Edumatica: Jurnal Pendidikan Matematika
Volume 15 Number 2 August 2025

Submited : 22 May 2025
Revised : 17 July 2025
Accepted : 25 July 2025

DOI: https://doi.org/10.22437/edumatica.vl5i2.44361

Analysis of Differential Item Functioning on Students' Mathematical Problem-Solving Ability Assessment Instruments

M. Rais Ridwan[1*], Samsul Hadi[2], Faihatuz Zuhairoh[3]
[1*]STKIP YPUP Makassar, Indonesia
[2]Universitas Negeri Yogyakarta, Indonesia
[3]STKIP YPUP Makassar, Indonesia
E-mail*: mraisridwan@yahoo.com

Abstract

The current problem-solving ability assessment tool lacks specified criteria for fair instrument items to measure students' abilities across genders and grade levels. To ensure score validity, unfair or biased assessment items should be removed from the test instrument. This study analyzed the quality of assessment items for mathematical problem-solving abilities, considering variables of gender and class. This research is an instrument development study with a quantitative descriptive approach. Data collection using an online survey obtained responses from 362 high school students to 10 instrument items. The data analysis technique utilized classical test theory, employing an item functioning analysis approach through the Mantel-Haenszel method. The item validity analysis indicated that not all items were classified as biased among gender groups. Two items were identified as biased according to grade level groups. The reliability estimate was 0.829, indicating a high level of consistency in the results obtained using different samples. The study's findings validate the bias item analysis approach and provide a reliable assessment tool for evaluating high school students' mathematics problem-solving abilities. This study also suggests using modern theoretical analysis to identify item features that distinguish students' ability, guessing levels, and carelessness in test item responses.

Keywords: bias assessment; differential item functioning; Mantel-Haenszel method; problem-solving ability; validity and reliability estimation

INTRODUCTION

Problem-solving is a crucial component of the 21st-century curriculum, as it develops essential skills that students need to thrive in the modern era. These skills are crucial for students to help solve complex problems as the modern era evolves and becomes technology-driven (Ismail et al., 2021). Problem-solving skills are not only crucial for academic success but also help students solve problems in everyday life and the workplace (Astuti, Aziz, Sumarti, & Bharati, 2019; Bahar & Maker, 2015). Problem-solving in realistic contexts enables students to explore and discover knowledge, thereby preparing them for development in a manner relevant to the times (Lim Li Gek, 2020). Problem-solving ability is a cognitive process consisting of identifying, planning, implementing strategies, and evaluating results in solving problems. Problem-solving ability is a domain of mathematical ability (Campbell, 2004; Ridwan et al., 2023a). The problem-solving stage consists of activities to understand the problem, develop strategies, implement plans, and review the results (Fahrudin et al., 2019; Firda et al., 2023; Ningsih & Hidayati, 2022; Parandrengi et al., 2024; Weng et al., 2007). Problem-solving ability is important in mathematics and nursing to apply learned concepts to real-world situations (Choi & Jeon, 2022; Yuristia & Musdi, 2020). Problem-solving skills are a mandatory requirement in recruiting workers in work and industry (Rios et al., 2020) and contribute to improving overall team performance (Scott, 2007).

Mathematical problem-solving ability refers to understanding, planning, implementing, and verifying solutions in solving mathematical problems. Mathematical problem-solving ability is a high-level cognitive ability that applies mathematical concepts to various real-world situations (Apriani et al., 2021; Istiqomah et al., 2019; Ningsih & Hidayati, 2022; Yuristia & Musdi, 2020). Some challenges in solving problems involving mathematical problem-solving skills consist of students' low ability to understand story problems (Apriani et al., 2021; Parandrengi et al., 2024), students' difficulties in working on non-routine problem forms due to deficiencies in working on practice questions (Kusuma & Untarti, 2020; Nasution et al., 2018), and students' lack of ability in working on problem items procedurally, namely understanding or modeling mathematical problems (Apriani et al., 2021). In addition, several factors influencing students' low mathematical problem-solving abilities are motivation, learning independence, self-understanding of mathematics, and using learning models. Higher motivation tends to lead to better performance at all stages of problem-solving solving (Kusuma & Untarti, 2020), higher learning independence indicates better problem-solving skills (Muhtarom et al., 2024; Yasin et al., 2020), and a positive self-understanding concept in mathematics correlates with better problem-solving abilities (Kharisudin & Cahyati, 2020). Then, the use of innovative learning models such as the IMPROVE (Introducing new concepts, Metacognitive questioning, Practicing, Reviewing and Reducing difficulties, Obtaining mastery, Verification, and Enrichment) model and problem-based learning models is efficacious in improving problem-solving abilities compared to conventional methods (Agustinsa et al., 2023; Gozali et al., 2022; Hasanah et al., 2021; Mansyur & Sunendar, 2020; Nasution et al., 2018; Ridwan et al., 2022b, 2022a; Ridwan, Retnawati, et al., 2021; Yasin et al., 2020).

Assessment of mathematical problem-solving ability is currently also an important aspect in determining the effectiveness and practicality of assessments carried out by teachers to improve students' abilities. The relationship between assessment strategies and mathematical problem-solving ability is an important aspect of exploring different assessment methods that affect students' abilities in solving mathematical problems. Like authentic assessments consisting of effective performance rubrics used to evaluate students' abilities (Rosli et al., 2013) and can improve the ability to solve mathematical problems compared to conventional methods (Darma et al., 2018; Firdausi & Supinah, 2021; Kadir, 2023). In addition, self-assessment can also motivate students to actively review their performance by identifying strengths and weaknesses in understanding mathematical problems (Barana et al., 2022). Then, integrating video games into problem-based learning can improve students' mathematical problem-solving abilities (Hwang et al., 2014; Ukobizaba et al., 2021). Formative assessment integrated into learning using the use of GeoGebra Classroom is also effective in improving mathematical problem-solving abilities (Rosyidi et al., 2024).

However, in addition to the assessment strategy, some aspects are also fundamental for teachers to pay attention to: assessment with measuring instrument components that can measure and are consistently used to assess students' mathematical problem-solving abilities. The measuring instrument components are based on the validity and reliability aspects of the mathematical problem-solving ability assessment instrument. Validity is defined as the accuracy of a measuring instrument in measuring what it claims to measure (Greeno, 2003; Ouzouni & Nakakis, 2011). Validity describes that the interpretation and use of test results can be justified and relied on (Linn, 2010; Ryan & DeMark, 2012; Sireci & Soto, 2016). In general, the validity of an assessment instrument is based on content and construct validity. Content validity provides evidence of the extent to which elements of the assessment instrument are relevant and represent constructs aligned with specific assessment objectives (Almanasreh et al., 2019). Meanwhile, construct validity is an assessment of the development of test instruments based on the suitability of the theoretical study construct model with empirical data (Coulacoglou & Saklofske, 2017). Several aspects of validity consist of item quality based on the item difficulty level, items can predict respondents' ability scores based on actual or false responses, and items have distractor answer options for students. The next aspect of validity is that items do not tend to favor male (or female) groups or be based on class levels, as reviewed from the aspect of item difficulty. Another quality aspect is that the test instrument has reliable or consistent characteristics that can be used by different students with the same school characteristics. Reliability refers to the consistency and reproducibility of measurements with assessment measuring instruments producing the same results against different measurement times and sample characteristics (Kaul et al., 2025).

Several studies are relevant to the results of the instrument quality analysis of measuring instruments for assessing students' mathematical problem-solving abilities. Research relevant to the results of research examining the analysis of instrument quality for assessing students' mathematical problem-solving abilities based on content validity and other item validity with a classical theory analysis approach based on difficulty level parameters, item identification of respondent ability scores, and estimation of instrument construct reliability (Ridwan, Istiyono, et al., 2021). The subsequent relevant research study is only based on content validity (Bambang et al., 2018; Kania et al., 2024; Rosyidi et al., 2024; Ulya et al., 2024; Widodo et al., 2021). The following relevant research study reported the results of the study, consisting of content validation, item suitability analysis using a modern theory analysis approach with a response model based on difficulty level parameters (Astuti, Supahar, Mundilarto, & Istiyono, 2020; Jatiningtyas, Kartono, & Mindyarto, 2022; Reffiane, Sudarmin, Wiyanto, & Saptono, 2021). The following relevant study reported the results of the instrument quality analysis also with a modern theory analysis approach to item quality based on the level of difficulty parameters and item discrimination power and the estimation of the instrument's reliability against the mathematical problem-solving assessment measuring instrument (Annisavitri et al., 2020; Sorby et al., 2022). Other validity aspect reports also use content validity, identification of item difficulty levels, and instrument construct validity based on the aspect of the instrument measuring measurable dimensions and estimation of instrument reliability (Fitriana & Supahar, 2019). Other relevant research results reports are based on content validity and construct and reliability analysis of the mathematical problem-solving ability assessment instrument (Wahyuni et al., 2018).

In general, the results of the literature review of relevant research obtained research results by several researchers did not specifically conduct item quality analysis based on the effectiveness of distractors or distractor answer choices and also identification of items that do not tend to benefit gender groups, namely male (or female) or based on class levels reviewed from the aspect of item difficulty level parameters. Identifying distractor effectiveness based on the proportion of test participants choosing distractors compared to the total test participants. Then, identify biased items using the differential item functioning (DIF) analysis approach using the Mantel-Haenszel method. By definition, DIF indicates that test participants in different groups with the same ability have different probabilities of answering certain test instrument items correctly (Gamerman et al., 2018; Yüksel et al., 2019). Such conditions indicate that the item is likely biased or unfair to specific groups. DIF identification is essential in test development to ensure the measuring instrument measures students' abilities in all groups. In addition, DIF analysis results are very much needed in the scope of education, psychology, social sciences, and health, which have significant implications for test results (Hidalgo & Gómez-

Benito, 2010; Runnels, 2013). The results of the item quality analysis by identifying answer choices other than the answer key aim to determine the construction of answer choices designed by researchers or test makers, chosen by test participants for specific reasons. Identification of subsequent items based on information on biased items with effects that can affect the test participant's ability score, so that the items are revised or eliminated and cannot be used for assessing mathematical problem-solving abilities. The novelty of this research provides information on the validation of a mathematical problem-solving ability assessment instrument using an analysis approach to item functional differences, aiming to obtain a standardized assessment measurement tool. The validation aspect of the assessment instrument has fair item criteria. It does not tend to favour one group of test participants or students over another based on gender or grade level differences.

The purpose of this study is to analyze the quality of the mathematical problem-solving ability assessment instrument items using the classical test theory analysis approach consisting of identifying the level of item difficulty, valid items measuring students' ability scores, distractor effectiveness, identifying biased items based on gender group and grade level responses, and estimating instrument reliability. The research questions (RQ) are as follows.

RQ1. How is the difficulty level of the mathematical problem-solving ability assessment instrument items mapped?

RQ2. How many valid items measure students' mathematical problem-solving ability scores?

RQ3. How many mathematical problem-solving ability assessment items have effective distractors for students to choose from?

RQ4. How many biased items measure students' mathematical problem-solving ability based on gender group and grade level responses?

RQ5. How does the assessment instrument's consistency level produce the same results against different measurement times and sample characteristics?

METHOD

Research design

The quantitative research design using the classical test theory approach consists of an analysis of the quality of instrument items based on the identification of item difficulty levels, items predicting respondent ability scores, the effectiveness of distractor answer choices, and the identification of biased items based on differences in item responses to gender groups and grade levels. The use of a classical test theory analysis approach takes into account the DIF method, namely the Mantel-Haenszel method. The DIF method is a classical approach with uniform effects (i.e., effects that do not consider ability level based on group membership), and the number of groups is equal to two (Magis et al., 2010; Mellenbergh, 1982). The classical test theory (CTT) approach is a measurement theory that focuses on the relationship between observed test scores and measurement error. The observed test score is the sum of the actual score (the true state of the unobservable variable) and the error score (random effects on the observed variable) (Brown, 2013; Steyer, 2015). The limitation of using CTT analysis is that the concept does not consider various sources of error, assuming that all measurement errors are random and uncorrelated with the actual score (Grabowski & Lin, 2019). However, using CTT for instruments with a larger number of items tends to have a higher level of consistency because it calculates the average random error more effectively (Algina & Swaminathan, 2015).

Instruments

The research instrument uses indicators for assessing mathematical problem-solving abilities with a test instrument length of 10 questions. The indicators for assessing mathematical problem-solving abilities consist of identifying and understanding problems, formulating and planning, implementing plans, and reviewing and reflecting. The four indicators are stages of problem solving following Polya's steps, namely the ability to understand problems, planning, implementation, and evaluation (Fahrudin et al., 2019; Ningsih & Hidayati, 2022; Parandrengi et al., 2024). The ability to identify and understand known and asked aspects is a sufficient requirement for solving mathematical problems (Mustofa et al.,

2020; Rakhmawati et al., 2019; Yuristia & Musdi, 2020). Formulating problems or constructing mathematical models requires constructing plans or strategies to solve mathematical problems (Mustofa et al., 2020; Yuristia & Musdi, 2020). The plan's design consists of accurate mathematical manipulation and calculations, which are procedures for solving mathematical problems (Mustofa et al., 2020; Rakhmawati et al., 2019; Yuristia & Musdi, 2020). The ability to review and reflect is the process of mathematical problem-solving activities by re-examining the solution and reflecting on the process that aims to verify the correctness of the solution (Mustofa et al., 2020; Permata et al., 2018; Rakhmawati et al., 2019; Yuristia & Musdi, 2020). The research instrument items were constructed based on the mathematical problem-solving indicators in Table 1. Five expert lecturers in mathematics education validated the ten problem-solving ability assessment instrument items (3 raters with doctoral degrees and two others with master's degrees). The Aiken validity results for each item have an Aiken index at the interval of 0.870 and 1 so that the items meet the valid criteria in terms of content, indicating that each item defines the assessment of the item indicator. The item validity criteria use the Aiken index table with a total of five raters, and the assessment category consists of 4 scores and a significant value of 0.050, which meets the minimum value of 0.870 (Aiken, 1985).

Table 1. *Construction of Mathematical Problem-Solving Ability Assessment Items*

Items	Indicators
P-SA1.1	Ability to understand problems and planning related to linear programming material.
P-SA1.2	Ability to understand problems, planning, implementation, and evaluation related to linear programming material.
P-SA2.3	Ability to understand problems, planning, implementation, and evaluation related to the opportunity material.
P-SA2.4	Ability to understand problems, planning, implementation, and evaluation related to the opportunity material.
P-SA2.5	Ability to understand problems, planning, implementation, and evaluation related to the opportunity material.
P-SA3.6	The ability to understand problems, planning, implementation, and evaluation related to the composition function material.
P-SA3.7	The ability to understand problems, planning, implementation, and evaluation related to the composition function material.
P-SA4.8	The ability to understand problems, planning, implementation, and evaluation related to linear equation material.
P-SA5.9	The ability to understand problems, planning, implementation, and evaluation related to the material on circle equations.
P-SA6.10	The ability to understand problems, planning, implementation, and evaluation related to the material on determining the area of a triangle.

Participants

Data collection for the trial of 10 items of mathematical problem-solving ability test instruments using an online survey of 362 students. The characteristics of the research sample consisted of 230 11th-grade students (94 male students, 40.87%; 136 female students, 59.13%) and 132 12th-grade students (78 male students, 59.09%; 54 female students, 40.91%). The determination of the research sample used purposive random sampling, namely, based on the objectives, and was carried out randomly. The purposive random sampling technique is a sampling technique based on consideration or selection (Andrade, 2020; Guarte & and Barrios, 2006) with a sampling technique that selects participants based on specific characteristics that are relevant to the research objectives (Williamson, 2018). The consideration of research sampling is based on the learning of problem-solving material by 11th and 12th-grade students, as well as gender differences. This technique ensures that the research sample meets the characteristics of the respondents to increase the relevance and quality of the collected data (Memon et al., 2025).

Data analysis

Data analysis using the response data of the 10-item trial instruments of 362 students. The item response data is dichotomous, meaning that each item has an actual value (response of 1) and an incorrect value (response of 0). The data analysis technique employs a classical test theory approach, which involves estimating item validity and instrument reliability. Analysis of instrument quality based on item validity and reliability estimation using the R Studio program. Aspects of item validity consist of identifying item difficulty, items predicting respondents' ability scores, the effectiveness of distractor answer choices, and identifying biased items based on differences in item responses to gender groups and class levels. The level of item difficulty uses the criteria for an item difficulty index greater than 0.700 (easy), greater than 0.300 and less than 0.700 (moderate), and an index less than 0.300 (difficult) (Allen & Yen, 1979). Then, the item criteria predict respondents' ability scores using a biserial point coefficient greater than or equal to 0.250 (Douglas-Morris et al., 2021; Varma, 2006). However, suppose the biserial point coefficient value is negative or less than 0.250. In that case, the item shows the test participant's response that answers correctly (wrongly) will get a low (high) total score. The following criterion is based on identifying the effectiveness of distractor answer choices using the distractor index criterion, which is greater than or equal to 0.020. Items with a distractor index less than 0.020 provide a condition where the distractor is not functioning correctly (Fernandes, 1984). Distractors with this index are revised and replaced by considering the distractor construction criteria that allow for reasons. Other item validity uses bias item analysis based on differences in gender group responses and class levels. Bias item analysis with DIF analysis uses a classical theory approach, namely the Mantel-Haenszel (MH) method with uniform effects and a purification process. The bias item criteria are based on a contrast DIF value greater than 0.640 and a significance value (p-value) less than 0.050 (Paek & Holland, 2015; Ridwan et al., 2023b). Next, identify the effects of DIF items using the ΔMH value criteria with significant effects (|ΔMH| values greater than or equal to 1,500), moderate effects (|ΔMH| values greater than or equal to 1,000 and less than 1,500), and small effects (|ΔMH| values less than 1,000) (Holland & Thayer, 1988; Zieky, 1993). Another aspect of instrument quality is reliability estimation, which requires a high level of consistency, as indicated by a Cronbach's Alpha coefficient of at least 0.700 (Sarstedt & Mooi, 2019).

RESULTS

The results of the analysis of the trial responses of 362 students to 10 items of the mathematical problem-solving ability test instrument using the classical theory approach consist of item quality based on the level of difficulty, identification of predicted respondent ability scores, and the effectiveness of choices other than the answer to the question to be selected by the test participants. The following item quality analysis results identify biased items using differences in item responses based on gender groups and class levels. The results of other instrument quality analyses are based on the level of consistency in the use of research samples.

Item difficulty level, identification of predicted respondent ability scores, and distractor effectiveness

The level of item difficulty describes the ability of test participants to answer correctly or incorrectly to an instrument item. Grouping the item difficulty level is based on the proportion of respondents giving correct responses to an item to the total number of respondents. Then, to identify the predicted score of respondents' ability using the biserial point coefficient criterion, the effectiveness of the distractor is based on the proportion of respondents responding to items other than the answer key to the total number of respondents. The results of item analysis using the classical test theory approach, based on the difficulty level, identification of the predicted score of respondents' ability, and the effectiveness of the distractor, are given in Table 2.

Table 2. *Results of the Analysis of the Quality of Problem-Solving Ability Assessment Items*

Items	Choices and answer keys	Difficulty index	Number of distractor responses	Point biserial coefficient	Decisions
P-SA1.1	A	-	0.171	-	

Items	Choices and answer keys	Difficulty index	Number of distractor responses	Point biserial coefficient	Decisions
	B*	0.638	-	0.505	The five items with a moderate difficulty level were valid in predicting respondents' ability scores, and all distractors functioned well.
	C	-	0.118	-	
	D	-	0.044	-	
	E	-	0.027	-	
P-SA1.2	A	-	0.133	-	
	B	-	0.138	-	
	C*	0.555	-	0.620	
	D	-	0.083	-	
	E	-	0.091	-	
P-SA2.3	A	-	0.097	-	
	B*	0.649	-	0.605	
	C	-	0.127	-	
	D	-	0.091	-	
	E	-	0.036	-	
P-SA2.4	A	-	0.166	-	
	B	-	0.113	-	
	C	-	0.102	-	
	D	-	0.091	-	
	E*	0.528	-	0.643	
P-SA2.5	A	-	0.124	-	
	B	-	0.138	-	
	C	-	0.135	-	
	D	-	0.130	-	
	E*	0.472	-	0.541	
P-SA3.6	A	-	0.169	-	Items with difficulty are valid in predicting respondents' ability scores, and distractors function well.
	B	-	0.152	-	
	C	-	0.166	-	
	D	-	0.273	-	
	E*	0.240	-	0.297	
P-SA3.7	A	-	0.127	-	The four items had a moderate difficulty level, were valid in predicting respondents' ability scores, and the distractors functioned well.
	B*	0.605	-	0.543	
	C	-	0.130	-	
	D	-	0.091	-	
	E	-	0.047	-	
P-SA4.8	A	-	0.196	-	
	B	-	0.229	-	
	C	-	0.180	-	
	D*	0.312	-	0.260	
	E	-	0.083	-	
P-SA5.9	A*	0.541	-	0.460	
	B	-	0.138	-	
	C	-	0.124	-	
	D	-	0.138	-	
	E	-	0.058	-	
P-SA6.10	A	-	0.146	-	
	B	-	0.165	-	
	C*	0.500	-	0.680	
	D	-	0.116	-	
	E	-	0.072	-	

(*Source: Own analysis*).

The findings from the item quality analysis in Table 2, derived from the percentage of test participants who answered correctly 10 questions of the mathematical problem-solving skill assessment, indicated that one item fell inside the difficult group. In contrast, the remaining items were classified as

moderate. Item P-SA3.6 has a difficulty index of 0.240, which is smaller than 0.300, which means that the proportion of test participants who answered correctly was only 24.00%, which is smaller than 30.00% or as many as 87 test participants who answered correctly out of 362 students. The other nine items are in the moderate category with a difficulty index greater than 0.300 and less than 0.700. In this case, test participants who answered correctly for the nine items have a proportion greater than 30.00% and less than 70.00%. The following item quality analysis results, namely identifying valid items predicting test participants' ability scores, showed that all items have a biserial point coefficient value greater than 0.250. The ten items identify the test taker's response that answers correctly (or incorrectly), which will get a high (or low) score, so the item provides an assessment according to the correct and incorrect answer responses. The results of other item quality analyses by identifying the effectiveness of distractors or choices other than the answer key obtained all items with distractors having an index greater than 0.020. These items have well-functioning distractors that pay attention to the construction criteria for answer choices that allow for reasons and can be selected.

Determining the validity of differential item functioning

The results of the evaluation of the quality of item assessment of the ability to solve the next mathematical problem identify whether the item is fair or biased by differences in responding to items based on gender groups and class levels.

Table 3. *Results of Analysis of Bias Based on Gender Groups*

Items	DIF contrast value	Significance value	DIF effect coefficient	Decisions
P-SA1.1	0.003	0.960	-0.141	The four items are not biased based on the two groups of students, with a small DIF effect that benefits the group of female students.
P-SA1.2	1.083	0.298	-0.901	
P-SA2.3	0.102	0.749	-0.370	
P-SA2.4	0.023	0.879	-0.268	
P-SA2.5	0.869	0.351	0.750	Items are not biased based on the two groups of students, with a small DIF effect that benefits the group of male students.
P-SA3.6	2.621	0.106	1.490	Items are not biased based on the two groups of students with a moderate DIF effect that benefits the group of male students.
P-SA3.7	1.704	0.192	-1.039	Items are not biased based on the two groups of students with a moderate DIF effect that benefits a group of female students.
P-SA4.8	0.167	0.683	0.347	Items are not biased based on the two groups of students, with a small DIF effect that benefits the group of male students.
P-SA5.9	0.307	0.580	-0.451	Items are not biased based on the two groups of students, with a small DIF effect that benefits a group of female students.

Items	DIF contrast value	Significance value	DIF effect coefficient	Decisions
P-SA6.10	0.194	0.659	0.532	Items are not biased based on both groups, with a small DIF effect that benefits the group of male students.

(*Source: Own analysis*).

The results of grain quality analysis using DIF with the Mantel-Haenszel method obtained the characteristics of refractive grains based on gender groups and class levels in Tables 3 and 4.

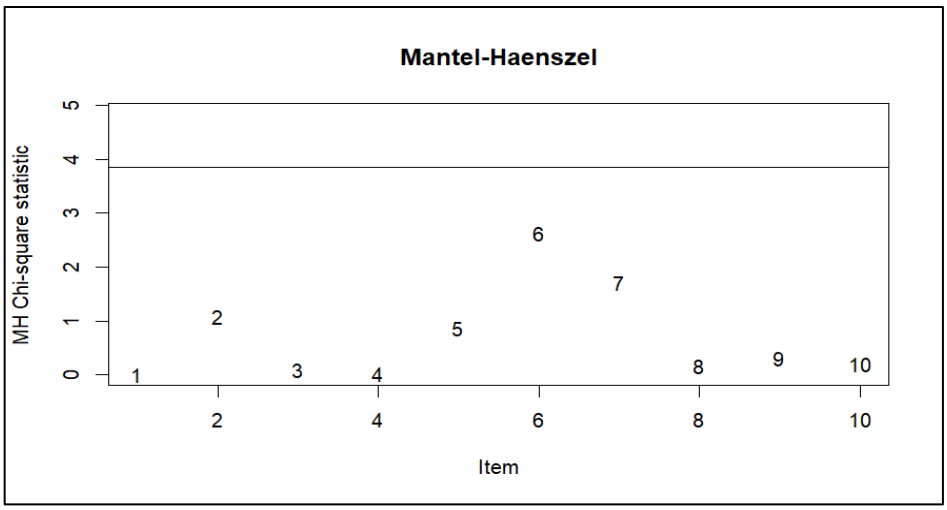
Table 4. *The Results of an Analysis of Bias Items Based on Class-Level Groups*

Items	DIF contrast value	Significance value	DIF effect coefficient	Decisions
P-SA1.1	0.004	0.951	0.062	Items are not biased based on the two groups of students, with a small DIF effect that benefits the group of 12th-grade students.
P-SA1.2	0.010	0.921	-0.060	Items are not biased based on the two groups of students, with a small DIF effect that benefits the group of students in grade 11.
P-SA2.3	0.072	0.788	0.365	Items are not biased based on the two groups of students, with a small DIF effect that benefits the group of 12th-grade students.
P-SA2.4	0.262	0.609	-0.567	Items are not biased based on the two groups of students, with a small DIF effect that benefits the group of students in grade 11.
P-SA2.5	0.008	0.928	0.046	The three items are not biased based on the two groups of students, with a small DIF effect that benefits the 12th-grade student group.
P-SA3.6	0.610	0.435	0.772	
P-SA3.7	0.004	0.948	0.063	
P-SA4.8*	6.234	0.013	1.740	Both items are identified based on the two groups of students with a significant DIF effect that benefits the 12th-grade student group.
P-SA5.9*	6.451	0.011	1.722	
P-SA6.20	0.644	0.422	-0.923	Items are not biased based on the two groups of students, with a small DIF effect that benefits the group of students in grade 11.

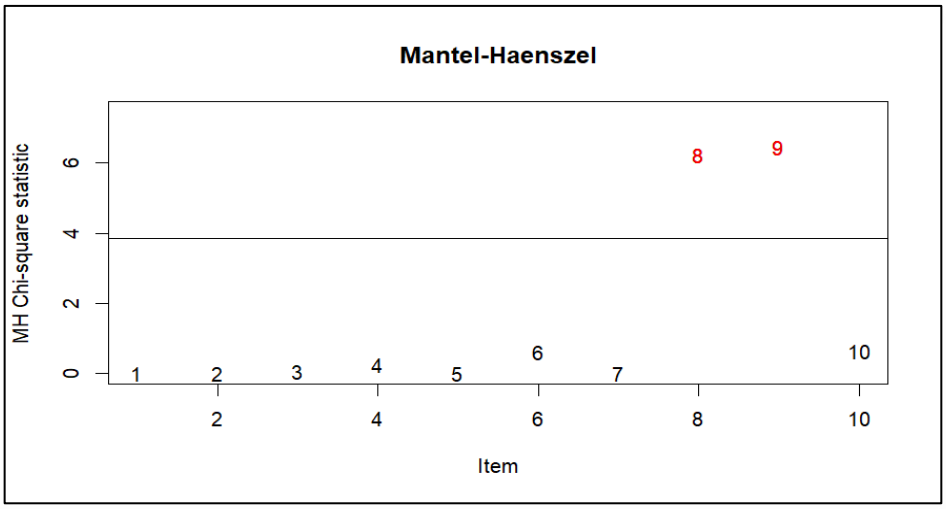
(*Source: Own analysis*).

The results of the DIF analysis are based on gender group responses and class levels in Tables 3 and 4, as well as Figure 1, each of which is obtained by the characteristics of a fair or non-biased item. They can be used to assess the ability to solve mathematical problems. The ten grains are not identified with a moderate and small DIF effect based on the response of the two gender groups. All items have a small DIF effect except the P-SA3.7 item, which has a moderate effect. Identification of other DIF analyses obtained six items with a response level indicating that it is easier for groups of male students than female students. Conversely, the other four items have an easier difficulty response for female students than for male students. On the other hand, there are two biased items with significant DIF effects based on the responses of the two class-level groups. Both bias items have a significant DIF effect, with a more difficult level of difficulty response for grade 11 students than 12th-grade students. The following eighth DIF analysis results give a small DIF effect consisting of 5 items with an easier difficulty level response for grade 12 students compared to 11th grade students. In contrast, other items benefit grade 11 students.

The results of the same DIF item analysis in Figure 1 with the Mantel-Haenszel method using a contrast value DIF value plot based on gender groups showed that all items have a statistical value less than 3,842 (with a significance value greater than 0.050). On the other hand, the results of other DIF items analysis based on class level groups obtained two items identified with statistical values of 6,234 (p-values = 0.013) and 6,451 (p-values = 0.011), each greater than 3,842 and p-value smaller than 0.050. The two items are the 8th item (P-SA4.8) and the 9th (P-SA5.9), each of which shows a biased item, so the items need to be revised or not used to assess the mathematical problem-solving ability of high school students.



(a)



(b)

Figure 1. Plot Identification of Bias Items Based on Groups (a) Gender and (b) Class Levels (*Source: Own analysis*)

Reliability estimation

The analysis of the quality of other instruments based on the consistency level of measuring instruments using different research samples obtained a Cronbach's Alpha coefficient of 0.829. The level of instrument reliability in a high category, which shows the consistency of measuring devices for research samples, differs from the same characteristics that can be used to assess the ability to solve students' mathematical problems.

DISCUSSION

The results of the quality analysis of 10 items of the mathematical problem-solving ability assessment instrument, using the responses of 362 students, with the classical test theory approach, obtained the item difficulty index in the moderate and complex categories. Nine items have a moderate difficulty level, and one other is difficult. Overall, for each item of the instrument, there is a tendency for students to answer correctly with a proportion of more than 30.00% and less than 70.00% of the total number of students. The subsequent item identification obtained all items with the criteria of being able to provide predictions of the respondent's ability score based on the suitability of the correct and incorrect responses. Each item includes information that students who answer the item correctly (or incorrectly) give a high (or low) ability score. The results of the construction of answer choices consist of answer keys and distractors for each item, each with an effective distractor providing reasons for being selected. Other validity results using the analysis of differences in item functioning against gender group responses showed that not all items were identified as biased based on male and female groups. Other DIF analysis results obtained two items that were identified as biased based on differences in grade 11 and 12 group responses. Both items were revised or could not be used to assess students' mathematical problem-solving abilities. The DIF analysis results show that for all items, it does not benefit one gender group (male or female), and two biased items affect students' ability scores that help the grade 11 (or grade 12) group. The reliability estimate of the instrument in the very reliable category is used to assess mathematical problem-solving abilities based on different research samples with the same characteristics.

Several studies are relevant to the results of the instrument quality analysis of measuring instruments for assessing students' mathematical problem-solving abilities. Research relevant to the results of research examining instrument quality analysis using the classical theory analysis (CTT) approach (Annisavitri et al., 2020; Ridwan, Istiyono, et al., 2021; Wahyuni et al., 2018) and modern

theory (IRT) (Astuti et al., 2020; Fitriana & Supahar, 2019; Jatiningtyas et al., 2022; Reffiane et al., 2021; Sorby et al., 2022). The fundamental differences between the two approaches are based on the research sample's dependency and the test's specificity. Reliability and validity estimates using the CTT analysis approach each depend on the use of a sample of test takers and a specific test item construct, thus limiting the generalizability of the results to the research sample and also other items (DeVellis, 2006; Meguellati et al., 2024). Then, based on the specificity aspect of the test, all test items are assumed to have the same weight so that the tendency of the measuring instrument is less accurate in measuring the actual score of the test-taker's ability. This assumption does not apply to IRT analysis, which has item parameters that vary based on the level of difficulty and item discrimination (DeVellis, 2006; Meguellati et al., 2024).

Relevant research studies obtained the results of the analysis of the quality of the assessment instruments for students' mathematical problem-solving abilities based on content validity (Astuti et al., 2020; Bambang et al., 2018; Fitriana & Supahar, 2019; Jatiningtyas et al., 2022; Kania et al., 2024; Reffiane et al., 2021; Ridwan, Istiyono, et al., 2021; Rosyidi et al., 2024; Ulya et al., 2024; Wahyuni et al., 2018; Widodo et al., 2021). Content validity is related to how well the instrument items define the measurable dimensions so that it is ensured that the items are relevant to the construct and represent all measurable aspects (Muliana et al., 2020; Spoto et al., 2025; Yusoff, 2019; Zapata-Ospina & García-Valencia, 2022). Evaluation of content validity involves expert judgment to identify items based on aspects of relevance and representativeness, with assessments carried out objectively based on qualitative and quantitative assessment aspects (Nordin et al., 2022; Thompson & Senk, 2017; Yao et al., 2007). Evaluation of item relevance for all relevant research studies uses the Aiken validity index. Aiken validity effectively identifies items with high (or low) levels of relevance and can be used to compare the assessments of various expert groups (Merino-Soto, 2023; Roebianto et al., 2023) with coefficients in the interval 0 and 1.

The following relevant research study reports the results of research with a modern theory analysis approach consisting of item fit analysis with the Partial Credit Model (PCM) (or Rasch model) and identification of the level of difficulty of mathematical ability assessment instrument items (Astuti et al., 2020; Fitriana & Supahar, 2019; Jatiningtyas et al., 2022; Reffiane et al., 2021). Item fit analysis with PCM uses the infit *t* criteria (Reffiane et al., 2021) and infit mean-square (MNSQ) (Astuti et al., 2020) while an item fits with the Rasch model uses the outfit MNSQ criteria (Fitriana & Supahar, 2019; Jatiningtyas et al., 2022). The difference in the MNSQ infit (and outfit) criteria is based on the level of sensitivity of unexpected responses to items targeted at (and far from) the test taker's ability level (Guo & Wind, 2021; Su et al., 2007). Practically, both statistical values are used in educational evaluations by determining the cutoff score and assessment accuracy (Akbari & Shahrokhi, 2024). Then, for the difficulty level criteria, the item measure criteria are used (Jatiningtyas et al., 2022), indices at intervals of 0.00 and 1.00 (Fitriana & Supahar, 2019), and indices of -2.00 and 2.00 (Reffiane et al., 2021). In this case, the level of difficulty of the test items is measured on a logit scale, with the distribution of test items along a difficulty scale adjusted to the test taker's abilities (Chan, 2009; Miyata et al., 2024; Takeda et al., 2024).

Other relevant studies report the results of the analysis of the quality of instrument items using the CTT analysis approach based on the parameters of the level of difficulty and the discrimination power of the test items against the mathematical problem-solving assessment measuring instrument (Annisavitri et al., 2020; Ridwan, Istiyono, et al., 2021; Sorby et al., 2022). In the CTT aspect, the level of item difficulty is defined as the percentage of test takers who respond to the question item correctly (i.e., for low scores, indicating items with a high level of difficulty) (Dickinson, 2015; Metsämuuronen, 2023; Xie & Cobb, 2020) while the item discrimination index is defined as the valid items that differentiate test-takers' abilities (Xie & Cobb, 2020) which is identified as the point-biserial correlation between item scores and overall scores. In this case, if the item has a high level of difficulty, then the lower the level of item validity, the greater the abilities between test takers (Sweeney et al., 2022). Other relevant research reports based on the construct validity aspect of the mathematical problem-solving ability assessment instrument (Fitriana & Supahar, 2019; Wahyuni et al., 2018). Construct validity is related to the accuracy of conclusions based on theoretical and operational concepts (Agarwal, 2011; Stone, 2019). The study of the construct validity aspect by Fitriana and Supahar (2019) is based on

identifying unidimensional measuring instruments using dominant factor criteria, with the first factor being able to explain a variance greater than 20.00%. Unidimensionality refers to the assumption that one latent trait underlies the responses to all items in a test or that the relationship between responses to all items is explained by only one variable representing the construct being measured (Anderson et al., 2017; Hattie, 2015). Another construct validity study by Wahyuni et al., (2018) with a confirmatory factor analysis approach to identify the level of item validity for each measured variable (Lewis, 2017; Rogers, 2024) and also identify the suitability of the measurement model to empirical data using the calculation of the suitability index and assessment of model parameters (Abdellahi et al., 2023; Graham et al., 2003; Mueller & Hancock, 2015).

Other relevant research reports use construct reliability estimates to identify the level of consistency of mathematical problem-solving ability assessment measuring instruments (Annisavitri et al., 2020; Fitriana & Supahar, 2019; Ridwan, Istiyono, et al., 2021; Sorby et al., 2022; Wahyuni et al., 2018). The criteria for assessing the consistency of the measuring instrument use the Cronbach's Alpha coefficient and Intraclass Correlation. Reliability estimates define the extent to which a measuring instrument consistently measures a construct based on differences in aspects of opportunity, method, or assessor (Blankson, 2020; Rindskopf, 2015). Reviewing the reliability aspects of the instrument is a primary consideration in designing the test to ensure that the test consistently measures the instrument construct based on aspects of population and conditions (Adams, 2005; Bodoff, 2008).

Although some relevant research has been conducted, quality analysis studies of mathematical problem-solving ability assessment instruments using classical test theory analysis approaches and modern test theories are lacking. However, the results of this study provide differences in studies related to item identification based on distractor indices and differences in item functionality that do not tend to favor gender groups and grade levels. The results of item evaluation based on distractor index criteria provide information that developers of assessment measurement tools, such as researchers or teachers, can consider when constructing an item with possible answer choices to be selected by test participants or students. Other item quality analysis results indicate that standard and standard assessment measurement tools meet the aspects for all fair items and do not tend to favor one group of test participants or students based on gender and grade level differences.

CONCLUSION

Standardized measuring instruments assess students' cognitive abilities based on validity and reliability estimation. Several aspects of validity consist of the results of item quality analysis based on identifying item difficulty levels, items' ability to predict respondents' ability scores, and identifying distractor effectiveness. Other aspects of validity are identifying unbiased items based on the responses of certain research sample groups. Then, the estimation of the instrument's reliability is based on the level of consistency of the measuring instrument for the use of different research samples with the same characteristics. The results of this quantitative study using the classical test theory approach obtained items by mapping the level of item difficulty based on moderate and difficult categories, all items can predict respondents' ability scores, and the construction of answer choices other than the answer key provides possible reasons for being chosen by test participants. The results of the DIF analysis provide information on identified biased items obtained by all items that can be used to assess mathematical problem-solving abilities based on the responses of groups of male and female students. Different conditions exist for identified biased items based on the responses of groups of grades 11 and 12 students, so that these biased items need to be revised or eliminated, which can affect the validity of the respondents' ability scores. The study's results contribute to the analysis approach of the quality of cognitive test instruments using classical test theory based on item difficulty mapping, item identification, predicting respondent ability scores, and distractors' effectiveness. The contribution of the next item analysis approach considers the characteristics of the research sample based on differences in gender group responses and class levels, with an analysis of differences in item functioning using the Mantel-Haenszel method. The study's limitations consist of using research samples in only one provincial area, so it does not represent the characteristics of a particular region. However, this study provides the results of an instrument item analysis that considers the characteristics of the research

sample based on gender groups and class levels. Research samples with the same test participant characteristics can still use this measuring instrument. Then, for further research recommendations, large-scale trials will be conducted using research samples with school characteristics representing Indonesia's western, central, and eastern regions. Other research recommendations include instrument quality analysis with a modern test theory approach using item response theory analysis that considers item characteristics based on difficulty levels, discrimination indexes, guessing levels, and carelessness factors in responding to items. The use of the results of the item response theory analysis provides comprehensive assessment results based on the four aspects of item parameters.

DECLARATIONS

Author Contribution	:	M. Rais Ridwan: Conceptualization, Formal analysis, Resources, Visualization, and Writing - Original Draft; Samsul Hadi: Data curation, Methodology, and Supervision; Faihatuz Zuhairoh: Investigation, Validation, and Writing - Review & Editing
Funding Statement	:	This research was self-funded by the researcher.
Conflict of Interest	:	The authors declare no conflict of interest.
Additional Information	:	Additional information is available for this paper.

REFERENCES

Abdellahi, E. A. S., Hadri, Z. E., & Iaousse, M. (2023). A New Algorithm to Compute the Correlation Matrix Implied By a Confirmatory Factor Analysis Model. *2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 1–5. https://doi.org/10.1109/IRASET57153.2023.10152922

Adams, R. J. (2005). Reliability as a Measurement Design Effect. *Studies in Educational Evaluation*, *31*(2–3), 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008

Agarwal, N. K. (2011). Verifying Survey Items for Construct Validity: A Two-Stage Sorting Procedure for Questionnaire Design in Information Behavior Research. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–8. https://doi.org/10.1002/meet.2011.14504801166

Agustinsa, R., Anjasari, V., & Yensy, N. A. (2023). The Effect of Problem-Based Learning Models Using Contextual Worksheets on Middle School Students' Mathematical Problem Solving Ability. *Edumatica : Jurnal Pendidikan Matematika*, *13*(1), 48–56. https://doi.org/10.22437/edumatica.v13i01.24387

Aiken, Lewis R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, *45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Akbari, A., & Shahrokhi, M. (2024). Unveiling Fairness in Scoring: A Thorough Method for Precise Cutoff Score Calculation in Education Assessment. *Quality Assurance in Education*, *32*(3), 493–509. https://doi.org/10.1108/QAE-12-2023-0208

Algina, J., & Swaminathan, H. (2015). Psychometrics: Classical Test Theory. In J. D. B. T.-I. E. of the S. & B. S. (Second E. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)* (pp. 423–430). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.42070-2

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

Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of Methods Used for Estimating Content Validity. *Research in Social & Administrative Pharmacy : RSAP*, *15*(2), 214–221. https://doi.org/10.1016/j.sapharm.2018.03.066

Anderson, D., Joshua, D. K., & Tindal, G. (2017). Exploring the Robustness of a Unidimensional Item Response Theory Model With Empirically Multidimensional Data. *Applied Measurement in Education*, *30*(3), 163–177. https://doi.org/10.1080/08957347.2017.1316277

Andrade, Chittaranjan. (2020). The Inconvenient Truth About Convenience and Purposive Samples. *Indian Journal of Psychological Medicine*, *43*(1), 86–88. https://doi.org/10.1177/0253717620977000

Annisavitri, R., Sa'dijah, C., Qohar, A., Sa'diyah, M., & Anwar, L. (2020). Analysis of Mathematical Literacy Test as a Problem-Solving Ability Assessment of Junior High School Students. *AIP Conference Proceedings*, *2215*(1), 60002. https://doi.org/10.1063/5.0000648

Apriani, I. F., Turmudi, T., Jupri, A., & Syaodih, E. (2021). How is the Mathematical Problem-Solving Ability of Elementary School Pre-Service Teachers? *Journal of Engineering Science and Technology*, *16*, 57–64.

Astuti, A. P., Aziz, A., Sumarti, S. S., & Bharati, D. A. L. (2019). Preparing 21st Century Teachers: Implementation of 4C Character's Pre-Service Teacher through Teaching Practice. *Journal of Physics: Conference Series*, *1233*(1), 12109. https://doi.org/10.1088/1742-6596/1233/1/012109

Astuti, A. T., Supahar, Mundilarto, & Istiyono, E. (2020). Development of Assessment Instruments to Measure Problem Solving Skills in Senior High School. *Journal of Physics: Conference Series*, *1440*(1), 012063. https://doi.org/10.1088/1742-6596/1440/1/012063

Bahar, A., & Maker, C. J. (2015). Cognitive Backgrounds of Problem Solving: A Comparison of Open-ended vs. Closed Mathematics Problems. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*(6), 1531–1546. https://doi.org/10.12973/eurasia.2015.1410a

Bambang, S. R. M., Salasi, R., Hasbi, M., & Mardhiah, M. Z. (2018). The Development of an Instrument to Explore Non-Routine Problem Solving Strategies Among Mathematics Education Students. *Journal of Physics: Conference Series*, *1088*(1), 12059. https://doi.org/10.1088/1742-6596/1088/1/012059

Barana, A., Boetti, G., & Marchisio, M. (2022). Self-Assessment in the Development of Mathematical Problem-Solving Skills. *Education Sciences*, *12*(2), 81. https://doi.org/10.3390/educsci12020081

Blankson, A. N. (2020). Reliability, Issues of. In *The Wiley Encyclopedia of Personality and Individual Differences* (pp. 165–168). https://doi.org/10.1002/9781119547167.ch98

Bodoff, D. (2008). Test Theory for Evaluating Reliability of IR Test Collections. *Information Processing & Management*, *44*(3), 1117–1145. https://doi.org/10.1016/j.ipm.2007.11.006

Brown, J. D. (2013). Classical test theory. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing* (pp. 323–335). Routledge Taylor & Francis Group. https://doi.org/10.4324/9780203181287-35

Campbell, J. I. D. (2004). *Handbook of Mathematical Cognition*. Psychology Press.

Chan, David W. (2009). Developing the Impossible Figures Task to Assess Visual-Spatial Talents Among Chinese Students: A Rasch Measurement Model Analysis. *Gifted Child Quarterly*, *54*(1),

59–71. https://doi.org/10.1177/0016986209352685

Choi, E., & Jeon, J. (2022). Factors Influencing Problem-Solving Competence of Nursing Students: A Cross-Sectional Study. *Healthcare (Basel, Switzerland)*, *10*(7), 1184. https://doi.org/10.3390/healthcare10071184

Coulacoglou, C., & Saklofske, D. H. (2017). *Psychometrics and Psychological Assessment: Principles and Applications*. Elsevier Academic Press.

Darma, I. K., Candiasa, I. M., Sadia, I. W., & Dantes, N. (2018). The Effect of Problem-Based Learning Model and Authentic Assessment on Mathematical Problem Solving Ability by Using Numeric Ability as the Covariable. *Journal of Physics: Conference Series*, *1040*(1), 012035. https://doi.org/10.1088/1742-6596/1040/1/012035

DeVellis, R. F. (2006). Classical Test Theory. *Medical Care*, *44*(11), S50–S59. https://doi.org/10.1097/01.mlr.0000245426.10853.30

Dickinson, J. R. (2015). An Empirical Comparison Of Measures Of Multiple-Choice Question Item Difficulty. In L. Robinson (Ed.), *Marketing Dynamism & Sustainability: Things Change, Things Stay the Same…. Developments in Marketing Science: Proceedings of the Academy of Marketing Science* (pp. 327–328). Springer International Publishing. https://doi.org/10.1007/978-3-319-10912-1_110

Douglas-Morris, J., Ritchie, H., Willis, C., & Reed, D. (2021). Identification-Based Multiple-Choice Assessments in Anatomy can be as Reliable and Challenging as Their Free-Response Equivalents. *Anatomical Sciences Education*, *14*(3), 287–295. https://doi.org/10.1002/ase.2068

Fahrudin, D., Mardiyana, & Pramudya, I. (2019). The Analysis of Mathematic Problem Solving Ability by Polya Steps on Material Trigonometric Reviewed from Self-Regulated Learning. *Journal of Physics: Conference Series*, *1254*(1), 012076. https://doi.org/10.1088/1742-6596/1254/1/012076

Fernandes, H. J. X. (1984). *Testing and Measurement*. National Education Planning, Evaluation, and Curriculum Development.

Firda, N., Suryadi, D., & Dahlan, J. A. (2023). Mathematical Problem-Solving Ability of Junior High School Students Based on Polya. *Edumatica : Jurnal Pendidikan Matematika*, *13*(3), 273–284. https://doi.org/10.22437/edumatica.v13i03.29287

Firdausi, & Supinah, R. (2021). Development of Authentic Assessment to Improve Students' Mathematical Problem Solving Ability. *Journal of Physics: Conference Series*, *1836*(1), 12065. https://doi.org/10.1088/1742-6596/1836/1/012065

Fitriana, D. A., & Supahar, S. (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. https://doi.org/10.33122/ijtmer.v2i3.81

Gamerman, D., Gonçalves, F. B., & Soares, T. M. (2018). Differential Item Functioning. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory: Three Volume Set* (pp. 67–86). Chapman and Hall/CRC. https://doi.org/10.1201/9781315119144-57

Gozali, I., Syamsuri, S., Nindiasari, H., & Fatah, A. (2022). The Effect of Problem Based Learning on Mathematical Disposition and Students' Problem-Solving Ability. *Edumatica : Jurnal Pendidikan Matematika*, *12*(2), 102–110. https://doi.org/10.22437/edumatica.v12i02.15772

Grabowski, K. C., & Lin, R. (2019). Multivariate Generalizability Theory In Language Assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative Data Analysis for Language Assessment Volume*

I: Fundamental Techniques (pp. 54–80). https://doi.org/10.4324/9781315187815-4

Graham, J. M., Abbie, C. G., & Thompson, B. (2003). Consequences of Not Interpreting Structure Coefficients in Published CFA Research: A Reminder. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(1), 142–153. https://doi.org/10.1207/S15328007SEM1001_7

Greeno, C. (2003). Measurement, or How Do We Know What We Know? Topic One: Validity. *Family Process*, *42*(3), 433–435. https://doi.org/10.1111/j.1545-5300.2003.00433.x

Guarte, J. M., & Barrios, E. B. (2006). Estimation Under Purposive Sampling. *Communications in Statistics - Simulation and Computation*, *35*(2), 277–284. https://doi.org/10.1080/03610910600591610

Guo, Wenjing, & Wind, Stefanie A. (2021). An Iterative Parametric Bootstrap Approach to Evaluating Rater Fit. *Applied Psychological Measurement*, *45*(5), 315–330. https://doi.org/10.1177/01466216211013105

Hasanah, A. N., Priatna, N., & Yulianti, K. (2021). The Ability of Mathematical Problem Solving of Junior High School Students in Situation based Learning and Discovery Learning. *Journal of Physics: Conference Series*, *1806*(1), 012070. https://doi.org/10.1088/1742-6596/1806/1/012070

Hattie, J. (2015). Dimensionality of Tests: Methodology. In J. D. B. T.-I. E. of the S. & B. S. (Second E. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences: Second Edition* (pp. 437–439). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.44016-X

Hidalgo, M. D., & Gómez-Benito, J. (2010). Differential Item Functioning. In P. Peterson, E. Baker, & B. B. T.-I. E. of E. (Third E. McGaw (Eds.), *International Encyclopedia of Education, Third Edition* (pp. 36–44). Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.00242-6

Holland, P. W., & Thayer, D. T. (1988). Differential Item Performance and the Mantel-Haenszel Procedure. In *Test validity* (pp. 129–145). Lawrence Erlbaum Associates, Inc. https://doi.org/10.1037/14047-004

Hwang, G.-J., Hung, C.-M., & Chen, N.-S. (2014). Improving Learning Achievements, Motivations and Problem-Solving Skills Through a Peer Assessment-based Game Development Approach. *Educational Technology Research and Development*, *62*(2), 129–145. https://doi.org/10.1007/s11423-013-9320-7

Ismail, A., Razali, S. S., Hashim, S., Abiddin, N. Z., Masek, A., & Abd Samad, N. (2021). The Integration of Problem Based Learning in Generating 21st Century Skills. *2021 IEEE 12th Control and System Graduate Research Colloquium (ICSGRC)*, 19–23. https://doi.org/10.1109/ICSGRC53186.2021.9515211

Istiqomah, F., Mardiyana, & Pramudya, I. (2019). Analysis of Student's Problem Solving Ability at Junior High School. *Journal of Physics: Conference Series*, *1211*(1), 012085. https://doi.org/10.1088/1742-6596/1211/1/012085

Jatiningtyas, P. D., Kartono, & Mindyarto, B. N. (2022). Test Instruments to Measure Non-Routine Mathematics Problem Solving Ability Grade IV Elementary School Students. *Journal of Education Research and Evaluation*, *6*(3), 407–414. https://doi.org/10.23887/jere.v6i3.48656

Kadir, K. (2023). Students' Mathematics Achievement Based on Performance Assessment through Problem Solving-Posing and Metacognition Level. *Mathematics Teaching Research Journal*, *15*(3), 109–135.

Kania, N., Kusumah, Y. S., Dahlan, J. A., Nurlaelah, E., Gürbüz, F., & Bonyah, E. (2024). Constructing

and Providing Content Validity Evidence through the Aiken's Vindex Based on the Experts' Judgments of the Instrument to Measure Mathematical Problem-Solving Skills. *REID (Research and Evaluation in Education)*, *10*(1), 64–79. https://doi.org/10.21831/reid.v10i1.71032

Kaul, R., Kaul, R., & Paul, P. (2025). Reliability study. In A. E. M. Eltorai, J. A. Bakal, & C. M. B. T.-T. C. Gibson (Eds.), *Handbook for Designing and Conducting Clinical and Translational Research* (pp. 247–249). Academic Press. https://doi.org/10.1016/B978-0-323-91790-2.00040-X

Kharisudin, I., & Cahyati, N. E. (2020). Problem-Solving Ability Using Mathematical Modeling Strategy on Model Eliciting Activities Based on Mathematics Self-Concept. *Journal of Physics: Conference Series*, *1567*(3), 32067. https://doi.org/10.1088/1742-6596/1567/3/032067

Kusuma, A. B., & Untarti, R. (2020). The Mathematical Problem-Solving In Algorithm Subject. *International Journal of Scientific and Technology Research*, *9*(4), 2982–2986.

Lewis, T. F. (2017). Evidence Regarding the Internal Structure: Confirmatory Factor Analysis. *Measurement and Evaluation in Counseling and Development*, *50*(4), 239–247. https://doi.org/10.1080/07481756.2017.1336929

Lim Li Gek, P. (2020). Teaching through Problem Solving. In *Mathematics Teaching in Singapore* (pp. 175–186). WORLD SCIENTIFIC. https://doi.org/doi:10.1142/9789811220159_0011

Linn, R. L. (2010). Validity. In P. Peterson, E. Baker, & B. B. T.-I. E. of E. (Third E. McGaw (Eds.), *International Encyclopedia of Education, Third Edition* (pp. 181–185). Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.00893-9

Magis, D., Béland, S., Tuerlinckx, F., & de Boeck, P. (2010). A General Framework and an R Package for the Detection of Dichotomous Differential Item Functioning. *Behavior Research Methods*, *42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

Mansyur, M. Z., & Sunendar, A. (2020). Improving Students' Mathematical Problem Solving Ability through Metacognitive Guidance Approach. *Edumatica : Jurnal Pendidikan Matematika*, *10*(2), 19–27. https://doi.org/10.22437/edumatica.v10i2.10494

Meguellati, S., Samia, A., Ferhat, A., Djelloul, A., & Khalifa, Z. A. (2024). A Critical Analysis of the Use of Classical Test Theory (CTT) in Psychological Testing: A Comparison with Item Response Theory (IRT). *Pakistan Journal of Life and Social Sciences*, *22*(2), 9442–9449. https://doi.org/10.57239/PJLSS-2024-22.2.00715

Mellenbergh, Gideon J. (1982). Contingency Table Models for Assessing Item Bias. *Journal of Educational Statistics*, *7*(2), 105–118. https://doi.org/10.3102/10769986007002105

Memon, M. A., Thurasamy, R., Ting, H., & Cheah, J.-H. (2025). Purposive Sampling: A Review and Guidelines for Quantitative Research. *Journal of Applied Structural Equation Modeling*, *9*(1), 1–23. https://doi.org/10.47263/JASEM.9(1)01

Merino-Soto, C. (2023). Aiken's V Coefficient: Differences in Content Validity Judgments. *MHSalud*, *20*(1), 23–32. https://doi.org/10.15359/mhs.20-1.3

Metsämuuronen, J. (2023). Seeking the Real Item Difficulty: Bias-Corrected Item Difficulty and Some Consequences in Rasch and IRT Modeling. *Behaviormetrika*, *50*(1), 121–154. https://doi.org/10.1007/s41237-022-00169-9

Miyata, K., Kondo, Y., Bando, K., Hara, T., & Takahashi, Y. (2024). Structural Validity of the Mini-Balance Evaluation Systems Test in Individuals With Spinocerebellar Ataxia: A Rasch Analysis Study. *Archives of Physical Medicine and Rehabilitation*, *105*(4), 742–749.

https://doi.org/10.1016/j.apmr.2023.12.015

Mueller, R. O., & Hancock, G. R. (2015). Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis. In J. D. B. T.-I. E. of the S. & B. S. (Second E. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences: Second Edition* (pp. 686–690). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.25009-5

Muhtarom, Zuhri, M. S., & Herlambang, B. A. (2024). Analysis of Mathematics Problem-Solving Ability of Junior High School Students in Terms of Learning Independence. *AIP Conference Proceedings*, *3046*(1), 020013. https://doi.org/10.1063/5.0194547

Muliana, N., Pada, A. U. T., & Nurmaliah, C. (2020). Content Validity of Conation Assessment. *Journal of Physics: Conference Series*, *1460*(1), 12057. https://doi.org/10.1088/1742-6596/1460/1/012057

Mustofa, B., Mardiyana, & Slamet, I. (2020). An Analysis of Problem Solving Ability in Linear Equation Systems with Two Variables. *Journal of Physics: Conference Series*, *1538*(1), 12099. https://doi.org/10.1088/1742-6596/1538/1/012099

Nasution, M. L., Yerizon, Y., & Gusmiyanti, R. (2018). Students' Mathematical Problem-Solving Abilities Through the Application of Learning Models Problem Based Learning. In Ramli, Yulkifli, Festiyed, A. M., S. R., Alizar, P. D.H., & P. A. (Eds.), *IOP Conference Series: Materials Science and Engineering* (Vol. 335, Issue 1). Institute of Physics Publishing. https://doi.org/10.1088/1757-899X/335/1/012117

Ningsih, S., & Hidayati, K. (2022). The Role of Abstraction Ability in Mathematical Problem Solving. *AIP Conference Proceedings*, *2575*(1), 050021. https://doi.org/10.1063/5.0109166

Nordin, M. R., Jamal, S. A., & Anuar, N. A. M. (2022). Content Validity of Research Instruments: Assessing Domestic Ecotourism in Protected Areas. *Enlightening Tourism*, *12*(2), 565–599. https://doi.org/10.33776/et.v12i2.7123

Ouzouni, C., & Nakakis, K. (2011). Validity and Reliability of Measurement Instruments in Quantitative Studies. *Nosileftiki*, *50*(2), 231–239.

Paek, I., & Holland, P. (2015). A Note on Statistical Hypothesis Testing Based on Log Transformation of the Mantel–Haenszel Common Odds Ratio for Differential Item Functioning Classification. *Psychometrika*, *80*(2), 406–411. https://doi.org/10.1007/s11336-013-9394-5

Parandrengi, R., Anwar, L., & Sa'dijah, C. (2024). Analysis of Problem-Solving Ability of High School Students in Jambi. *AIP Conference Proceedings*, *3106*(1), 050002. https://doi.org/10.1063/5.0215320

Permata, L. D., Kusmayadi, T. A., & Fitriana, L. (2018). Mathematical Problem Solving Skills Analysis about Word Problems of Linear Program Using IDEAL Problem Solver. *Journal of Physics: Conference Series*, *1108*(1), 012025. https://doi.org/10.1088/1742-6596/1108/1/012025

Rakhmawati, I. A., Budiyono, & Saputro, D. R. S. (2019). An Analysis of Problem Solving Ability Among High School Students in Solving Linear Equation System Word Problems. *Journal of Physics: Conference Series*, *1211*(1), 012098. https://doi.org/10.1088/1742-6596/1211/1/012098

Reffiane, F., Sudarmin, S., Wiyanto, W., & Saptono, S. (2021). Developing an Instrument to Assess Students' Problem-Solving Ability on Hybrid Learning Model Using Ethno-STEM Approach through Quest Program. *Pegem Journal of Education and Instruction*, *11*(4), 1–8. https://doi.org/10.47750/pegegog.11.04.01

Ridwan, M. R., Hadi, S., & Jailani. (2022a). Identification of Effectiveness Measurements and Bias

Publication of Literature Results Study: A Cooperative Learning Models on Mathematics Learning Outcomes of Vocational School Students in Indonesia. *Journal on Efficiency and Responsibility in Education and Science*, *15*(3), 189–200. https://doi.org/10.7160/eriesj.2022.150306

Ridwan, M. R., Hadi, S., & Jailani, J. (2022b). A Meta-Analysis Study on the Effectiveness of a Cooperative Learning Model on Vocational High School Students' Mathematics Learning Outcomes. *Participatory Educational Research*, *9*(4), 396–421. https://doi.org/10.17275/per.22.97.9.4

Ridwan, M. R., Hadi, S., & Jailani, J. (2023a). A Meta-Analysis of Numerical Aptitude's Effect on Learning Outcomes and Mathematical Ability. *TEM Journal*, *12*(1), 434–444. https://doi.org/10.18421/TEM121-53

Ridwan, M. R., Hadi, S., & Jailani, J. (2023b). Measurement of Psychometric Properties Numerical Aptitude Assessment Scale for Prospective High School Students: A Rasch Model Analysis. *TEM Journal*, *12*(4), 2416–2429. https://doi.org/10.18421/TEM124-54

Ridwan, M. R., Istiyono, E., & Widihastuti, W. (2021). Test Items Analysis of Mathematical Problem Solving Ability using a Classical Test Theory Approach. *Jurnal Pendidikan MIPA*, *22*(1), 98–111. https://doi.org/10.23960/jpmipa/v22i1.pp98-111

Ridwan, M. R., Retnawati, H., Hadi, S., & Jailani. (2021). The Effectiveness of Innovative Learning on Mathematical Problem-Solving Ability: A Meta-Analysis. *International Journal of Research in Education and Science (IJRES)*, *7*(3), 910–932. https://doi.org/10.46328/ijres.2287

Rindskopf, D. (2015). Reliability: Measurement. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition* (pp. 248–252). https://doi.org/10.1016/B978-0-08-097086-8.44050-X

Rios, J. A., Ling, G., Pugh, R., Becker, D., & Bacall, A. (2020). Identifying Critical 21st-Century Skills for Workplace Success: A Content Analysis of Job Advertisements. *Educational Researcher*, *49*(2), 80–89. https://doi.org/10.3102/0013189X19890600

Roebianto, A., Savitri, S. I., Aulia, I., Suciyana, A., & Mubarokah, L. (2023). Content Validity: Definition and Procedure of Content Validation in Psychological Research. *TPM - Testing, Psychometrics, Methodology in Applied Psychology*, *30*(1), 5–18. https://doi.org/10.4473/TPM30.1.1

Rogers, P. (2024). Best Practices for Your Confirmatory Factor Analysis: A JASP and Lavaan Tutorial. *Behavior Research Methods*, *56*(7), 6634–6654. https://doi.org/10.3758/s13428-024-02375-7

Rosli, R., Goldsby, D., & Capraro, M. M. (2013). Assessing Students' Mathematical Problem-Solving and Problem-Posing Skills. *Asian Social Science*, *9*(16), 54–60. https://doi.org/10.5539/ass.v9n16p54

Rosyidi, A. H., Sari, Y. M., Fardah, D. K., & Masriyah, M. (2024). Designing Mathematics Problem-Solving Assessment with Geogebra Classroom: Proving the Instrument Validity. *Journal of Education and Learning (EduLearn)*, *18*(3), 1030–1038. https://doi.org/10.11591/edulearn.v18i3.21191

Runnels, J. (2013). Measuring Differential Item and Test Functioning Across Academic Disciplines. *Language Testing in Asia*, *3*(1), 9. https://doi.org/10.1186/2229-0443-3-9

Ryan, J. M., & DeMark, S. (2012). Variation in Achievement Scores Related to Gender, Item Format, and Content Area Tested. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale Assessment*

Programs for All Students (pp. 64–82). Routledge. https://doi.org/10.4324/9781410605115-9

Sarstedt, M., & Mooi, E. (2019). *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics (3rd ed.)*. Springer-Verlag GmbH. https://doi.org/10.1007/978-3-662-56707-4

Scott, S. (2007). Team Performance and the Problem-Solving Approach. *Journal of Industrial Technology*, *23*(4).

Sireci, S. G., & Soto, A. (2016). Validity and Accountability: Test Validation for 21st-Century Educational Assessments. In H. Braun (Ed.), *Meeting the Challenges to Measurement in an Era of Accountability* (pp. 149–167). Routledge. https://doi.org/10.4324/9780203781302-13

Sorby, S. A., Duffy, G., & Yoon, S. Y. (2022). Math Instrument Development for Examining the Relationship between Spatial and Mathematical Problem-Solving Skills. *Education Sciences*, *12*(11), 828. https://doi.org/10.3390/educsci12110828

Spoto, A., Nucci, M., Prunetti, E., & Vicovaro, M. (2025). Improving Content Validity Evaluation of Assessment Instruments through Formal Content Validity Analysis. *Psychological Methods*, *30*(2), 203–222. https://doi.org/10.1037/met0000545

Steyer, R. (2015). Classical (Psychometric) Test Theory. In J. D. B. T.-I. E. of the S. & B. S. (Second E. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)* (pp. 785–791). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.44006-7

Stone, C. (2019). A Defense and Definition of Construct Validity in Psychology. *Philosophy of Science*, *86*(5), 1250–1261. https://doi.org/10.1086/705567

Su, Y.-H., Sheu, C.-F., & Wang, W.-C. (2007). Computing Confidence Intervals of Item Fit Statistics in the Family of Rasch Models Using the Bootstrap Method. *Journal of Applied Measurement*, *8*(2), 190–203.

Sweeney, S. M., Sinharay, S., Johnson, M. S., & Steinhauer, E. W. (2022). An Investigation 0f the Nature and Consequence of the Relationship Between IRT Difficulty and Discrimination. *Educational Measurement: Issues and Practice*, *41*(4), 50–67. https://doi.org/10.1111/emip.12522

Takeda, R., Miyata, K., Tamura, S., Kobayashi, S., & Iwamoto, H. (2024). Item Distribution of the Berg Balance Scale in Older Adults with Hip Fracture: A Rasch Analysis. *Physiotherapy Theory and Practice*, *40*(1), 136–143. https://doi.org/10.1080/09593985.2022.2109541

Thompson, D. R., & Senk, S. L. (2017). Examining Content Validity of Tests Using Teachers' Reported Opportunity to Learn. *Investigations in Mathematics Learning*, *9*(3), 148–155. https://doi.org/10.1080/19477503.2017.1310572

Ukobizaba, F., Nizeyimana, G., & Mukuka, A. (2021). Assessment Strategies for Enhancing Students' Mathematical Problem-solving Skills: A Review of Literature. *Eurasia Journal of Mathematics, Science and Technology Education*, *17*(3), em1945. https://doi.org/10.29333/ejmste/9728

Ulya, H., Sugiman, S., & Rosnawati, R. (2024). Design and Content Validity of Mathematics Creative Problem-Solving Ability Instrument for Junior High School Students. *Eurasia Journal of Mathematics, Science and Technology Education*, *20*(6), em2462. https://doi.org/10.29333/ejmste/14661

Varma, S. (2006). *Preliminary Item Statistics Using Point-Biserial Correlation and P-Values (1st Ed.)*. Educational Data Systems Inc.

Wahyuni, V., Kartono, K., & Susiloningsih, E. (2018). Development of Project Assessment Instruments to Assess Mathematical Problem Solving Skills on A Project-Based Learning. *Journal of Research and Educational Research Evaluation*, *7*(2), 147–153. https://doi.org/10.15294/jere.v7i2.24501

Weng, L. C., Dai, Y. T., Huang, H. L., & Chen, S. L. (2007). Brief Review of Meanings, Measures, and Effects in Problem Solving. *Journal of Nursing*, *54*(4), 83–87.

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

Williamson, K. (2018). Populations and Samples. In K. Williamson & G. Johanson (Eds.), *Research Methods: Information, Systems, and Contexts: Second Edition* (pp. 359–377). Chandos Publishing. https://doi.org/10.1016/B978-0-08-102220-7.00015-7

Xie, D., & Cobb, C. L. (2020). Item Analysis. In *The Wiley Encyclopedia of Personality and Individual Differences* (pp. 159–163). https://doi.org/10.1002/9781119547167.ch97

Yao, G., Wu, C., & Yang, C. (2007). Examining the Content Validity of the WHOQOL-BREF from Respondents' Perspective by Quantitative Methods. *Social Indicators Research*, *85*(3), 483–498. https://doi.org/10.1007/s11205-007-9112-8

Yasin, M., Huda, S., Putra, F. G., Syazali, M., Umam, R., & Widyawati, S. (2020). IMPROVE Learning Model and Learning Independence: Influence and Interaction on Mathematics Problem-Solving Abilities in Islamic Boarding School. *Journal of Physics: Conference Series*, *1467*(1), 12003. https://doi.org/10.1088/1742-6596/1467/1/012003

Yüksel, S., Demir, P., & Alkan, A. (2019). Factors Causing Occurrence of Artificial Dif: A Simulation Study for Dichotomous Data. *Communications in Statistics - Simulation and Computation*, *48*(7), 2004–2011. https://doi.org/10.1080/03610918.2018.1429622

Yuristia, N., & Musdi, E. (2020). Analysis of Early Mathematical Problem-Solving Ability in Mathematics Learning for Junior High School Student. *Journal of Physics: Conference Series*, *1554*(1), 12026. https://doi.org/10.1088/1742-6596/1554/1/012026

Yusoff, M. S. B. (2019). ABC of Content Validation and Content Validity Index Calculation. *Education in Medicine Journal*, *11*(2), 49–54. https://doi.org/10.21315/eimj2019.11.2.6

Zapata-Ospina, J. P., & García-Valencia, J. (2022). Validity Based on Content: A Challenge in Health Measurement Scales. *Journal of Health Psychology*, *27*(2), 481–493. https://doi.org/10.1177/1359105320953477

Zieky, M. (1993). Practical Questions in the Use of DIF Statistics in Test Development. In *Differential item functioning* (pp. 337–347). Lawrence Erlbaum Associates, Inc. https://doi.org/10.1075/z.62.13kok