Journal of Philosophy and Research in Islamic Mathematics Education
J-PRIMa
Vol. 1, No. 1, 2025, pp 49–62 e-ISSN:
https://doi.org/10.33474/
online at: https://riset.unisma.ac.id/index.php/j-prima/article/view/25135

Measuring mathematical problem-solving ability: a validity and reliability testing scheme for developing quality tests

Irbah Tsabitah[1]

[1]SMA Islam Almaarif Singosari Malang, Indonesia
*Correspondence: IrbahTsabitah@gmail.com

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

Abstract. *Problem-solving skills are important because they help describe students' thinking patterns and influence how they deal with descriptive problems in everyday life. Therefore, in addition to developing students' mathematical problem-solving skills, education practitioners also need to develop an appropriate measurement instrument. The purpose of this study was to produce a valid and reliable test instrument for measuring the problem-solving skills of high school students. The research approach was Research and Development, which adapted the Borg & Gall development model. Seven research steps were taken to develop a test instrument. There were 84 eleventh-grade students at a private high school in Malang Regency as the test sample. Qualitative data was obtained from experts for validity, then the data was analyzed qualitatively and descriptively. Quantitative data was obtained from limited trials and field trials. The trial results data was then analyzed using the SPSS application to determine its empirical validity and reliability. Next, item analysis was conducted using Microsoft Excel to determine the level of difficulty and the discriminating power of each item. The results of this study indicate that the developed problem-solving ability test instrument meets the criteria of validity, reliability, has a varied level of difficulty, and has good discriminating power.*

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

***How to cite**:* Tsabitah, I. (2025). Measuring mathematical problem-solving ability: a validity and reliability testing scheme for developing quality tests. *Journal of Philosophy and Research in Islamic Mathematics Education (J-PRIMa)*, *1*(1), 49–62. Retrieved from https://riset.unisma.ac.id/index.php/j-prima/article/view/25135

Introduction

In education, the assessment process plays a vital role and has many uses (Budiman & Jailani, 2014; Miller & Sternberg, 1983; Sprague, 1939). Assessment results can be used to evaluate student abilities (Miller & Sternberg, 1983). In addition, effective assessment can also help teachers to determine the effectiveness of learning in relation to student abilities, so that teachers can use the assessment results as a reference to develop better teaching and improve student learning outcomes (Wick, 1980). Based on Permendikbud Number 104 of 2014, one of the objectives of conducting learning outcome assessments is to determine whether participants have mastered attitude, knowledge, and skill competencies. If not, then they need to be improved through remedial learning and enrichment programs.

Therefore, it is necessary to develop representative instruments as assessment tools to measure individual abilities and the achievements of all students in the class (Sprague, 1939). The development of a test and self-assessment instrument can also be used by researchers to obtain individual student data such as talents, academic achievements, and personality (Gall et al., 1996a). The information obtained from the test results can help teachers and students make good educational decisions (Schaughency et al., 2012). In developing instruments, there are many activities, such as compiling questions, conducting field tests, and analyzing test results. The participation of subject matter experts and measurement experts is also needed in this process (Knupp & Haris, 2012).

According to Gall et al. (1996), the first step in developing an assessment instrument is to select an interesting construct, such as mathematics learning outcomes, and then determine specific activities for measuring learning outcomes. In this study, the construct selected was students' mathematical problem-solving skills. There are many discussions that emphasize the importance of students' mathematical problem-solving skills. The National Council of Teachers of Mathematics (1980) states that there are five skills that students must have in mathematics learning, including: (1) problem-solving skills, (2) reasoning and proof, (3) mathematical connections, (4) mathematical communication, and (5) mathematical representation.

Problem solving is a fundamental skill that students must have so that they can make decisions in finding solutions to mathematical problems (Arifin in Nisa & Faradiba, 2023). Problem solving plays a central role in mathematics teaching and learning and helps connect mathematics with other fields of study and the physical world (Bassler et al., 1975). Problem-solving skills are important because they help describe students' thinking patterns and influence how they deal with descriptive problems in everyday life (Grouws, 1981; Rochmat & Suparman, 2021). In addition, mathematical problem-solving skills are the fuel for developing other student abilities such as metacognitive abilities and cognitive abilities such as knowledge, understanding, analysis, and evaluation (De-quan, 2003; Rosskopf, 1958).

This study will develop a test instrument that will be used to measure the mathematical problem-solving abilities of high school students. To determine whether students' mathematical problem-solving abilities are "good" or "poor," we must analyze the results of the students' problem-solving ability tests. To conclude that the assessment findings have met these objectives, researchers must consider evidence (test instruments) that supports reliability and validity for specific purposes (AERA, APA, & NCME, 1999). Several other opinions also state the same thing, that a good test instrument has strong reliability and validity results (Gall et al., 1996b; Schaughency et al., 2012).

There is one study that is almost similar to this study, namely the study conducted by Ayu Fitriana (2019), but the test instruments developed in that study were not in the form of contextual problems. Another study was conducted by Forsyth & Spratt (1980), but the mathematical problem-solving test instrument developed was in the form of multiple choice, so not much information could be obtained regarding students' mathematical problem-solving abilities. So far, there has been no research focusing on the development of a descriptive mathematical problem-solving ability instrument that contains contextual problems for high school students on the topic of linear programming. In fact, linear programming is one of the topics in mathematics that is closely related to contextual problems because linear programming is a branch of applied mathematics.

The above explanation forms the background for the researcher in developing a mathematical problem-solving ability test instrument. The result of the development of this test instrument is a valid and reliable linear programming problem-solving ability test instrument. In addition, the final result of the development of this mathematical problem-solving ability instrument can also be used as a summative assessment question in the classroom or as another research instrument to measure students' mathematical problem-solving abilities and interpret the achievement of mathematical problem-solving ability indicators by students.

Method

The purpose of this study was to produce a valid and reliable test instrument for measuring the problem-solving abilities of senior high school students. Therefore, this study focused on developing a test instrument for measuring the problem-solving abilities of senior high school students. The type of research used is Research and Development (RnD) by adapting seven of the ten research and development steps initiated by Borg & Gall (1996). The seven steps of this research are as follows: (1) research and information gathering, (2) planning, (3) initial product development, (4) limited testing, (5) product revision, (6) field testing, and (7) final revision.

The initial step taken by the researcher was to conduct a small study related to the concept and indicators of problem-solving skills, analyze the materials studied in high school that were relevant to problem-solving skills, and determine the form of the test questions to be developed. Next, the researcher began designing and compiling an outline and test questions for problem-solving skills, as well as scoring guidelines. The indicators used as the basis for compiling the outline and test questions are shown in Table 1, namely the problem-solving skill indicators developed by Polya (2004).

Table 1. Polya's Problem-Solving Ability Indicators

Polya's Problem-Solving Stages	**Indicators of Mathematical Problem-Solving Skills**
1. Understanding the problem	Students are able to find the information provided and what is asked in the linear programming question.
2. Planning the solution	Students are able to determine the steps needed to solve the problem and the reasons behind them
3. Implementing the problem-solving plan	Students are able to solve the problem using the steps that have been determined

4. Reviewing the entire solution	Students are able to check the accuracy of the results obtained from each step taken in solving the problem

After the initial instrument has been compiled, the next step is to conduct a limited trial. However, before that, the researcher will conduct a content validation test first. Validation is carried out to justify whether the instruments that have been compiled are appropriate for measuring students' mathematical problem-solving abilities, and also to determine whether the scores obtained can interpret the students' achievement indicators (Allen & Yen, 1979; American Educational Research Association American Psychological Association & National Council on Measurement in Education, 1999; Geisinger et al., 2012; Kerlinger, 1986; Nunally, 1978; Schaughency et al., 2012). This content validation test was conducted by experts and practitioners to obtain evidence of the validity of the test instrument content (Berk, 1990; Delgado-Rico et al., 2012). In this study, the validation test was conducted by two experts and a practitioner with more than 10 years of teaching experience. After the validation test was conducted, the researchers made improvements to the product based on the input and suggestions from the validators.

After the instrument was declared valid, the next step was to conduct a limited trial with 27 students and a field trial with 84 students in grade XI at Almaarif Singosari Islamic High School. The data from the limited trial and field trial will be analyzed quantitatively to determine validity and reliability. An empirical validity test will be conducted to determine whether the developed instrument is of good quality (American Educational Research Association American Psychological Association & National Council on Measurement in Education, 1999; Gall et al., 1996b; Schaughency et al., 2012). After that, the criteria for test items will be determined, including the level of difficulty and discriminating power (Osterlind & Wang, 2012a, 2012b). This is done to determine the quality of the instrument. The entire process in this study is cyclical, in which if the field test results show that the instrument product is not feasible, then revisions, checks by validators, and retesting must be carried out.

In this study, there are two types of data that will be collected, namely qualitative and quantitative data. The instrument used to collect qualitative data is a validation sheet by experts. The validation sheet is used to measure the theoretical validity of the initial test instrument product developed by the researcher. There are three aspects of theoretical validity that will be assessed by experts, including: (1) content, (2) language, and (3) writing. The data from the experts' assessments is qualitative data that will be analyzed using descriptive qualitative methods. There are four criteria for the validity of a test instrument product, namely: (1) suitable for use without revision, (2) suitable for use with minor revisions, (3) suitable for use with major revisions, (4) not suitable for use. The decision on whether the developed instrument is "suitable" or "unsuitable" is based entirely on the assessment results of expert validators and practitioner validators.

The quantitative instrument in this study is a mathematical problem-solving test that has been validated by experts. Quantitative data collection was carried out by conducting limited trials and field trials. The data from the limited and field trials will be tested for empirical validity, reliability, level of difficulty, and discriminating power. To measure the validity of the test instrument product, a *Pearson Bivariate* correlation test will be conducted using *SPSS.* This will be followed by measuring the reliability of the test instrument product by conducting a *Cronbach's Alpha* test using *SPSS.* After that, an analysis of the test items will be conducted to determine the level of difficulty and discriminating power of the test instrument product using *Microsoft Excel.*

The validity criteria for test items are based on the significance level or comparison ofr_{hitung} andr_{Tabel} . If the significance level of *the Pearson Bivariate* Correlation test is < 0.05, then the test items in the test instrument are declared valid. Similarly, if$r_{\text{hitung}} > r_{\text{Tabel}}$, then the test items in the test instrument are declared valid. The reliability criteria for test items are based on a comparison of *Cronbach's Alpha* coefficient withr_{Tabel} . If the *Cronbach's Alpha* coefficient is$> r_{\text{Tabel}}$, then the test items are declared reliable. Furthermore, the criteria for the level of difficulty of test items can be seen in Table 2, and the criteria for discriminating power can be seen in Table 3.

Table 2. Criteria for the Difficulty Level of Test Items

Difficulty Index (p)	**Difficulty Level Criteria**
$p \leq 0{,}30$	Difficult
$0{,}30 < p \leq 0{,}70$	Moderate
$0{,}70 < p \leq 1$	Easy

(Arifin, 2009; Arikunto, 2020)

Table 3. Item Discrimination Criteria for Test Items

Discrimination Index (DP)	**Discrimination Power Criteria**
$DP < 0{,}20$	Poor (Item discarded and replaced with another)
$20 \leq DP0 < 30$	Poor (Requires further discussion, needs improvement, and should be targeted for improvement)
$30 \leq DP0 < 40$	Fairly Good (Accepted with revisions)
$DP \geq 40$	Good (Accepted)

(Basuki & Hariyanto, 2014)

Problem-solving ability test questions of good quality should meet the criteria of validity, reliability, varying levels of difficulty, and have a minimum acceptable level of discriminating power. The final product of a problem-solving ability test instrument that meets these four criteria can be used widely.

Results and Discussion

Based on the results of a small research conducted in the early stages of the study, indicators for developing a problem-solving ability test instrument were obtained in Table 4, which refers to the indicators of mathematical problem-solving ability developed by Polya (2004). Linear programming material was chosen because it is included in the branch of applied mathematics that contains contextual problems. In addition, solving contextual linear programming problems requires more than one solution procedure. The test questions are arranged in such a way that there are w essay questions, each of which contains everyday problems. A sample of the linear programming problem-solving test questions can be seen in Figure 1.

Table 4. Mathematical Problem-Solving Ability Indicators

Polya's Problem-Solving Stages	**Indicators of Mathematical Problem-Solving Ability**
1. Understanding the problem	Students identify and write down the information in the question and the problem presented
2. Planning the solution	Students determine the steps for solving the problem systematically. They begin by stating the contextual problem in terms of a system of inequalities, drawing a graph of the system of inequalities and determining the solution set, determining the points that limit the solution set using the elimination-substitution method, and determining the maximum or minimum value using the substitution method.
3. Implementing the problem-solving plan	Students solve problems according to the designed solution procedure.
4. Double-checking the entire solution	Students check the correctness of the results obtained from each step taken in solving the problem and write down the conclusions from the solution process that has been carried out.

Suatu pesawat udara memiliki tempat duduk tidak lebih dari 50 penumpang. Setiap penumpang kelas utama boleh membawa bagasi 14 Kg, sedangkan untuk kelas ekonomi 6 Kg. Pesawat itu hanya dapat membawa bagasi 420 Kg. Jika harga untuk kelas utama adalah Rp.500.000 per orang, dan kelas ekonomi Rp.300.000 per orang., maka keuntungan maksimum yang dapat diperoleh oleh maskapai tersebut adalah...

Figure 1. Problem-Solving Ability Test Questions

At the planning stage, the researcher also prepared an expert validation sheet as an instrument to measure the theoretical validity of the test items. The expert validation sheet contains three aspects, namely the content of the test items and the language. The series of validation results by each expert can be seen in Table 5 below.

Table 5. Validation Results by Experts

Validator	Validation Results	Recommendations
Expert 1	Valid (suitable for use with minor revisions)	Improvements are needed in the use of inequality symbols in the scoring guidelines
Expert 2	Valid (suitable for use with minor revisions)	Sentences in the questions need to be revised. Test questions should use imperative sentences.
Practitioner 1	Valid (suitable for use with revisions)	The solution area of the inequality system is unclear, so it needs to be revised. Correct *typos*.

The limited trial data analysis in this study includes validity testing, reliability testing, difficulty level testing, and discrimination power testing. First, to determine the validity of the initial test instrument, the researcher conducted a *Pearson bivariate* correlation test using *SPSS*. The validity test results can be seen in Table 6.

Table 6. Results of the Validity Test of the Initial Product of the Problem-Solving Ability Test Questions in Limited Trials

		Item_1	Item_2	Total Score
Item_1	Pearson Correlation	1	.710**	.943**
	Sig. (2-tailed)		.000	.000
	N	27	27	27
Item_2	Pearson Correlation	.710**	1	.904**
	Sig. (2-tailed)	.000		1
	N	27	27	27
Score	Pearson Correlation	.943**	.904	1
	Sig. (2-tailed)	.000	.000	
	N	27	27	27

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

Table 7 shows the significance of the two items, "$0.000 < 0.05$." The value of "$\mathrm{r}_{\text{hitung}}$ " for item 1 is 0.943 and for item 2 is 0.904. Meanwhile, "$\mathrm{r}_{\text{tabel}}$ " at a significance level of 5% with a number of "$N = 27$ " is 0.381. Because "$\mathrm{r}_{\text{hitung}} >$ "$\mathrm{r}_{\text{tabel}}$, it can be concluded that items 1 and 2 have met the validity criteria. Next, the results of the reliability test of the instrument will be presented in Table 4.

Table 7. Results of the Reliability Test of the Initial Product of the Problem-Solving Ability Test in a Limited Trial

Cronbach's Alpha	Number of Items
.815	2

$r_{tabel} = 0{,}381$Table 4 shows that the *Cronbach's Alpha* coefficient for the test questions is 0.815 with a sample size of 2. Since the *Cronbach's Alpha* coefficient is> r_{tabel} , it can be concluded that the problem-solving ability test instrument is reliable or consistent. Next, the results of the difficulty level analysis will be presented in Table 8.

Table 8. Difficulty Level of the Initial Problem-Solving Test Questions in the Limited Trial

Question Number	Difficulty Index	Category
1	0.44	Moderate
2	0.22	Difficult

The data in Table 5 shows that both questions have varying levels of difficulty. The difficulty level of question number 1 falls into the moderate category with a difficulty index **of**$(p) = 0{,}44$ **.** Meanwhile, the difficulty level of question number 2 falls into the difficult category with a difficulty index of$(p) = 0{,}22$. The final item analysis for the limited trial test data is discriminating power. The results of the discriminating power test are presented in Table 9.

Table 9. Discrimination Power of the Initial Test Items for Problem-Solving Ability in the Limited Trial

Item Number	Discrimination Index (DP)	Category
1	0.67	Good
2	0.48	Good

The data in Table 9 shows that both questions have good discriminating power with an item discrimination index of $1(DP) = 0{,}67$ and an item discrimination index of $2(DP) = 0{,}48$. Therefore, both items in the problem-solving ability test instrument can be directly used for the next stage of research, namely field testing. Overall, the statistics of the initial product analysis results in the limited trial phase can be seen in Table 10.

Table 10. Statistics of Initial Product Analysis Results for Problem-Solving Ability Questions in the Field Test

Statistical Scale	Value
Mean	33.19
Standard Deviation	24.192
Median	31ut
Reliability Coefficient	0.815
Average Difficulty Level	0.33
Average Discrimination Power	0.575

Based on the data in Table 10, the average score of students in the limited trial was 33.19. The average difficulty level was 0.33, which means that the difficulty level of the initial test questions was moderate. The average discrimination power was 0.575, which means that the discrimination power of the initial test questions was good and accepted without revision.

Not much different from the limited trial data analysis, the field trial data analysis conducted by the researcher included validity testing, reliability testing, difficulty level testing, and discrimination power testing. First, to determine the validity of the initial test instrument, the researcher conducted a *Pearson bivariate* correlation test using *SPSS*. The validity test results can be seen in Table 11.

Table 11. Results of the Validity Test of the Final Test Items for Problem-Solving Skills in the Field Test

		Item_1	Item_2	Total Score
	Pearson Correlation	1	.710**	.944**
Item_1	Sig. (2-tailed)		.000	.000
	N	84	84	84
	Pearson Correlation	.710**	1	.926**
Item_2	Sig. (2-tailed)	.000		1
	N	84	84	84
	Pearson Correlation	.943**	.904**	1
Score	Sig. (2-tailed)	.000	.000	
	N	84	84	84

Table 7 shows the significance of the two items$0.000 < 0.05$. The value ofr_{hitung} for item 1 is 0.944 and for item 2 is 0.926. Meanwhile,r_{tabel} with a significance level of 5% and a number of$N = 84$ is 0.2120. Because $r_{\text{hitung}} > r_{\text{tabel}}$, it can be concluded that items 1 and 2 have met the validity criteria. Next, the results of the instrument reliability test are presented in Table 12.

Table 12. Reliability Test Results of the Final Product of the Problem-Solving Ability Test in the Field Trial

Cronbach's Alpha	Number of Items
.853	2

$r_{\text{tabel}} = 0{,}2120$Table 12 shows that the *Cronbach's Alpha* coefficient for the test questions is 0.853 with a sample size of 2. Since the *Cronbach's Alpha* coefficient is$> r_{\text{tabel}}$, it can be concluded that the problem-solving ability test instrument is reliable or consistent. Next, the results of the difficulty level analysis will be presented in Table 13.

Table 13. Difficulty Level of the Final Product of the Problem-Solving Ability Test Questions in

the Field Trial

Question Number	Difficulty Index	Category
1	0.44	Moderate
2	0.25	Difficult

The data in Table 13 shows that both questions have varying levels of difficulty. The difficulty level of question number 1 falls into the moderate category with a difficulty index **of**$(p) = 0{,}44$ **.** Meanwhile, the difficulty level of question number 2 falls into the difficult category with a difficulty index of$(p) = 0{,}25$. The final item analysis for the limited trial test data is discriminating power. The results of the discriminating power test are presented in Table 14.

Table 14. Discrimination Power of the Final Product of the Problem-Solving Ability Test Items in the Field Trial

Question Number	Discrimination Index (DP)	Category
1	0.62	Good
2	0.53	Good

The data in Table 14 shows that both questions have good discriminating power with an item discrimination index for question number 1 of$(DP) = 0{,}62$ and an item discrimination index for question number 2 of$(DP) = 0{,}53$. Therefore, both questions in the problem-solving ability test instrument can be directly used for the next stage of research, namely field testing. Overall, the statistics of the initial product analysis results in the limited trial phase can be seen in Table 15.

Table 15. Statistics of the Final Product Analysis Results of Problem-Solving Ability Questions in the Field Test

Statistical Scale	Value
Mean	34.73
Standard Deviation	23.265
Median	30.50
Reliability Coefficient	0.853
Average Difficulty Level	0.345
Average Discrimination Power	0.575

Based on the data in Table 15, the average score of students in the limited trial was 34.73. The average difficulty level was 0.345, which means that the difficulty level of the initial test questions was moderate. The average discrimination power was 0.575, which means that the discrimination power of the initial test questions was good and accepted without revision. From the presentation of the data analysis results above, we can see that the test instrument for problem-solving skills in linear programming material has met the criteria of validity, reliability, with varying levels of difficulty, and good discrimination. A comparison of the results of the analysis of the initial product in the limited trial with the results of the analysis of the final product in the field trial also showed no significant differences. This means that the test instrument for mathematical problem-solving skills in linear programming material is of good quality.

In this study, the test questions developed were in essay form because essay questions give students the freedom to write down their solutions to problems, allowing researchers to analyze the results of students' work to gain a deeper understanding of their mathematical problem-solving abilities and making it easier for researchers to accurately interpret students' problem-solving abilities. The results of this study are supported by previous research by Forsyth & Spratt (1980) that multiple-choice questions in mathematics assessment tend to provide less diagnostic information than essay questions, so there are limitations in measuring students' problem-solving abilities.

In this study, the problem-solving ability test items contained contextual problems. Based on the opinion of Bottge & Hasselbring (1993), contextual mathematical problems can significantly develop students' mathematical problem-solving abilities. The use of contextual problems can help students through the process of mathematization and develop it within themselves (Ani & Rosyidi, 2020).The use of contextual problems in mathematics can provide students with experiences that help them solve mathematical problems, even if they can relate them to everyday life and implement them to solve other problems (Asmah & Rahaju, 2022).

An instrument is considered to be of good quality if it meets the criteria of validity and reliability (Schaughency et al., 2012). There are two validity tests in this study, namely content validity and empirical validity. Content validity testing is very important in the test selection process. If the content validity is weak, the test results will be very inaccurate (Gall et al., 1996). Poggio et al. (1987) revealed that theoretical validity tests (content and instructional aspects) are conducted by experts. In line with this, Gall et al. (1996) revealed that the validity of an instrument's content is proven by experts.

Meanwhile, empirical validity is conducted through direct field trials. In this study, empirical validity was conducted twice, namely a limited trial with 27 students and a field trial with 84 students. The data from the two trials were analyzed to determine whether the developed test instrument could accurately describe abilities in both small and large populations. This is in line with the opinion of LeCompte & Aguilera-Black Bear (2012) that validity is a principle that requires that the instrument used actually measures the intended concept (construct validity), and the results obtained from one population can be generalized to other populations.

The reliability of test instruments also needs to be measured to determine the consistency of a test instrument in conducting measurements. Reliability testing in research helps ensure the consistency, stability, and equivalence of research findings (Jacob, 2019). Reliability needs to be measured to determine the power of significance testing and assumptions about actual variance, error variance, and observed variance (Williams & Zimmerman, 1989). Reliability test results can help teachers estimate measurement errors, thereby helping them make appropriate decisions regarding students (Frisbie, 1988). In addition, test instrument reliability is essential if it is to be used for professional assessment, as test instrument reliability can interpret all possible measurements (Crossley et al., 2002).

In addition to being valid and reliable, an instrument should meet item criteria. Item criteria include difficulty level, discriminating power, and answer distribution (Osterlind & Wang, 2012b, 2012a). In this study, the tests were limited to difficulty level and discriminating power. Answer distribution was not examined because the instrument developed was in the form of essay questions. The results of this study show that the level of difficulty of the questions varied from moderate to difficult. This is in line with the opinion that an effective test instrument for measuring ability or learning outcomes is characterized by a balance in the level of difficulty of the questions (Warju et al., 2020). Meanwhile, the discriminating power of the developed instrument met the criteria for being good and acceptable. This is in line with research conducted that good test items must be able to distinguish between high and low student abilities (Magdalena et al., 2021).

Conclusion

Good quality essay test questions should meet the criteria of validity and reliability. In addition, good quality essay test questions have characteristics such as varying levels of difficulty and good discriminating power. This study went through the 7 research steps developed by Borg & Gall. The results of the study show that the instrument for testing students' mathematical problem-solving abilities falls into the "Valid" and "Reliable" categories. In addition, the results of the study on test item characteristics show that the difficulty level of the developed test varies from moderate to difficult with good discriminating power. Based on the above explanation, the developed test instrument for problem-solving abilities is of good quality

Acknowledgement

The researcher would like to express sincere gratitude to all parties who have supported the completion of this study. Special appreciation is extended to the experts and practitioners who provided valuable insights during the validation process of the test instrument. Their constructive feedback greatly contributed to the refinement and quality of the final product. The researcher also conveys deep thanks to the principal, mathematics teachers, and eleventh-grade students of SMA Islam Almaarif Singosari, Malang, for their cooperation and participation during the limited and field trials. Without their willingness to be involved, this research would not have been successfully carried out.

Lastly, heartfelt thanks are addressed to colleagues, mentors, and everyone who offered encouragement, assistance, and motivation throughout the research and writing process. Their support has been an essential part of completing this work.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Brooks/Cole Publishing Company.

American Educational Research Association American Psychological Association & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*.

Ani, S. I., & Rosyidi, A. H. (2020). Planning Contextual Problem Solving: Guided by Formulas or Context? *Edumatica Journal of Mathematics Education*.

Arifin, Z. (2009). *Learning Evaluation: Principles, Techniques, Procedures*. PT. Remaja Rosdakarya.

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

Asmah, N. F., & Rahaju, E. B. (2022). Contextual Problem-Solving Skills in Linear Programming Material for High School Students with Field Dependent and Field Independent Cognitive Abilities. *Scientific Journal of Mathematics Education*, *11*(3).

Ayu Fitriana, D. (2019). Developing an Assessment Instrument of Mathematical Problem-Solving Skills in Senior High School. *International Journal of Trends in Mathematics Education Research*, *2*(3), 138–141. http://ijtmer.com

Bassler, O., Beers, M., & Richardson, L. (1975). Comparison of Two Instructional Strategies for Teaching the Solution of Verbal Problems. *Journal for Research in Mathematics Education*, *6*, 170–178.

Basuki, I., & Hariyanto, H. (2014). *Learning Assessment.* Remaja Rosdakarya.

Berk, R. A. (1990). Importance of Expert Judgment in Content Related Validity Evidence. *Western Journal of Nursing Research*, *12*(5). https://doi.org/https://doi.org/10.1177/019394599001200507

Bottge, B., & Hasselbring, T. (1993). A Comparison of Two Approaches for Teaching Complex, Authentic Mathematics Problems to Adolescents in Remedial Math Classes. *The Exceptional Child*. https://doi.org/10.1177/001440299305900608

Budiman, A., & Jailani. (2014). Development of Higher Order Thinking Skills (HOTS) Assessment Instruments in Mathematics for Grade VIII Junior High School Semester 1. *Journal of Mathematics Education Research*, *1*(2). https://doi.org/10.21831/jrpm.v1i2.2671

Crossley, J., Davies, H., Humphris, G., & Jolly, B. (2002). Generalisability a key to unlock professional assessment. *Medical Education*. https://doi.org/10.1046/j.1365-2923.2002.01320.x

Delgado-Rico, E., Carretero-Dios, H., & Ruch, W. (2012). Content Validity Evidences in Test Development: An Applied Perspective). *International Journal of Clinical and Health Psychology España*.

De-quan, Z. (2003). An Experimental Study on Meta-cognition Development Based on Problem-solving Study. *Journal of Southwest China Normal University*. https://api.semanticscholar.org/CorpusID:123653858

Forsyth, R. A., & Spratt, K. F. (1980). Measuring Problem Solving Ability in Mathematics with Multiple-Choice Items: The Effect of Item Format on Selected Item and Test Characteristics. *Journal of Education Measurement,* 1. https://doi.org/https://doi.org/10.1111/j.1745-3984.1980.tb00812.x

Frisbie, D. (1988). Reliability of Scores From Teacher-Made Tests. *Educational Measurement: Issues and Practice*.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996a). *Educational research: An introduction (Borg & Gall)* (6th ed.). Longman Publishing.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996b). *Educational research An introduction (Borg & Gall)* (6th ed.). Longman Publishing.

Geisinger, K., Shaw, L. H., & McCormick, C. (2012). The Validation of Tests in Higher Education. In C. Secolsky & D. B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 194–207). Routledge.

Grouws, D. A. (1981). Problem Solving: A Panoramic Approach. *School Science and Mathematics*, *81*(4). https://doi.org/https://doi.org/10.1111/j.1949-8594.1981.tb15487.x

Jacob, J. (2019). Reliability How? When? What? *International Journal of Advances in Nursing*

Management.
Ministry of *Education* and *Culture*. (2014). *Regulation of the Minister of Education and Culture of the Republic of Indonesia Number 104 concerning Assessment of Learning Outcomes by Educators in Primary and Secondary Education.* https://luk.staff.ugm.ac.id/atur/bsnp/Permendikbud104-2014PenilaianHasilBelajar.pdf
Kerlinger, F. N. (1986). *Principles of behavioral research (Translated by L.R. Simatupang)*. Gajahmada University Press. https://doi.org/10.9734/BJESBS/2014/6427
Knupp, T., & Haris, D. J. (2012). Building Content and Statistical Test Specifications. In D. Brian (Ed.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education*. Routledge.
LeCompte, M. D., & Aguilera-Black Bear, D. E. (2012). Revisiting Reliability and Validity in Higher Education Research and Program Evaluation. In C. Secolsky & D. B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (p. 613).
Magdalena, I., Anggraini, I. A., & Khoiriah, S. (2021). Analysis of Discrimination Power and Difficulty Level in Roman Numeral Questions for Grade 4 Students at SDN Tobat 1 Balaraja. In *Journal of Education and Social Sciences* (Vol. 3, Issue 1). https://ejournal.stitpn.ac.id/index.php/nusantara
Miller, T. L., & Sternberg, L. (1983). Measurement in Educational Research. *Exceptional Education Quarterly*, *4*(3), 18–26. https://doi.org/10.1177/074193258300400305
National Council of Teachers of Mathematics. (1980). *An Agenda for Action: Recommendations for School Mathematics of the 1980s.* The Council. https://www.nctm.org/flipbooks/standards/agendaforaction/html5/index.html
Nisa, F., & Faradiba, S. S. (2023). Mathematical Literacy Profile of Students Based on Problem-Solving Ability Levels in PISA Questions. *Jurnal Cendekia: Journal of Mathematics Education*, *7*(2), 1003–1019. https://doi.org/10.31004/cendekia.v7i2.2211
Nunally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw Hill. https://doi.org/https://doi.org/10.1177/014662167900300216
Osterlind, S. J., & Wang, Z. (2012a). Item Response Theory in Measurement, Assessment, and Evaluation for Higher Education. In C. Secolsky & D. B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (p. 150). Routledge.
Osterlind, S. J., & Wang, Z. (2012b). Statistical Modeling in Measurement, Assessment, and Evaluation for Higher Education. In C. Secolsky & D. B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (p. 150). Routledge.
Poggio, J. P., David Miller, M., & Glasnapp, D. R. (1987). The Adequacy of Retrospective Judgments to Establish Instructional Validity. *Educational and Psychological Measurement*, *47*(3). https://doi.org/https://doi.org/10.1177/001316448704700332
Polya, G. (2004). *How to Solve It: a new aspect of mathematical methods.* Princeton University Press.
Rochmat, S., & Suparman, S. (2021). Analysis of E-MODULE Requirements to Improve Students' Mathematical Ability in Solving Problems in Essay Questions. *Journal of Mathematics Education Innovation (JIPM),* 3*(*2), 106–112. https://doi.org/10.37729/jipm.v3i2.1597
Rosskopf, M. F. (1958). The Strategy of Concept Attainment. *Teachers College Record*, *60*(1). https://doi.org/https://doi.org/10.1177/016146815806000103
Schaughency, E., Smith, J. K., Meer, J. van der, & Berg, D. (2012). Classical Test Theory and Higher Education: Five Questions. In C. Secolsky & D. B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 117–131). Routledge.
Sprague, J. B. (1939). Diagnostic Testing to Improve Mathematical Ability in Grade X. *American Journal of Education*, *47*. https://doi.org/https://doi.org/10.1086/440390
Warju, Ariyanto, S. R., Soeryanto, & Trisna, R. A. (2020). Analysis of the Quality of HOTS-Type Items on Students' REM System Competence in Vocational High Schools. *Journal of Technology and Vocational Education*, *17*(1).
Wick, J. W. (1980). Measurement and Evaluation in Education and Psychology. *American Psychological Association*.
Williams, R. H., & Zimmerman, D. W. (1989). Statistical Power Analysis and Reliability of Measurement. *Journal of General Psychology*. https://doi.org/10.1080/00221309.1989.9921123