Journal of Philosophy and Research in Islamic Mathematics Education
J-PRIMa
Vol. 1, No. 1, 2025, pp 110–119 e-ISSN:
https://doi.org/10.33474/
online at: https://riset.unisma.ac.id/index.php/j-prima/article/view/25146

Innovative strategies for developing assessment tools to enhance reflective thinking skills in the learning evaluation process

Triayuningsih Permata Suci[1*]

[1]Graduate Program in Mathematics Education, University of Islam Malang, Indonesia
Email: *22202072009@unisma.ac.id*

Received: 17-09-2025; Revised: 27-10-2025; Accepted: 10-11-2025

Abstract. *The evaluation process plays an important role in achieving learning objectives. One evaluation technique that can be used in the learning process is the use of test questions. Test questions that are integrated with reflective thinking skills have the potential to encourage students to reflect, evaluate and draw conclusions regarding the steps taken in solving a problem. This research aims to develop an evaluation instrument in the form of test questions that can measure students' reflective thinking abilities. The research method applied is development research by adopting the 4D model (define, design, develop, and disseminate). The research sample consisted of 10 eighth-grade students in one of the junior high schools in Malang Regency, selected using random sampling techniques. The results of this research involve creating a mathematical problem instrument that focuses on reflective thinking skills, especially on the Pythagorean theorem material. The validity and reliability test shows that the three test items are declared valid and have good reliability. The results of the limited trial showed that the test questions had succeeded in measuring students' thinking abilities effectively. It should be noted that this research is still limited to initial trials at the development stage. Nevertheless, this research makes an important contribution in overcoming the limitations in current evaluations in measuring students' reflective thinking abilities.*

Keywords*: evaluation, instrument, reflective thinking ability.*

How to cite: Suci, T. P. (2025). Innovative strategies for developing assessment tools to enhance reflective thinking skills in the learning evaluation process. *Journal of Philosophy and Research in Islamic Mathematics Education (J-PRIMa)*, *1*(1), 110–119. Retrieved from https://riset.unisma.ac.id/index.php/j-prima/article/view/25146

Introduction

The evaluation process plays a central role in a teacher's duties, as it not only reflects student achievement but also determines the direction of future learning. In this context, evaluation is not merely about giving grades, but rather a guide to optimizing students' learning experiences and designing learning approaches that suit individual needs (Alfarizi & Fuady, 2023; Hasanah & Setiawan, 2023; Nisa & Faradiba, 2023) . This is in line with Suardipa & Primayana (2020) and Wedastuti et al (2023), who state that the basic purpose of evaluation or assessment is to determine the effectiveness and efficiency of learning activities, with the main indicator being the success in achieving learning objectives. 's view regarding learning evaluation emphasizes

that this objective is inseparable from the roles of teachers and students. This is in line with the opinion of Cahyaningrum et al (2023) that the purpose of evaluation is to see the extent of an educator's success in implementing the learning process. Educators conduct evaluations to assess their ability to organize learning, considering the extent to which the methods, strategies, and materials are in line with the institution's objectives (Amhar et al., 2023; Muspawi, 2021; Susanti & Faradiba, 2022). On the other hand, students undergo evaluation to provide teachers with information about their shortcomings and weaknesses during learning, as well as to identify material that has not been understood (Negara et al., 2023). Although various learning evaluation techniques such as exams, assignments, and project assignments are commonly used, these evaluations often do not adequately cover students' reflective thinking skills.

Students' reflective thinking skills are an important aspect that needs to be considered in the evaluation process. A more holistic evaluation should not only be limited to measuring factual knowledge, but also provide space for students to reflect, evaluate, and respond to learning materials in greater depth. According to Wulansari et al (2019), reflective thinking activities in students are active thinking activities in which students connect the knowledge they have acquired in solving problems to reach conclusions. Indicators of reflective thinking skills according to Surbeck, Han, and Mayor (in Fuady, 2017) include responding deeply (reacting), comparing, and contemplating. Improving reflective thinking skills requires a learning approach that encourages discussion, problem solving, and personal reflection. Evaluation instruments that can measure students' reflective thinking skills need to be carefully designed to reflect the complexity and depth of this aspect (Masamah, 2017). One form of instrument that can effectively measure students' reflective thinking skills is through test questions integrated with story situations or case studies. This is in line with research conducted by Miftahussa'adiah et al (2020) and Rumiati et al (2023) which explains that story questions can influence students' thinking abilities. This type of test question not only measures understanding of a concept, but also encourages students to reflect, evaluate, and detail the views or actions taken in a particular context. Story questions integrated with reflective thinking skills can provide a more comprehensive picture of the extent to which students can apply their knowledge in real-life situations (Atuni et al., 2023), while also providing space for students to express their personal thoughts and views (Ratnasari & Setiawan, 2022). These instruments can include questions that challenge students to elaborate on their thoughts when faced with ethical dilemmas, complex situations, or everyday problems that require critical consideration and self-evaluation (Hanafi et al., 2022; Widdah & Setiawan, 2023). Thus, test questions that are integrated with reflective thinking skills do not only measure students' knowledge, but also provide a more in-depth picture of the critical and reflective thinking skills that students possess.

The results of the research observation conducted in March at one of the junior high schools in Malang Regency produced several significant findings. One of the main findings was that students' reflective thinking skills were still relatively low. The main causes of this low level of ability were a lack of cognitive stimulation and a lack of critical thinking practice in the learning process at school. Responding to the need for improvement in this aspect, it is clear that innovation in the development of instruments based on reflective thinking skills is an urgent necessity to overcome the gaps in the current learning evaluation system. These instruments need to be able to measure students' reflective aspects more carefully and comprehensively. Therefore, the development of these instruments includes adjustments to evaluation techniques, one of which is through the use of story-based questions that can stimulate students to analyze and solve problems more deeply.

Method

This type of research is research and development. This research was conducted at a junior high school in Malang Regency. The product developed was an instrument consisting of essay-type questions testing reflective thinking skills. To obtain a development prototype, this research adapted the 4-D development model, which consists of four stages: define, design, develop, and disseminate. The define stage aimed to determine and define the reflective thinking ability test instruments needed. The design stage aimed to design the prototype test instruments to be developed, including a test item grid covering reflective thinking ability indicators based on Surberck, Han, and Mayor, and to create a scoring rubric. In addition to these components, the researchers also compiled a validation sheet instrument to be given to experts. The development stage involved expert judgment and was followed by limited testing.

Figure 1. 4D Model Development

The limited trial involved 10 eighth-grade students selected using random sampling techniques. This technique was chosen because it was considered to have a high chance of providing results close to the population parameters (Arieska & Herdiani, 2018). In addition, the random sampling process can reduce bias in sample selection. The results of the limited trial were then analyzed and revised (if necessary) to proceed to the next stage, namely the field trial. However, in this study, the field trial was not conducted and only the limited trial at the development stage was carried out.

The data for this study was collected from qualitative and quantitative data. Qualitative data consisted of expert validation instruments to measure the validity of test items before testing (Suci et al., 2023). Thus, the initial theoretical product data was assessed through expert judgment to identify, evaluate, and confirm the suitability of the product with the results. The expert assessment contained statements consisting of the content, language, and writing of the questions. Quantitative data consists of expert assessment scores for test items that have been previously validated and improved according to expert suggestions/input. The validity and reliability of the test items were tested using *SPSS software.* The validity criteria for test items were based on the significance value or comparison of $r_{(calculated)}$ with $r_{(table)}$ .If the significance value was $< 0{,}05$ or $r_{(calculated)} > r_{(table)}$,then the test item was declared valid. Meanwhile, the reliability of the test items is based on the reliability coefficient value (*Cronbach's Alpha* value*)*. If the reliability coefficient of the test item is $> 0{,}6$, then the item is declared reliable. If the reliability coefficient of the test item is $< 0{,}6$, then the item is declared unreliable.

Results and Discussion

Results

The definition stage was conducted to investigate the problems faced in greater depth. Several analyses were carried out at this stage, including: 1) curriculum analysis; 2) student characteristics analysis; and 3) student reflective thinking skills analysis. The curriculum analysis used the Merdeka Curriculum. Based on observations, the characteristics of students at one of the junior high schools selected by the researcher as the location for the study indicate that eighth-grade students are aged 12–14 years, have different backgrounds and economic circumstances, and have different levels of understanding and skills. This variability can affect students' learning experiences and thinking skills, including reflective thinking skills, which are one of the higher-order thinking skills. The results of observations and interviews with students showed that students had low reflective thinking skills. It was found that the questions given in the learning process did not stimulate students to think deeply. This finding was related to students' difficulties in understanding the material, especially story problems in geometry, as revealed in interviews with mathematics teachers. The next topic discussed was the Pythagorean theorem.

The design stage was carried out to create a prototype of the questions to be developed. These questions will be tailored to contextual problems or those that students often encounter in their daily lives. A total of three questions were developed, and each question was tailored to the indicators of reflective thinking skills. The indicators used by the researcher are presented in Table 1.

Table 1. Reflective Thinking Indicators based on Surberck, Han, and Mayor

Indicator	Sub-Indicator
Reacting	Describing the problem in the question
	Convey all the information provided.
	Connecting the problem between the knowledge possessed and a specific explanation
Comparing	Making comparisons between the problem situation faced and previous experiences.
	Planning strategies to solve the problem.
Contemplating	Implementing the resolution strategy.
	Stating conclusions by connecting them to the context of the problem.
	Identifying errors in the answers.
	Correcting and explaining errors in the answers.
	Concluding accurately.

(Wulansari et al., 2019)

Next, create an assessment rubric as a guide for clear and specific criteria for assessing student performance. The assessment rubric used is presented in Table 2 below.

Table 2. Assessment Rubric for the Reflective Thinking Ability Test Questions

No	Indicator	Score	Student Assessment Rubric
1	*Reacting*	3	Writing down known information correctly and completely
		2	Writing down known information correctly but incompletely
		1	Writing down known information but inaccurately
		0	Not writing down any information
2	*Comparing*	3	Can provide correct and systematic solutions and can connect the information provided with the appropriate steps
		2	Can provide correct and systematic solutions based on the information provided, but some steps are still not appropriate
		1	Can provide a solution but cannot connect the information provided
		0	Does not provide a solution
3	*Contemplating*	3	−Evaluating/checking the validity of an argument based on the concepts/properties used appropriately −Drawing correct conclusions
		2	−Evaluating/checking the validity of an argument based on the concepts/properties used appropriately −Drawing conclusions that are inaccurate
		1	−Evaluating/checking the validity of an argument based on the concepts/properties used but inaccurately −Drawing conclusions that are inaccurate
		0	−Unable to evaluate/examine the validity of an argument based on the concepts/properties used −Does not draw conclusions

A validation sheet instrument is also developed at this stage, which covers content, language, and question writing aspects. This instrument is used to measure the theoretical validity of the developed questions.

The development stage involves expert judgment. The results of the expert assessment explain that the items are valid but still need improvement. Some suggestions or input from experts regarding improvements include: 1) the marking of corner points on the image is too small; 2) alternative solutions for each question; and 3) the writing of area units. After all improvements were made and consulted again with experts, the product was declared suitable for testing in the next stage, which was a limited trial.

The limited trial involved 10 eighth-grade students. The initial flow of this trial was to give the students test questions that had been declared valid and feasible. There were three test questions. One of the questions is presented in Table 3. After the students completed the test questions, the next step was to analyze their answers to obtain validity and reliability data.

Table 3. Reflective Ability Test Questions on the Pythagorean Theorem

Question 1
In the city park, Rafka is flying a kite. The length of the string he has is 130 meters. When the kite is above a certain point in the park, Rafka measures the horizontal distance from himself to the point below the kite as 50 meters.
a. Draw an illustration based on the question above!
b. Calculate the height of the kite above the city park!
c. Are you sure about your answer? Give your reasons!

The results of the item validity test are shown in Table 4 and the results of the test reliability test are shown in Table 5.

Table 4. Results of the Item Validity Test

		Question 1	**Question 2**	**Question 3**	**Total**
Question 1	Pearson Correlation	1	.506	.460	.822**
	Sig. (2-tailed)		.135	.181	.004
	N	10	10	10	10
Question 2	Pearson Correlation	.506	1	.564	.869**
	Sig. (2-tailed)	.135		.089	.001
	N	10	10	10	10
Question 3	Pearson Correlation	.460	.564	1	.756*
	Sig. (2-tailed)	.181	.089		.011
	N	10	10	10	10

Total	Pearson Correlation	.822**	.869**	.756*	1
	Sig. (2-tailed)	.004	.001	.011	
	N	10	10	10	10

Table 5. Test Item Reliability Test Results

Cronbach's Alpha	Number of Items
.729	3

The statistics of the analysis results of the reflective thinking ability test questions in the limited trial stage are shown in Table 6.

Table 6. Limited Trial Results

Statistical Scale	Score
Mean	7.90
Median	8.00
Standard Deviation	0.738

Disscussion

Several analyses were conducted during the definition stage, the first of which was a curriculum analysis. The curriculum used was the Merdeka Curriculum, which gives schools or teachers the freedom to design curricula according to the needs, context, and characteristics of students (Arfany & Faradiba, 2022; Marisa, 2021; Sobirin et al., 2023). Second, the characteristics of junior high school students aged between 12 and 14 years. According to Piaget, junior high school students in this age range are classified as being in the formal operational stage, where they are able to think abstractly without having to rely on objects or events that exist directly. The reasoning process at this stage occurs when students' cognitive structures have developed to the point where they are able to use symbols, ideas, abstractions, and generalizations. Diversity in background and economy, along with variations in understanding and skills, are the focus of researchers' attention. This has the potential to influence students' character and behavior when facing certain problems (Azizah, 2023) because in learning, students are often faced with various problems that require solutions. To successfully solve a problem, students need to combine their learning experiences and thinking skills. Students are required not only to analyze the situation before making a decision, but also to relate their prior knowledge to the information obtained and to evaluate and explain the actions taken (Fuady, 2017). The results of observations and interviews with students describe that their reflective thinking skills tend to be low. This finding highlights the lack of stimulation in the questions given during the learning process, which causes students to have difficulty thinking deeply. This challenge is more specifically related to understanding the material, especially in dealing with story-based problems related to geometry,

as revealed in interviews with mathematics teachers. To overcome this obstacle, the next topic to be studied is the Pythagorean theorem. Understanding that students' reflective thinking skills are a determining factor, the research will focus on developing integrated test questions that can improve students' understanding and thinking skills in the context of geometry, particularly in relation to the Pythagorean theorem.

The design stage and product prototyping involve the development of test questions that serve as evaluation tools. These test questions are designed with consideration of the relevant problem context and are integrated with indicators of reflective thinking skills. The next step is the preparation of an assessment rubric that serves as a guide with clear and specific criteria for assessing student performance. In addition, at this stage, a validation sheet instrument is also compiled, covering aspects of content, language, and question writing. This validation sheet is an essential tool for measuring the theoretical validity of each question item developed. By integrating these steps into the design process, it is hoped that the resulting product prototype can be an accurate and effective evaluation tool for measuring students' reflective thinking skills.

During the development stage, *expert judgment* was conducted on the developed test items. The expert assessment results showed that the test items were valid, although there were several aspects that needed improvement. A limited trial involved 10 eighth-grade students to evaluate the test items that were considered valid and feasible. The validity test results in Table 4 show that the significance values for questions 1, 2, and 3 are $0 < 0.05$. The three test questions have $r_{(calculated)}$ values of 0.004, 0.001, and 0.011, respectively. By comparing these values with the $r_{(table)}$ value with N=10, it can be concluded that $r_{(count)}$ is greater than $r_{(table)}$ for each test item, including numbers 1, 2, and 3. Therefore, it can be concluded that the three test items meet the validity criteria that have been set. The results of this limited trial provide confidence that the test questions developed can provide a valid measurement of students' ability to respond to story questions on the material being tested. In Table 5, *Cronbach's Alpha* coefficient for the test questions after the limited trial was 0.729 ($> 0{,}6$). This means that the test questions tested have good reliability. Based on the measurement results in Table 6, it can be concluded that the average score of the limited trial was 7.90. This figure indicates that the test questions have successfully measured students' reflective thinking skills effectively.

This research and development aims to create an effective evaluation tool to measure students' reflective thinking skills on the Pythagorean theorem material. Based on the results report, the test instrument that has been developed has successfully met the validity standards (theoretically and empirically reliable) and is considered reliable.

Conclusion

The test questions designed must meet the validity and reliability standards in accordance with the purpose of the test. The results of the research and development, which were carried out in three stages following the 4D development model, show that the test instrument for measuring reflective thinking skills on the Pythagorean theorem material has met the criteria for theoretical and empirical validity, as well as reliability. Thus, it can be concluded that the test instrument that has been developed meets the standards for measuring students' reflective thinking skills on the Pythagorean theorem material. Although the development of this test has only reached the limited trial stage and measurement of reflective thinking skills on the Pythagorean theorem material, further research can involve field trials and measurements of other abilities such as creative thinking, opening up opportunities for broader and more in-depth research.

References

Alfarizi, M., & Fuady, A. (2023). Analysis of the Effectiveness of STEM-Based Learning in Improving Mathematical Conceptual Understanding. *Journal of Mathematics Education*, *7*.

Amhar, Y., Abidin, Z., & Faradiba, S. S. (2023). Analysis of Success Rates: Results Validation of the Tangkas Matematika 14 Level Book. *Journal of Mathematics Education*, *7*.

Arfany, F. P., & Faradiba, S. S. (2022). Characterization of Students' Mathematical Conceptual Errors in Solving Relation and Function Problems. *Cendekia Journal: Journal of Mathematics Education*, *6* (3), 3044–3058. https://doi.org/10.31004/cendekia.v6i3.1780

Arieska, P. K., & Herdiani, N. (2018). Selection of Sampling Techniques Based on Relative Efficiency Calculations. *Jurnal Statistika Universitas Muhammadiyah Semarang*, *6*(2).

Atuni, S. N. J., Setiawan, Y. E., Faradiba, S. S., Teeside, S. V. S., & Zoker, E. M. (2023). The Process of Students' Mathematical Literacy in Solving System of Two Variables Linear Equation Based on Level of Ability. *Numerical: Journal of Mathematics and Mathematics Education*, *7* (2

Azizah, N. L. (2023). The Influence of Assertive Behavior on Students' Personality Types in Mathematics Learning. *Journal of Mathematics Education*, *7*.

Cahyaningrum, I. Y., Fuady, A., & Faradiba, S. S. (2023). Characterization of Visual and Symbolic Mathematical Representations of Ninth Grade Students on Transformation Material. *Cendekia Journal: Journal of Mathematics Education*, *7* (3), 2646–2659. https://doi.org/10.31004/cendekia.v7i3.1944

Choiroh, M. (2021). Evaluation of Arabic Language Learning Based on E-Learning Media. *Naskhi Journal: Journal of Arabic Language and Education Studies*, *3* (1), 41–47. https://doi.org/10.47435/naskhi.v3i1.554

Destiana, D., Suchyadi, Y., & Anjaswuri, F. (2020). Development of Assessment Instruments to Improve the Quality of Productive Learning in Elementary Schools. *Journal of Elementary School Teacher Education and Teaching (JPPGuseda)*, *03*(02), 119–123.

Fuady, A. (2017). Reflective Thinking in Mathematics Learning. JIPMat, *1*

Hanafi, M., Syamsuri, S., & Anwar, M. (2022). Development of High Order Thinking Skills Mathematics Question Instruments Based on Brookhart's Context of Pandeglang Batik Motifs for MTs Students. *Mathematics Education Media*, *10*(1), 43–59.

Hasanah, F. R., & Setiawan, Y. E. (2023). Math Anxiety and Metacognitive Awareness Inventory of Junior High School Students. *AKSIOMA: Journal of Mathematics Education Study Program*, *12* (3), 3038. https://doi.org/10.24127/ajpm.v12i3.7203

Marisa, M. (2021). Curriculum Innovation "Independent Learning" in the Era of Society 5.0. *Santhet: Journal of History, Education, and Humanities*, *5*(1), 66–78.

Masamah, U. (2017). Improving High School Students' Reflective Mathematical Thinking Skills Through Problem-Based Learning in Terms of Initial Mathematical Ability. *Journal of Mathematics Education Research*, *1*(1), 1–18.

Miftahussa'adiah, M., Alberida, H., & Handayani, D. (2020). Development of Critical Thinking Assessment of Circulatory System Material for Grade XI High School Students. *SIMBIOSA*, *9* (1)

Muspawi, M. (2021). Strategies for Improving Teacher Performance. *Scientific Journal of Batanghari University Jambi*, *21* (1), 101. https://doi.org/10.33087/jiubj.v21i1.1265

Negara, F. P., Abidin, Z., & Faradiba, S. S. (2023). Improving Students' Mathematics Self-Efficacy Through Problem-Based Learning. *Cendekia Journal: Journal of Mathematics Education*, *7* (1), 455–466. https://doi.org/10.31004/cendekia.v7i1.1943

Nisa, F., & Faradiba, S. S. (2023). Mathematical Literacy Profile of Students Based on Problem-Solving Ability Levels in PISA Questions. *Cendekia Journal: Journal of Mathematics Education*, *7* (2), 1003–1019. https://doi.org/10.31004/cendekia.v7i2.2211

Ratnasari, J. R., & Setiawan, Y. E. (2022). Numeracy Literacy of Students in Solving Quadrilateral and Trapezoid Problems. *AKSIOMA: Journal of Mathematics Education Program*, *11* (3)

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

Sobirin, J. H., Setiawan, Y. E., & Faradiba, S. S. (2023). Mathematical Representations Based on Field-Dependent and Field-Independent Cognitive Styles. *JME (Journal of Mathematics Education)*, *8* (2), 137–149. https://doi.org/10.31327/jme.v8i2.1981

Suardipa, I. P., & Primayana, K. H. (2020). The Role of Learning Evaluation Design in Improving Learning Quality. *Widyacarya: Journal of Education, Religion and Culture*, *4* (2), 88–100. https://doi.org/10.55115/widyacarya.v4i2.796

Suci, T. P., Fuady, A., & Faradiba, S. S. (2023). Effectiveness of Augmented Reality (AR) Use in Reviewing Creative Mathematical Thinking Skills. *Journal of Mathematics Education*, *7*.

Susanti, E., & Faradiba, S. S. (2022). Analysis of Students' Mathematical Connection Skills in Solving Mathematical Problems Based on the Metacognitive Awareness Inventory. *Cendekia Journal: Journal of Mathematics Education*, *6* (2), 1203–1209. https://doi.org/10.31004/cendekia.v6i2.1344

Wedastuti, N. K., Fuady, A., & Sunismi, S. (2023). Analysis of the Quality of Final Semester Exam Questions Using the Anates Program Application. *JIPM (Journal of Mathematics Education)*, *12* (1)

Widdah, H., & Setiawan, Y. E. (2023). Analysis of Mathematical Problem-Solving Ability in Relation to Students' Motivation to Learn Mathematics. *AKSIOMA: Journal of Mathematics Education Program*, *12* (2)

Wulansari, M. D., Purnomo, D., & Utami, R. E. (2019). Analysis of Eighth Grade Students' Reflective Thinking Skills in Solving Mathematical Problems in Terms of Visual and Auditory Learning Styles. *Imajiner: Journal of Mathematics and Mathematics Education*, *1* (6).