Journal of Philosophy and Research in Islamic Mathematics Education
J-PRIMa
Vol. 1, No. 1, 2025, pp 93–109 e-ISSN:
https://doi.org/10.33474/
online at: https://riset.unisma.ac.id/index.php/j-prima/article/view/25140

Assessing students' mathematical problem-solving skills: framework for developing tests on the pythagoras theorem

Febrianti Rofidatul Hasanah[1]

[1]SD Plus Muhammadiyah 1 Waru Pamekasan, Indonesia
*Correspondence: febriantirofidatulhasanah@gmail.com

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

Abstract. *Good mathematical problem-solving skills are acquired through practice and training. This can be achieved by regularly administering tests that measure students' mathematical problem-solving skills. The development of this test aims to obtain a test that is capable of measuring students' mathematical problem-solving skills. The Research & Development (R&D) approach was used to develop a test instrument for mathematical problem-solving skills on the Pythagorean theorem. Seven stages adapted from the Borg & Gall development model were followed to obtain a good test instrument that met the test objectives. The total sample size for the trial was 87 eighth-grade students from a junior high school in Pamekasan Regency. The sample was selected using random sampling techniques. The instrument consisted of an expert validation sheet and a test sheet that was developed. Data were collected through expert validation and pilot testing. The expert validation results were analyzed descriptively to measure the theoretical validity of the developed test. Meanwhile, the pilot test data were analyzed using Classical Test Theory (CTT) parameters, which included empirical validity, reliability, difficulty level, and discrimination power tests. The results showed that each test item was valid, reliable, had different levels of difficulty, and had good discriminating power. Thus, the developed test instrument met the criteria for a good test and could be used to measure mathematical problem-solving skills.*

***Keywords**: Mathematical Problem-Solving Ability, Test Development, Reliability, Validity.*

How to cite: Hasanah, F. R. (2025). Assessing students' mathematical problem-solving skills: framework for developing tests on the pythagoras theorem. *Journal of Philosophy and Research in Islamic Mathematics Education (J-PRIMa)*, *1*(1), 93–109. Retrieved from https://riset.unisma.ac.id/index.php/j-prima/article/view/25140

Introduction

Mathematical problem solving is one of the important skills that needs to be developed to improve student learning success (Badrulaini, 2018; Cahyaningrum et al., 2023; Chamot et al., 1992; Guven & Cabakcor, 2013; Hasanah & Setiawan, 2023; Nisa & Faradiba, 2023). Previous studies have shown that the mathematical problem-solving skills of students in Indonesia are relatively low, especially in geometry and the Pythagorean theorem (Musa et al., 2017; Ratnasari & Setiawan, 2022; Simbolon & Sapri, 2022; Susanti & Faradiba, 2022; Wulandari et al., 2021). Many factors influence students' mathematical problem-solving abilities. Some of these are affective factors such as mathematical disposition, self-efficacy, and mathematical anxiety (Hasanah & Setiawan, 2023; Negara et al., 2023; Sari & Faradiba, 2022; Wedastuti & Faradiba, 2023). The main factors causing low mathematical problem-solving skills are students' lack of understanding of the problems given and the material taught (Arfany & Faradiba, 2022; Hegarty et al., 1995; Verschaffel et al., 1999; Wulandari et al., 2021). Both factors occur because students are not accustomed to practicing or being presented with mathematical problem-solving questions (Amhar et al., 2020; Wedastuti, Sunismi, et al., 2023; Wijayanti et al., 2023).

One effort to overcome this is by administering mathematical problem-solving tests on a regular basis. The test instruments provided must be able to measure and train students' mathematical problem-solving skills. Therefore, the questions in the test are not compiled randomly and as they are, but are compiled in accordance with the learning objectives and the objectives of the test (Faisal, 2022; Magdalena et al., 2020; Secolsky & Denison, 2012). If a test is developed with the aim of measuring students' mathematical problem-solving abilities, then the test items must meet the indicators of mathematical problem-solving abilities.

A test instrument is categorized as a good evaluation tool if the items on the test are not only in line with the learning objectives and test objectives, but also meet the criteria of validity, reliability, level of difficulty, and good discriminating power (Fiska et al., 2021; Magdalena et al., 2020, 2021; Secolsky & Denison, 2012; Setiyawan & Wijayanti, 2020). Validity is a principle that requires an instrument to truly measure the intended concept, that the hypothesized causes truly produce the observed effects, and that the results obtained from a population using the instrument can be generalized to other populations (Lecompte & Bear, 2017). Therefore, validity is the most critical consideration in a test item that is being developed (Secolsky & Denison, 2012). Meanwhile, reliability is a principle that describes the consistency of scores over time, test form, or internally within the test itself (Secolsky & Denison, 2012). In other words, reliability measures the extent to which test items can provide reliable interpretations (Alfarizi & Fuady, 2020). Validity and reliability represent the credibility and quality of a test item (Lecompte & Bear, 2017). Thus, the better the validity and reliability of a test item, the better its quality.

A good test instrument should contain test items with varying levels of difficulty in proportion (Aiken, 1989). In other words, there should be a balance in the number of items classified as easy, medium, and difficult in a test instrument (Annisah et al., 2020; Magdalena et al., 2021). The level of difficulty is not determined by the perspective of the item writer, but rather by the ability of the student to answer the item (Clauser & Hambleton, 2011; Magdalena et al., 2021). Measuring the level of difficulty is necessary to determine the quality of each item on a test (Annisah et al., 2020). The higher the difficulty index of a test item, the lower its difficulty level ( ), and vice versa. Test items with difficulty indices that are exceptionally high or low compared to other test items need to be reviewed and revised/replaced (Clauser & Hambleton, 2011; Magdalena et al., 2021). Good test items should not fall into the "very easy" or "very difficult" categories (Annisah et al., 2020; Clauser & Hambleton, 2011; Wedastuti, Fuady, et al., 2023). Items that are too easy tend not to stimulate students' ability to solve mathematical problems (Magdalena et al., 2020; Nurhalimah et al., 2022).

Discriminatory power is also one of the indicators of the quality of a test item. A good test item can distinguish or "discriminate" between students in the upper group and students in the lower group (Clauser & Hambleton, 2011). The discriminatory power of a test item is based on the item discrimination index. The higher the discrimination index of an item, the better its ability to distinguish between groups of students who have understood the material and those who have not (Clauser & Hambleton, 2011; Magdalena et al., 2021).

The purpose of this research and development is to develop a test instrument that can measure students' mathematical problem-solving abilities on the Pythagorean theorem material. The test instrument will be developed in such a way that it meets the criteria for a good evaluation tool, namely, it is theoretically and empirically valid, reliable, has a varied level of item difficulty, and has good discriminating power (Fiska et al., 2021; Magdalena et al., 2020, 2021; Secolsky & Denison, 2012; Setiyawan & Wijayanti, 2020). The test instrument to be developed will only contain essay-type questions because the purpose of the test is to measure or describe students' mathematical problem-solving abilities in the Pythagorean theorem material.

Several developments of mathematical problem-solving ability test instruments have been carried out previously. Widodo et al. (2021) developed a mathematical problem-solving ability test instrument on geometry material for junior high school level. Unfortunately, the test development process carried out by Widodo et al. (2021) was limited to the expert validation stage and had not been tested. The results of the research by Widodo et al. (2021) show that the test instrument on geometry material developed has met the criteria of validity and practicality with a V-Aiken value of more than 0.532. A mathematical problem-solving ability test instrument was also developed by Lukman et al. (2023) on the topic of algebra. However, the test instrument development stage was limited to the content validity stage. Therefore, we developed a mathematical problem-solving test instrument containing Pythagorean theorem material from the initial research stage to the field trial stage, resulting in a final product that is ready for use.

Method

This research is classified as *Research & Development* (R&D). The product developed is a mathematical problem-solving test instrument compiled based on Polya's (1985) stages and indicators of problem-solving. The stages and indicators of problem-solving ability are presented in Table 1.

Table 1. Polya's(1985) Problem-Solving Indicators

Problem-Solving Stage	**Indicators**
Understanding the problem	Able to find information contained in the question and asked in the question.

Problem-Solving Stage	Indicators
Planning the solution	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 verify the accuracy of the results obtained from each step taken in solving the problem.

The mathematical problem-solving test instrument was designed and developed through several stages adapted from the Borg & Gall (1983) development model. The model consists of 10 stages of development. However, we adapted it so that there are seven stages used, namely 1) research and information gathering, 2) planning, 3) initial product development, 4) limited testing, 5) initial product revision, 6) field testing, and 7) final product revision. The research and information gathering stage aims to examine the problems encountered in greater depth based on relevant theoretical studies (Borg & Gall, 1983). In addition to literature studies, researchers analyze the problems encountered, the conditions of the students, and the learning environment. Observations and interviews with teachers and students are conducted to analyze the three aforementioned aspects. Task analysis and material analysis were also carried out by breaking down the Learning Outcomes (CP) and Learning Objectives (TP), which were then translated into test question indicators.

At the planning stage, the researcher compiled a test question grid, test questions, and guidelines for scoring mathematical problem-solving test questions. These three things were compiled based on the results obtained from the research and information gathering stages. The expert validation sheet instrument was also compiled at this stage based on three assessment aspects, namely the content aspect of the test questions, the language aspect, and the writing aspect of the test questions.

In the initial product development stage, theoretical validation was carried out through *expert judgment* or by requesting the assessment of several experts (Hamdi et al., 2022; Putra & Zarita, 2022). The experts/validators who assessed the initial product were mathematics education experts with more than 11 years of teaching experience. The results of theoretical validation are followed up by making improvements to the initial product based on the suggestions and input from the experts. After going through the revision process, the initial product is consulted again with the experts until the product is declared suitable for use.

A limited trial of the initial product was conducted involving 20 eighth-grade students at a junior high school in Pamekasan Regency as respondents or trial samples. The results of the limited trial were analyzed and used as a basis for making improvements (if any) to the initial product developed. The parameters measured in the trial results included empirical validity, reliability, difficulty level, and test item discrimination. If the analysis results showed that the initial product was valid, reliable, had a varied difficulty level, and good discrimination, then it proceeded to the next stage.

The field trial stage was conducted on the limited trial product that had met the four parameters measured. The field trial sample consisted of 67 eighth-grade students. The field trial results were analyzed and measured for empirical validity, reliability, level of difficulty, and discriminating power. Next, revisions are made to the final product (if any revisions are needed) before the product is finally used. In general, the flow of developing mathematical problem-solving test questions on the Pythagorean theorem material is illustrated in Figure 1.

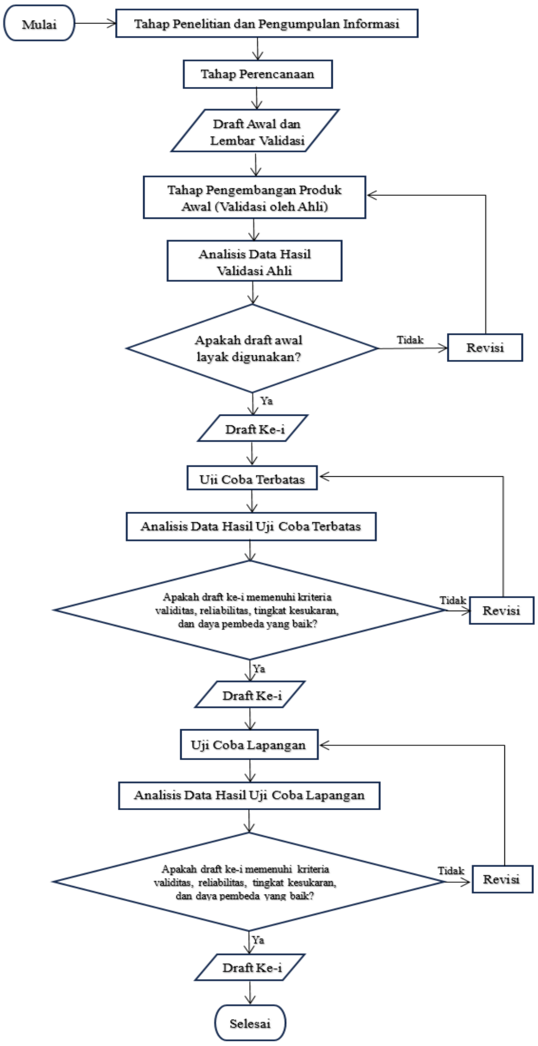


Figure 1. Test Item Development Process

The data collected consisted of qualitative and quantitative data. The qualitative data collection instrument was an expert validation sheet used to measure the validity or feasibility of test items before they were tested (Tsabitah & Sunismi, 2020). Therefore, the data from the theoretical validation of the initial test items through *expert judgment* is qualitative data, which is then analyzed using descriptive analysis. The validation criteria for each test item consist of four options, namely suitable for use without revision, suitable for use with minor revisions, suitable for use with major revisions, and not yet suitable for use.

The quantitative data collection instrument consisted of test questions that had been theoretically validated and revised according to the suggestions of validators/experts. Quantitative data were obtained through limited trials and field trials. The sample or pilot test respondents were eighth-grade students at a junior high school in Pamekasan Regency. The pilot test data, both from the limited pilot test and the field pilot test, were analyzed using the parameters *of classical test theory (*CTT) (Secolsky & Denison, 2012). The CTT parameters measured in the test items developed were limited to empirical validity, reliability, difficulty level, and item discrimination (DeVellis, 2006; Magno, 2009; Secolsky & Denison, 2012). This is because the test items developed are essay-type items, so there is no need to measure the effectiveness of distractors.

Empirical validity and reliability testing of test items was conducted using *SPSS software*. Meanwhile, the difficulty level and discriminating power of test items were analyzed using the program with the assistance of *Microsoft Excel*. 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 item was declared valid. Meanwhile, the reliability of test items is based on their reliability coefficient (*Cronbach's Alpha* value*)*. If the reliability coefficient of a test item is$> 0{,}6$, then the item has good reliability, and vice versa. To assess the level of difficulty of test items, the criteria in Table 2 are used, with p being the difficulty index of the item.

Table 2. Categories of Difficulty Level (p) of Test Items

Difficulty *Index (* (p)	**Difficulty Level**
$p \leq 0{,}30$	Difficult
$0{,}30 < p \leq 0{,}70$	Moderate
$p > 0{,}70$	Easy

(Arikunto, 2020)

The criteria for classifying test items based on discriminative power are based on the discrimination index as shown in Table 3.

Table 3. Test Item Discrimination Power Categories

Discrimination Index (DB)	Category
$DB \geq 0{,}40$	Good (accepted)
$0{,}30 \leq DB < 0{,}40$	Fairly Good (accepted with improvements)
$0{,}20 \leq DB < 0{,}30$	Poor (requires further discussion, needs improvement and should be targeted for improvement)
$DB < 0{,}20$	Not Good (discarded/rejected and replaced with another item)

The mathematical problem-solving test instrument developed for the Pythagorean theorem material will be considered good if it is proven to be valid, reliable, has varying levels of difficulty, and has good discriminatory power. Test questions that meet these four parameters become the final product that can be used widely.

Results and Discussion

The research and information gathering stage was conducted to examine in depth the problems encountered, namely the low mathematical problem-solving ability of students in solving problems, especially descriptive questions in the form of story problems. An analysis of the learning conditions and environment of students was also conducted at this stage. The results of observations and interviews with students showed that they were not accustomed to solving descriptive mathematical problems such as story problems. The questions given so far were limited to multiple choice, fill-in-the-blank, and descriptive questions, but not in the form of story problems. The descriptive questions given were limited to the application of concepts to simple questions. The results of observations and interviews with mathematics teachers also show that students' ability to understand material and problems in the form of story problems is relatively low, especially in geometry. There is concern that giving problems such as story problems in class will take up a lot of time and result in some material/sub-material not being completed or taught by the end of the semester.

Task and material analysis was also conducted during the research and information gathering stage. The next topic to be taught is the Pythagorean theorem. Therefore, the Learning Outcomes (CP) and Learning Objectives (TP) related to this material were outlined to become question indicators. The learning objective chosen for the test questions developed was to solve everyday problems related to the Pythagorean theorem. This learning objective was chosen because the evaluation must be in the form of everyday problems, which are usually presented in the form of story problems. The purpose of the instrument developed is to measure students' mathematical problem-solving abilities.

Therefore, at this stage, a literature study was also conducted related to the development topic, such as problem-solving ability indicators, evaluation tool measurement, test development models, and so on.

At the planning stage, the test blueprint, test items, and scoring guidelines were developed based on the results of research and information gathering in the previous stage. Therefore, the test blueprint, test items, and scoring guidelines were based on Polya's (1985) indicators of mathematical problem-solving skills. The test items are arranged in such a way that they contain Polya's four indicators of mathematical problem solving. The test instrument developed consists of two essay questions containing pictures. The test scoring guidelines contain alternative answers that may be written by students. Thus, there is more than one alternative answer in the scoring guidelines. An expert validation sheet was also compiled at this stage, containing three assessment aspects, namely the content aspect of the test items, the language aspect, and the writing aspect of the test items. The expert validation sheet was used to measure the theoretical validity of the developed test items. A summary of the validation results by each validator is presented in Table 4.

Table 4. Summary of Expert Validation Results

Validator	Validation Results	Suggestions/Input
Expert I	Valid (suitable for use with minor revisions)	The question indicators should be made more concise and begin directly with operational verbs
Expert II	Valid (suitable for use with minor revisions)	Corrected spelling errors (*typos*), including in the writing of area units
Expert III	Valid (suitable for use with minor revisions)	The color of the letters indicating the names of the corners in the image has been changed to black so that they are more visible when printed

The test question grid, test items, and scoring guidelines were revised according to the validators' suggestions and input. The question indicators were made more concise, typing errors were corrected, and the letters indicating the vertices of flat shapes in the pictures in the questions were made black. After the revision process was carried out, the researchers consulted with experts again until the product was declared suitable for testing.

A limited trial was conducted on a small group of 20 students. The results of the limited trial were analyzed to obtain data on the validity, reliability, difficulty level, and discriminating power of the test items. The results of the item validity test are shown in Table 5.

Table 5. Results of Item Validity Testing in Limited Testing

		Item No. 1	Item No. 2	Score
Item No. 1	Pearson Correlation	1	$.928^{**}$	$.981^{**}$
	Sig. (2-tailed)		0.000	0.000
	N	20	20	20
Item No. 2	Pearson Correlation	$.928^{**}$	1	$.982^{**}$
	Sig. (2-tailed)	0.000		0.000
	N	20	20	20
Score	Pearson Correlation	$.981^{**}$	$.982^{**}$	1
	Sig. (2-tailed)	0.000	0.000	
	N	20	20	20

In Table 5, the significance values for items 1 and 2 are $0{,}000 < 0{,}05$. The $r_{calculated}$ values for each test item are 0.981 and 0.982. Meanwhile, the $r_{(table)}$ value with $N = 20$ and a significance of 5% is 0.444. Thus, $r_{(calculated)} > r_{(table)}$ applies to both item number 1 and number 2. Therefore, it can be concluded that both test items have met the validity criteria. The results of the test item reliability test in the limited trial stage are presented in Table 6.

Table 6. Test Item Reliability Test Results in the Limited Trial

Cronbach's Alpha	Number of Items
0.982	2

In Table 6, *Cronbach's Alpha* coefficient for test questions after limited testing was 0.982 ($> 0{,}6$). This means that the test questions tested have good reliability. The results of the analysis of the difficulty level of each test question are presented in Table 7.

Table 7. Results of the Analysis of Test Item Difficulty Levels in the Limited Trial

Question Number	p	Difficulty Level
1	0.61	Moderate
2	0.34	Moderate

Based on Table 7, both items have different difficulty indices but are classified in the same difficulty category. The difficulty level of both test items is moderate. Thus, a slight improvement is needed on one of the test items so that the difficulty level of the developed test instrument becomes more varied. The results of the item discrimination analysis can be seen in Table 8.

Table 8. Results of Item Discrimination Analysis in Limited Testing

Item Number	Discrimination Index (DB)	Item Discrimination Category
1	0.39	Fairly Good
2	0.67	Good

Based on Table 8, item number 1 has fairly good discrimination. Therefore, based on the test item discrimination category in Table 3, test item number 1 is acceptable but needs improvement. Item number 2 already has good discrimination, so it can be used directly in the field trial stage without needing improvement. In general, the statistics of the analysis results of the initial product of the mathematical problem-solving test questions at the limited trial stage are summarized in Table 9.

Table 9. Statistics of the Analysis Results of the Initial Test Items for Problem-Solving Ability

Statistical Scale	Value
Mean	47.10
Standard Deviation	24.85
Median	51.00
Reliability Coefficient	0.982
Average Difficulty Level	0.475
Average Discrimination Power	0.530

Table 9 shows that the average score of students was 47.10 and the average difficulty index of the test questions was 0.475. This means that the difficulty level of the test questions was moderate. Meanwhile, the average discrimination index or discriminating power of the questions was 0.530 $(\geq 0{,}40)$, which means that the test questions have good discriminatory power. In other words, the developed mathematical problem-solving test questions have been able to distinguish well between students who have mastered and those who have not mastered the Pythagorean material.

Test items that needed revision based on the analysis of the limited trial data were corrected before being used in the field trial. After the initial test items underwent the revision process, the field trial was conducted. At this stage, there were 67 students who became the subjects of the field trial. The field trial data were analyzed as in the limited trial, namely empirical validity, reliability, difficulty level, and test item discrimination. The results of the empirical validation of the test items in the field trial are summarized in Table 10.

Table 10. Results of Item Validity Testing in the Field Test

		Item No. 1	Item No. 2	Score
Item No. 1	Pearson Correlation	1	.693**	.920**
	Sig. (2-tailed)		0.000	0.000
	N	67	67	67
Item No. 2	Pearson	.693**	1	.920**

		Item No. 1	Item No. 2	Score
	Correlation			
	Sig. (2-tailed)	0.000		0.000
	N	67	67	67
Score	Pearson Correlation	.920**	.920**	1
	Sig. (2-tailed)	0.000	0.000	
	N	67	67	67

In Table 10, the significance values for items 1 and 2 are $0{,}000 < 0{,}05$. The $r_{calculated}$value for each test item is 0.920. Meanwhile, the $r_{(table)}$value with $N = 67$ and a significance level of 5% is 0.240. Thus, $r_{(calculated)}$is greater than $r_{(table)}$for both items 1 and 2. Therefore, it can be concluded that both test items meet the validity criteria. The results of the test item reliability test in the field trial stage are presented in Table 11.

Table 11. Test Item Reliability Test Results in the Field Trial

Cronbach's Alpha	Number of Items
0.819	2

Table 11 shows that the *Cronbach's Alpha* coefficient of the test questions after the field trial was 0.819 ($> 0{,}6$). This means that the revised and tested test questions have good reliability. The results of the analysis of the difficulty level of each test question in the field trial are presented in Table 12.

Table 12. Results of the Analysis of Test Item Difficulty Levels in the Field Test

Item Number	Difficulty Index (p)	Difficulty Level
1	0.59	Moderate
2	0.30	Difficult

Table 12 shows that the two items have different difficulty indices. Item 1 is classified as moderate, while item 2 is classified as difficult. The revised test items that were tested in the field trial had a more varied level of difficulty than those in the limited trial. The results of the analysis of the discriminating power of test items in the field trial are presented in Table 13.

Table 13. Results of the Analysis of Item Discrimination in the Field Test

Question Number	Discrimination Index (DB)	Item Discrimination Category
1	0.41	Good
2	0.70	Good

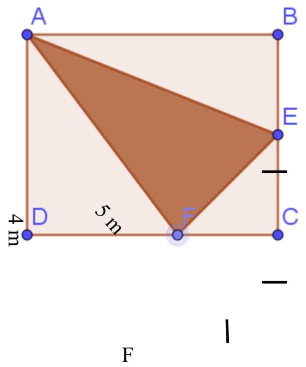
Based on Table 13, both test items developed have an item discrimination index of $\geq 0{,}40$, meaning that both test items have good discriminating power. Based on the test item discrimination power category in Table 3, both test items can be accepted without modification. Thus, the final product of the mathematical problem-solving test on the Pythagorean theorem can be used for wider testing. In general, the statistics of the final product analysis of the mathematical problem-solving test based on field trials are summarized in Table 14.

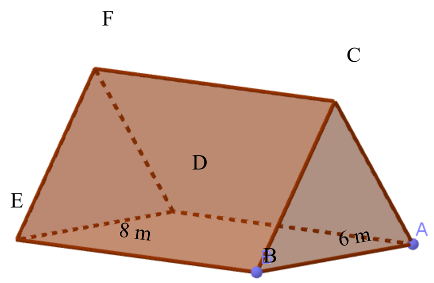
Table 14. Statistics of the Analysis Results of the Final Product of the Mathematical Problem-Solving Ability Test

Statistical Scale	Value
Mean	44.70
Standard Deviation	23.86
Median	40.00
Reliability Coefficient	0.819
Average Difficulty Level	0.445
Average Discrimination Power	0.555

In Table 14, the average test score of students is 44.70 and the average difficulty index of the test questions is 0.445. This means that the difficulty level of the test instrument developed is in the moderate category. Meanwhile, the average discrimination index or discrimination power of the questions is 0.555 ($\geq$ 0,40), which means that the test questions have good discrimination power. In other words, the developed mathematical problem-solving test questions are able to distinguish well between students who have mastered and those who have not mastered the Pythagorean material. Based on the results of the discrimination test, it was also concluded that the test instruments developed were able to distinguish well between students with good mathematical problem-solving abilities and students with poor mathematical problem-solving abilities. Based on all of the above analysis results, it can be concluded that the mathematical problem-solving test questions on the Pythagorean theorem that were developed meet the criteria for a good test and can be used widely. The final product of the mathematical problem-solving test that was developed is shown in Table 15.

Table 15. Mathematical Problem-Solving Ability Test Questions on the Pythagorean Theorem

<table>
<tr><th>Question</th></tr>
<tr><td>The following image is an illustration of a stage used in an art performance at SMP Negeri 1 Pakong.


The stage is rectangular in shape. The part indicated by the shape AFE is the stage floor covered with brown carpet. Determine the area of the brown carpet used!</td></tr>
<tr><td>Wildy drew the roof frame of his house as shown in the following picture. The front and back are isosceles triangles, while the sides are rectangles.


The roof frame of Wildy's house is made of wood. If the wood is indicated by the lines in the picture and the roof height is 4 m, determine the total length of wood needed!</td></tr>
</table>

Mathematical problem-solving skills are an important component of mathematics learning (Chamot et al., 1992; Guven & Cabakcor, 2013; NCTM, 2000; Widdah & Setiawan, 2023). Good mathematical problem-solving skills are not acquired suddenly, but require practice and training (Amhar et al., 2020). This practice can be realized by regularly giving mathematical problem-solving test questions. However, the test questions used need to be compiled and developed in such a way that they are truly in line with the learning objectives and the objectives for which the questions were developed (Faisal, 2022; Magdalena et al., 2020; Secolsky & Denison, 2012).

An evaluation tool, including test questions, can be considered good if it meets several criteria, two of which are validity and reliability (Magdalena et al., 2020; Secolsky & Denison, 2012). Validity and reliability tests are very important for assessing the credibility and quality of a test question (Lecompte & Bear, 2017). The higher the level of validity and reliability of an evaluation tool, such as a test question, the lower the level of measurement error in the test question (Dassopoulos et al., 2007; Lecompte & Bear, 2017). Therefore (, before a test item can be used widely, it must be tested for validity and reliability (Magdalena et al., 2020; Secolsky & Denison, 2012). Validity testing of a developed test consists of theoretical and empirical validity. Theoretical validity is measured through expert review (Hamdi et al., 2022; Putra & Zarita, 2022). Meanwhile, empirical validity is measured based on test data (Hamdi et al., 2022; Lecompte & Bear, 2017; Putra & Zarita, 2022). The results of the validity (theoretical and empirical) and reliability tests in the field trial show that the test questions on mathematical problem-solving skills in the Pythagorean theorem material that were developed are valid and reliable. In other words, the developed product has met two indicators to be considered a good evaluation tool and meets the criteria for product development feasibility (Lecompte & Bear, 2017; Secolsky & Denison, 2012).

The results of the field trial show that the difficulty level of the test items is moderate to difficult. The test instrument developed contains two items with different and proportional difficulty levels. The balance of difficulty levels of test items is important to consider when developing a test instrument (Aiken, 1989; Magdalena et al., 2021; Misri et al., 2021). A test that contains items with the same level of difficulty in each item is considered poor (Magdalena et al., 2021; Misri et al., 2021; Nurhalimah et al., 2022). The test results also show that the test instrument developed does not contain items with easy criteria. Test items that fall under the easy criterion tend not to stimulate students' mathematical problem-solving abilities (Magdalena et al., 2021; Nurhalimah et al., 2022). Therefore, the test instrument developed has met one of the criteria for a good test based on its level of difficulty.

The final product developed has good discriminatory power. The discrimination index of the two test items$\geq 0{,}40$, which means that it meets the criteria for being good. This shows that the two test items are able to effectively distinguish between students who have mastered the material and those who have not (Magdalena et al., 2021; Secolsky & Denison, 2012; Zarokanellou et al., 2023). The discrimination index value for item number 2 is higher than the discrimination index for item number 1. The higher the discrimination index of an item, the more valid and reliable the test score will be (Secolsky & Denison, 2012; Zarokanellou et al., 2023). In other words, the higher the discrimination index of an item, the better the item developed (Clauser & Hambleton, 2011; Magdalena et al., 2021). This means that even though both test items have good discriminatory power, item number 2 is better able to distinguish or "discriminate" between students with good mathematical problem-solving abilities and those with poor abilities.

The purpose of this research and development is to obtain a good test instrument to measure students' mathematical problem-solving abilities on the Pythagorean theorem material. Based on the results presented, the test instrument developed has met the criteria of being valid (suitable for use) theoretically and empirically, reliable, has varying levels of difficulty, and has good discriminating power. Thus, the developed test instrument has met the criteria for a good test and can be used on a wider scale (Fiska et al., 2021; Magdalena et al., 2020, 2021; Secolsky & Denison, 2012; Setiyawan & Wijayanti, 2020).

Acknowledgement

The researchers would like to express their sincere gratitude to all parties who contributed to the completion of this study. Special thanks are extended to the experts who provided valuable input during the validation process, as well as to the eighth-grade students and teachers of the junior high school in Pamekasan Regency for their participation and support during the pilot testing. The researchers also acknowledge the assistance of colleagues and the school administration, whose cooperation and encouragement greatly facilitated the development of the test instrument. Their contributions were essential to the successful completion of this research.

References

Aiken, L. R. Jr. (1989). Assessment of Intellectual Functioning. New York: Plenum Press. https://api.semanticscholar.org/CorpusID:142525255

Alfarizi, M., & Fuady, A. (2020). Analysis of the Effectiveness of STEM-Based Learning in Improving Mathematical Conceptual Understanding. Journal of Mathematics Education, 7(2), 173–184.http://journal.stkip-andi-matappa.ac.id/index.php/histogram/index

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

Setiyawan, R. A., & Wijayanti, P. S. (2020). Analysis of the Quality of Instruments for Measuring Students' Problem-Solving Abilities During Online Learning in the Pandemic. Lebesgue Journal: Scientific Journal of Mathematics Education, Mathematics, and Statistics, 1(2), 130–139. https://doi.org/10.46306/lb.v1i2

Annisah, S., Boeriswati, E., Wildaniati, Y., & Supriatin, A. (2020). Test Instrument Development of Mathematical Problem Solving Skills. International Journal of Advanced Science and Technology, 29(6), 1483–1492. http://digilib.iain-palangkaraya.ac.id/id/eprint/3087

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

Arikunto, S. (2020). Fundamentals of Educational Evaluation. Jakarta: Bumi Aksara.

Badrulaini. (2018). The Relationship between Problem-Solving Skills and Learning Independence on Students' Mathematics Learning Outcomes. Tambusai Education Journal, 2(2), 847–855.

Basuki, I., & Hariyanto, H. (2014). Learning Assessment. Bandung: PT. Remaja Rosdakarya.

Borg, W. R., & Gall, M. D. (1983). Educational Research: An Introduction (4th ed.). Longman.

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, 07, 2646–2659. https://doi.org/10.31004/cendekia.v7i1.1944

Chamot, A. U., Dale, M., Michael O'malley, J., & Spanos, G. A. (1992). Learning and Problem Solving Strategies of ESL Students. Bilingual Research Journal, 16(3-4), 1-28 https://doi.org/10.1080/15235882.1992.10162635

Clauser, J. C., & Hambleton, R. K. (2011). Item Analysis Procedures for Classroom Assessments in Higher Education. https://api.semanticscholar.org/CorpusID:148008416

Dassopoulos, T., Nguyen, G. C., Bitton, A., Bromfield, G. P., Schumm, P. L., Wu, Y., Elkadri, A., Regueiro, M., Siemanowski, B., & Torres, E. A. (2007). Assessment of Reliability and Validity of IBD Phenotyping within the National Institutes of Diabetes and Digestive and

Kidney Diseases (NIDDK) IBD Genetics Consortium (IBDGC). Inflammatory Bowel Diseases, 13(8), 975–983. https://doi.org/10.1002/ibd.20144

DeVellis, R. F. (2006). Classical Test Theory. Medical Care, 44(11). https://journals.lww.com/lww-medicalcare/fulltext/2006/11001/classical_test_theory.11.aspx Faisal, A. (2022). Test Administration in Learning Evaluation. Darussalam Journal, 23(1). http://dx.doi.org/10.58791/drs.v23i1.175

Fiska, J. M., Hidayati, Y., Qomaria, N., & Hadi, W. P. (2021). Analysis of Daily Science Test Items Using ANATES Software in the Classical Test Theory Approach. Journal of Natural Science Educational Research, 4(1), 65-76. https://doi.org/10.21107/nser.v4i1.8133

Guven, B., & Cabakcor, B. O. (2013). Factors Influencing Mathematical Problem-Solving Achievement of Seventh Grade Turkish Students. Learning and Individual Differences, 23, 131–137. https://doi.org/https://doi.org/10.1016/j.lindif.2012.10.003

Hamdi, H., Muchsin, M., & Saiful, M. (2022). Development of Kinematics Teaching Materials to Improve the Multimodal Representation Skills of Prospective Physics Teachers. Jurnal Real Riset, 4(2), 2774–7263. https://doi.org/10.47647/jrr

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

Hegarty, M., Mayer, R. E., & Monk, C. A. (1995). Comprehension of Arithmetic Word Problems: a comparison of successful and unsuccessful problem solvers. Journal of Educational Psychology, 87(1), 18–31. https://doi.org/10.1037/0022-0663.87.1.18

Lecompte, M. D., & Bear, D. E. A.-B. (2017). Revisiting Reliability and Validity in Higher Education Research and Program Evaluation.https://api.semanticscholar.org/CorpusID:114404146

Lukman, H. S., Setiani, A., & Agustiani, N. (2023). Development of Mathematical Problem-Solving Ability Test Instruments Based on Krulik and Rudnick's Theory: Content Validity Analysis. Cendekia Journal: Journal of Mathematics Education, 7(1), 326–339. https://doi.org/10.31004/cendekia.v7i1.1761

Magdalena, I., Fauziah, S. N., Faziah, S. N., & Nupus, F. S. (2021). Analysis of Validity, Reliability, Difficulty Level, and Discrimination Index of Final Exam Questions for Theme 7, Grade III, SDN Karet 1 Sepatan. BINTANG: Journal of Education and Science, 3(2). https://ejournal.stitpn.ac.id/index.php/bintang

Magdalena, I., Hifziyah2 , M., Aeni3 , V. N., & Rahayu, R. P. (2020). Development of Student Test Instruments for Elementary School Level in Tangerang Regency. In Journal of Education and Social Sciences, 2(2). https://ejournal.stitpn.ac.id/index.php/nusantara

Magno, C. (2009). Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data. In The International Journal of Educational and Psychological Assessment, 1(1). https://ssrn.com/abstract=1426043

Misri, M. A., Saifuddin, S., Akbar, R. O., & Kamelia, N. R. (2021). Development and Evaluation of a HOTS-based Test for Matrix Topic: A classical test and item response theory. Beta: Journal of Mathematics Education, 14(2). https://doi.org/10.20414/betajtm.v14i2.383

Musa, M. R., Ikhsan, M., & Zaura, B. (2017). Improving Critical Thinking Skills through the Application of the Van Hiele Theory-Based Learning Model in Grade IX of Banda Aceh Public Junior High School. Scientific Journal of Mathematics Education Students, 2(2).

National Council of Teachers of Mathematics. (2000). Principles and Standards for School Mathematics. National Council of Teachers of Mathematics.

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

Nurhalimah, S., Hidayati, Y., Rosidi, I., Wiwin, D., & Hadi, P. (2022). The Relationship Between Item Validity and Discrimination Power and Difficulty Level of Multiple Choice Questions in PAS. Journal of Natural Science Educational Research, 4(3).

https://doi.org/10.21107/nser.v4i3.8682
Sari, O. P. D., & Faradiba, S. S. (2022). Direct and Indirect Mathematical Dispositions: Characterization of Mathematical Dispositions Based on Gestures. Cendekia Journal: Journal of Mathematics Education.
Polya, G. (1985). How to Solve It: a new aspect of mathematics methods. Princeton University Press.
Putra, A. T. A., & Zarita, R. (2022). Development of Instruments for National and Religious Awareness for Students in Kendari City. Zawiyah: Journal of Islamic Thought, 8(8). https://dx.doi.org/10.31332/zjpi.v8i1.3836
Ratnasari, J. R., & Setiawan, Y. E. (2022). Students' Numeracy Literacy in Solving Quadrilateral and Trapezoid Problems. AKSIOMA: Journal of Mathematics Education Program, 11(3), 2533. https://doi.org/10.24127/ajpm.v11i3.5714
Secolsky, C., & Denison, D. B. (2012). Handbook on Measurement, Assessment, and Evaluation in Higher Education.
Simbolon, S., & Sapri, S. (2022). Analysis of Learning Difficulties of Fourth Grade Students in Flat Shapes at Elementary School. EDUKATIF : Journal of Education, 4(2), 2510–2515. https://doi.org/10.31004/edukatif.v4i2.2081
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, 06(02), 1203–1219. https://doi.org/10.31004/cendekia.v6i2.1344
Tsabitah, I., & Sunismi, S. (2020). Validity of Interactive Digital Worksheets with Augmented Reality on 3-Dimensional Vector Material. Journal of Mathematics Education, 7(2), 124–132. http://journal.stkip-andi-matappa.ac.id/index.php/histogram/index
Verschaffel, L., de Corte, E., Lasure, S., van Vaerenbergh, G., Bogaerts, H., & Ratinckx, E. (1999). Learning to Solve Mathematical Application Problems: A design experiment with fifth graders. Mathematical Thinking and Learning. Mathematical Thinking and Learning, 1(3), 195–229. https://doi.org/10.1207/s15327833mtl0103_2
Wedastuti, N. K., & Faradiba, S. S. (2023). Analysis of Student's Mathematical Thinking in Solving Hots Problems Viewed From Mathematics Anxiety. International Journal of Trends in Mathematics Education Research, 6(1), 33–40. https://doi.org/10.33122/ijtmer.v6i1.173
Wedastuti, N. K., Fuady, A., & Sunismi, S. (2023). Analysis of Final Exam Question Quality Using the Anates Program Application. JIPM (Scientific Journal of Mathematics Education), 12(1), 1–11. https://doi.org/10.25273/jipm.v12i1.15413
Wedastuti, N. K., Sunismi, Nf., & Fuady, A. (2023). Development of PISA Model Literacy-Based E-Modules. Kwangsan: Journal of Educational Technology, 11(1), 459–471. https://doi.org/10.31800/jtp.kw.v11n1.p459--471
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 Study Program, 12(2), 2641 – 2650. https://doi.org/10.24127/ajpm.v12i2.7113
Widodo, S. A., Ibrahim, I., Hidayat, W., Maarif, S., & Sulistyowati, F. (2021). Development of Mathematical Problem-Solving Tests on Geometry for Junior High School Students. Jurnal Elemen, 7(1), 221–231. https://doi.org/10.29408/jel.v7i1.2973
Wijayanti, N. S., Maulana, M., & Isrok'atun, I. (2023). Students' Problem-Solving Skills in the Comprehensive Mathematics Instruction Approach. JNPM (National Journal of Mathematics Education), 7(1), 55–64. https://doi.org/10.33603/jnpm.v7i1.7610
Wulandari, S., Syahbana, A., Tanzimah, T., Shang, Y., Weinhandl, R., & Sharma, R. (2021). Analysis of Students' Thinking Level in Solving Pythagoras' Theorem Problems Based on Van Hiele's Theory. Malikussaleh Journal of Mathematics Learning (MJML), 4(2), 124–130. https://doi.org/10.29103/mjml.v4i2.3905
Zarokanellou, V., Tafiadis, D., Gryparis, A., Prentza, A., Voniati, L., & Ziavra, N. (2023). The Effect of Real Word Stimuli Versus Non-Word Stimuli on Oral Diadochokinetic Rates Across The Life Span: An item discrimination analysis. International Journal of Language and Communication Disorders. https://doi.org/10.1111/1460-6984.12962