of thinking, and a source of knowledge. These functions are used as the basis for setting mathematics learning objectives in schools. The objectives of mathematics are to guide students to understand mathematical concepts, explain the relationships between concepts, and apply concepts or algorithms easily, accurately, efficiently, and appropriately in solving problems (Nisa & Faradiba, 2023; Wedastuti & Faradiba, 2023 (Tsabitah, 2020; Wedastuti, Fuady, et al., 2023; Sunismi, 2017). In addition, reasoning skills, mathematical manipulation, and mathematical communication are important aspects of mathematics learning (Khasanah et al., 2023; Wedastuti, Fuady, et al., 2023).Hodiyanto (2017:11) states that mathematical communication involves students' expressions in conveying and interpreting mathematical ideas through various media. Therefore, mathematical communication is considered an essential basic competency in the context of mathematics learning and mathematics education.

The use of instruments in measuring abilities cannot be separated from the process of testing the validity of these instruments (Sugiyono, 2017). To increase validity and reliability, these instruments must undergo further testing involving a group of students in a trial. This is because validity and reliability play a crucial role in maintaining and improving the quality of instruments, thereby avoiding potential flaws (Arikunto, 2010). The higher the validity and reliability of an instrument, the more accurate the data obtained from the research will be (Sugiyono, 2017). Howard and Hendry, as stated by Rosseni (2005), argue that the consistency of an instrument can be measured through repeated testing of the same or similar subjects or respondents. Thus, reliability relates to the level of measurement accuracy, which is a necessary, but not sufficient, condition for assessing the validity of measurement results (Arikunto, 2010). Low reliability is considered to limit validity, while high reliability still cannot guarantee optimal validity (Sugiyono, 2017). The reliability of items/questions in research instruments is usually measured using Cronbach's alpha (Nunnally & Bernstein, 1994). In addition, the validity and reliability of items/questions in instruments can be evaluated using the Rasch Model, a measurement model that considers the validity and reliability of each prospective respondent's response to each item/question and the level of difficulty of the item/question (Bond & Fox, 2007). Thus, this study aims to measure the validity and reliability of mathematical communication questions.

Method

This research is classified as *Research & Development* (R&D) (Amhar et al., 2020; Wedastuti, Sunismi, et al., 2023; Rumiati et al, 2023). Development research, often referred to as research and development (R&D), is a type of research that focuses on efforts to create or improve new products, processes, or methods. The main objective of development research is to produce innovation and progress in a particular field (Mustaqim, 2016; Widdah & Faradiba, 2022). The research was conducted in August 2023 at a high school in Malang. Before the instrument was implemented, it was tested for validity by expert validators and a small group (Arikunto, 2010). In this study, data collection was carried out by giving mathematical communication questions directly to small groups. To check the validity of the data, the instrument sheets were also examined by expert validators for comparison purposes in making decisions (Sugiyono, 2017). The product developed was a mathematical communication test instrument compiled based on the stages and indicators of mathematical communication. The stages and indicators of mathematical communication skills are presented in Table 1.

Table 1. Mathematical Communication Indicators

Stages of Mathematical Communication	Indicators
Understanding the problem	Being able to find the information contained in the question and asked in the question.
Planning the solution to the problem	Being able to determine the steps needed to solve the problem and the reasons behind them.
Implementing the problem-solving plan	Being able to solve the problem using the steps that have been determined.
Reviewing the entire solution	Able to check the accuracy of the results obtained from each step taken in solving the problem.

The test instrument in this study was in the form of essay questions. This test instrument referred to mathematical communication skills. This test instrument consisted of 5 test questions. The questions prepared were adjusted to the indicators to be achieved in students' mathematical communication skills. According to Sugiyono (2017:173), the validity of an instrument refers to its ability to correctly measure what should be measured. An evaluation tool is considered valid (legitimate or valid) if it is able to measure accurately as desired, or it is said to be valid if it is able to measure in accordance with reality. In the context of this study, the validation process involves three aspects, namely content validation, construct validation, and empirical validation. Sugiyono (2017:176) argues that test-based instruments need to meet the criteria of construct validity and content validity. Content validity is tested by comparing the content of the instrument with the teaching material, while construct validity relates to the ability of each question to achieve the test objectives. The success of the test objectives depends on the ability of each question to measure relevant indicators. The test instruments that have been prepared through the preparation stage then undergo a content validity process by experts before being used. This aims to ensure that the test instruments to be used meet the requirements as a valid data collection tool. Content validity was carried out by lecturers from the UNISMA mathematics education study program who have adequate knowledge and experience in the field of mathematics.

To test empirical validity, the test questions were piloted in class X IPS 3, and the results of the pilot test were then tested for validity. In this study, to determine the validity of the test questions, *the product moment* correlation formula proposed by Pearson was used as follows.

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{(N\sum X^2 - (\sum X)(N\sum Y^2 - (\sum Y))}}$$

Explanation:

r_{xy} : correlation coefficient between variable X and variable Y

$\sum X$: number of questions

$\sum Y$: total score

$\sum XY$: the sum of the multiplication of the item score by the total score

$\sum X^2$: sum of the squares of the item scores

$\sum Y^2$: sum of squares of total scores
N : number of subjects

To determine the validity level of the evaluation tool, the criteria in Table 2 below can be used.

Table 2. Classification of Correlation Coefficients (r_{xy})

Correlation Coefficient Index (r_{xy})	**Classification**
$0 \leq r_{xy} \leq 0,20$	Very low
$0.20 < r_{xy} \leq 0,40$	Low
$0.40 < r_{xy} \leq 0,60$	Moderate
$0.60 < r_{xy} \leq 0,80$	High
$0.80 < r_{xy} \leq 1,00$	Very high

In this study, data analysis was performed using the Independent Sample t-Test parametric statistical method. The entire data processing process was carried out using *SPSS* software *version 23*. The Independent Sample T-Test was applied to identify differences in mathematical communication skills, detailing the results or findings relevant to these variables.

Results and Discussion

Researchers conducted validity tests on this measuring instrument with the aim of ensuring that the five items had a single dimension, specifically in measuring mathematical communication. The validity of an instrument can be tested through various evidence, including content validity and construct validity. In this context, the study focused more on construct validity as the main consideration. Construct validity emphasizes the extent to which the measuring instrument reflects measurement results that are in accordance with its definition, which should be based on existing theory. If the definition of the variable has been well constructed based on the appropriate theory, and the questions or statements in the items are appropriate, then the instrument is considered valid in terms of construct validity (Fraenkel, Wallen, & Hyun, 2012).

In the analysis using SPSS 23 software with Pearson Correlation on the total score for each question, a fairly high correlation coefficient was obtained for all items, namely 0.775 (high) for question number 1, 0.631 (high) for question number 2, 0.876 (very high) for question number 3, 0.689 (high) for question number 4, and 0.835 (very high) for question number 5. To assess validity, the calculated correlation value must exceed the table correlation value. Based on the r product moment table with N = 20 and a significance level of 5%, the correlation table value is 0.3783. Therefore, it can be concluded that all items are considered valid, because the correlation value for each item exceeds the correlation value on the table. The validity calculations for the questions can be found in Table 3 (Adapted from Fraenkel, Wallen, & Hyun, 2012).

Table 3. Validity of Mathematical Communication Items

		ITEM1	ITEM2	ITEM3	QUESTION 4	QUESTION 5	TOTAL
QUESTION1	Pearson Correlation	1	.412	.500*	.759**	.447*	.775**
	Sig. (2-tailed)		.071	.025	.000	.048	.000
	N	20	20	20	20	20	20
QUESTIONS	Pearson Correlation	.412	1	.342	.277	.292	.631**
	Sig. (2-tailed)	.071		.140	.238	.212	.003
	N	20	20	20	20	20	20
QUESTION3	Pearson Correlation	.500*	.342	1	.428	.973**	.876**
	Sig. (2-tailed)	.025	.140		.059	.000	.000
	N	20	20	20	20	20	20
QUESTION4	Pearson Correlation	.759**	.277	.428	1	.361	.689**
	Sig. (2-tailed)	.000	.238	.059		.117	.001
	N	20	20	20	20	20	20
QUESTION 5	Pearson Correlation	.447*	.292	.973**	.361	1	.835**
	Sig. (2-tailed)	.048	.212	.000	.117		.000
	N	20	20	20	20	20	20
TOTAL	Pearson Correlation	.775**	.631**	.876**	.689**	.835**	1
	Sig. (two-tailed)	.000	.003	.000	.001	.000	
	N	20	20	20	20	20	20

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

The reliability of an instrument can be assessed through several reliability test methods, such as test-retest, equivalence, and internal consistency. In the internal consistency category, there are various test techniques, including split-half, KR 20, KR 21, and Cronbach's alpha. Each test technique has instrument criteria that can be tested using its respective method. In this study, the reliability of the instrument was tested using the test-retest method, in which one type of instrument was tested several times on the same subject or respondent. The reliability value of the instrument was measured from the correlation coefficient between the results of the first test and the subsequent tests. The instrument is considered reliable if the correlation coefficient is positive and significant. The correlation between the results of the first test and the subsequent test is analyzed using Product Moment correlation to find the correlation coefficient value. Using SPSS 23 software, the calculation results in the Reliability Statistics column using the Cronbach's Alpha method show a test reliability coefficient value of 0.812. This value indicates that the instrument used in this study has a very high level of accuracy, in accordance with the desired reliability criteria, namely $0.80 < r \leq 1.00$. Further information regarding the reliability test can be found in Table 4

Table 4. Reliability Table

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
item1	62.90	58.726	.664	.764
question2	63.50	58,474	.393	.844
question3	67.90	46,937	.770	.717
question4	63.60	61,726	.548	.792
question5	67.70	49,379	.704	.741

Instruments play a crucial role in research as a means of facilitating the data collection process. Instruments with good validity and reliability will produce accurate data, creating a strong basis for drawing conclusions that correspond to the reality of the research object (Sugiyono, 2017). Instrument validity is a measure of the extent to which an instrument can accurately measure what it is supposed to measure. Validity can be examined through content validity, construct validity, and criterion validity. Content and construct validity are measured by experts in the field. An instrument is considered valid in terms of content and construct if experts do not provide additional suggestions or input and accept the content, format, and construct of the instrument (Arikunto, 2010).

Criterion validity is assessed by comparing the instrument with its criteria, and this comparison can be tested using a correlation test. The higher the validity coefficient value approaching +1.00, the more valid the instrument is indicated to be (Sugiyono, 2017). Reliability, on the other hand, measures the extent to which a measurement can be trusted and consistent in its consistency. Instruments with two or more answer choices can be considered reliable if several measurements of the same subject (test-retest) produce relatively similar results or if a single measurement with a different instrument (equivalent) also produces relatively similar results (Arikunto, 2010).

Instruments with two answer choices are considered reliable if the item correlation coefficient (ri) value is greater than the table value (rt). Meanwhile, instruments with more than two answer choices are considered reliable if the Cronbach's Alpha reliability coefficient is in the range of 0.70-0.90 (Nunnally & Bernstein, 1994). Therefore, the validity and reliability of instruments are the main foundations for producing reliable data to support the validity of research and interpretation of results (Sugiyono, 2017).

Conclusion

The results of the construct validity and reliability evaluation of the mathematical communication instrument show that the five items have unidimensional properties, meaning that they only measure a single factor. The conclusion that can be drawn is that these five items are acceptable. This is because the five items of this instrument meet several criteria as good items, namely (1) they have positive factor loadings, (2) the validity value (rhitung) is greater than the table value (rtabel) with a significance level of 5%, which is obtained at 0.3783, and (3) only have a correlation between item measurement errors of no more than three, or in other words, the items are unidimensional.

References

Adamson, K. A. & Prion, S. (2013). Reliability: measuring internal consistency using Cronbach's α, Clinical Simulation in Nursing, 9, pp. 179-180.

Amhar, Y., Abidin, Z., & Faradiba, S. S. (2020). ANALYSIS OF SUCCESS RATES: RESULTS OF THE VALIDATION OF THE TANGKAS MATEMATIKA 14 LEVEL BOOK. *Journal of Mathematics Education*, *7*(2), 133–141. http://journal.stkip-andi-matappa.ac.id/index.php/histogram/index

Arikunto, Suharsimi. (2010). Research procedures: a practical approach. Jakarta: Rineka Cipta.

Arslan, C. (2009). Anger, self-esteem, and perceived social support in adolescence. Social Behavior and Personality: an international journal, 37(4), 555-564.

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch Model: Fundamental Measurement in the Human Sciences. New York: Routledge.

Branden, N. (1992). Power of Self Esteem. Barnes & Noble Books. Joreskog, K.G. &

Ministry of National Education of the Republic of Indonesia. (2006). Permendiknas RI Number 22 of 2006 concerning Content Standards. Jakarta: Ministry of National Education of the Republic of Indonesia.

Fraenkel, J. L., Wallen, N. E., & Hyun, H. H.. (2012). How to design and evaluate research in education eighth edition. New York: Mc Graw Hill.

Hodiyanto, H. (2017). Problem-Based Mathematics Learning and Improving Students' Mathematical Communication Skills. Journal of Mathematics Education Research, 5(1), 11-20.

Khasanah, I., Abidin, Z., & Faradiba, S. S. (2023). Analysis of Students' Errors in Solving Three-Dimensional Problems Using Watson's Criteria in Terms of Learning Styles and Gender Differences. *Cendekia Journal: Journal of Mathematics Education,* 7*(*2), 1976–1987. https://doi.org/10.31004/cendekia.v7i2.1923

Lina Rumiati, Surahmat, Sunismi, & Sikky El Walida. (2023). The Influence of Logical-Mathematical Intelligence on the Numeracy Literacy Ability of Junior High School Students Assisted by E-Modules. *Histogram: Journal of Mathematics Education*, *7*(2), 10–22.

Martín-Albo, J., Núñez, J. L., Navarro, J. G., & Grijalvo, F. (2007). The Rosenberg Self-Esteem Scale: translation and validation in university students. The Spanish Journal of Psychology, 10(02), 458-467.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric Theory (3rd ed.). New York: McGraw-Hill.

Rosseni, D. (2005). Measurement in Education. Kuala Lumpur: McGraw-Hill Education.

Ryan, L., & Caltabiano, M. L. (2009). Development of a new resilience scale: The Resilience in Midlife Scale (RIM Scale). Asian Social Science, 5(11), P39.

Sorbom, D. (1999). LISREL 8.70 for Windows (computer software). Lincoln-wood, IL: Scientific Software International, Inc.

Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency, Journal of Personality Assessment, 80 (1), pp. 99-103.

Suherman. (2003). Curriculum Development: Theory and Practice. Bandung: Remaja Rosdakarya.

Sugiyono. (2014). Statistics for Research. Bandung: Alfabeta.

Sugiyono. (2017). Educational Research Methods: Quantitative, Qualitative, and R&D Approaches. Bandung: Alfabeta.

Sunismi. (2017). Development of a Collaborative Learning Model for Mathematics Based on the Calculus II Course Blog Media. *National Seminar on Mathematics and Its Applications, Airlangga University*.

Sunismi, W. D., M. S. A., & A. M. I. (2022). University Readiness, Involvement and Lecturers' Knowledge for Implementing Newest Curriculum in Indonesia: A Survey in an Islamic University in Indonesia. *International Journal of Social Science Research and Review*, *5*(1), 127–133.

Tavakol, M. & Dennick, R. (2011). Making sense of Cronbach's alpha, International Journal of Medical Education, 2, pp. 53-55

Tsabitah, I. (2020). VALIDITY OF INTERACTIVE DIGITAL LKPD WITH AUGMENTED REALITY IN 3-DIMENSIONAL VECTOR MATERIAL. *Journal of Mathematics Education*, *7*(2), 124–132. http://journal.stkip-andi-matappa.ac.id/index.php/histogram/index

Umar, J. (2010). CFA Logic (Confirmatory Factor Analysis). Lecture material Werner, E. E. (1995). Resilience in development. Current directions in psychological science, 81-85

Wedastuti, N. K., Fuady, A., & Sunismi, S. (2023). Analysis of Final Exam Question Quality Using the Anates Program Application. *JIPM (Journal of Mathematics Education), 12(*1), 1. https://doi.org/10.25273/jipm.v12i1.15413

Wedastuti, N. K., Sunismi, Nf., & Fuady, A. (2023a). DEVELOPMENT OF E-MODULES BASED ON PISA MODEL LITERACY. *Kwangsan: Journal of Educational Technology, 11(*1), 459. https://doi.org/10.31800/jtp.kw.v11n1.p459--471