22 (1), 2021, 98-111

Jurnal Pendidikan MIPA

e-ISSN: 2550-1313 | p-ISSN: 2087-9849

http://jurnal.fkip.unila.ac.id/index.php/jpmipa/

Test Items Analysis of Mathematical Problem Solving Ability using a Classical Test Theory Approach

Muhammad Rais Ridwan[1,2*], Edi Istiyono[2], Widihastuti[2]

[1]Department of Mathematics Education, STKIP YPUP Makassar, Indonesia
[2]Department of Educational Research and Evaluation, Yogyakarta State University, Indonesia

Abstract: This study aims to analyze the item characteristics of the mathematics problem-solving ability test instrument using the Classical Test Theory model. The data collection was based on the results of the test documentation as many as 359 students with dichotomous data. Qualitative validation analysis by experts uses the panel method, and quantitative analysis uses the Aiken index and Content Validity Ratio (CVR), while quantitative validation uses biserial point correlation. The test reliability index used the Alpha-Cronbach method. The results of qualitative validation show that all items correspond to the indicators of solving mathematical problems, but technically writing a few items needs improvement. Quantitative analysis using the Aiken index and CVR shows all items are valid with good validity. The reliability of the test is stable, with a coefficient of 0.83. Then, the test instrument consists of items with very difficult, difficult, and easy levels. All items were able to distinguish the test taker's ability and the effectiveness of the distractor to function correctly.

Keywords: classical test theory, biserial point correlation, validity, reliability.

Abstrak: *Penelitian ini bertujuan untuk menganalisis karakteristik butir soal instrumen tes kemampuan pemecahan masalah matematika menggunakan model Classical Test Theory. Pengumpulan data berdasarkan hasil dokumentasi tes sebanyak 359 siswa dengan bentuk data dikotomus. Analisis validasi kualitatif oleh ahli dengan metode panel dan analisis kuantitatif menggunakan indeks Aiken dan Content Validity Ratio (CVR) sedangkan validasi kuantitatif menggunakan korelasi poin biserial. Indeks reliabilitas tes menggunakan metode Alpha-Cronbach. Hasil validasi kualitatif menunjukkan semua butir memiliki kesesuaian dengan indikator pemecahan masalah matematika, namun secara teknis penulisan beberapa butir soal perlu perbaikan. Analisis kuantitatif menggunakan indeks Aiken dan CVR menunjukkan semua butir valid dengan validitas baik. Reliabilitas tes stabil dengan koefisien sebesar 0.83. Kemudian, instrumen tes terdiri dari butir dengan tingkat sangat sukar, sukar dan mudah. Semua butir soal mampu membedakan kemampuan peserta tes dan juga efektivitas distraktor berfungsi dengan baik.*

Kata kunci: *teori tes klasik, korelasi poin biserial, validitas, reliabilitas*.

Muhammad Rais Ridwan et al.
Corresponding Email:
mrais.2019@student.uny.ac.id

DOI: http://dx.doi.org/10.23960/jpmipa/v22i1.pp98-111
Received: 27 March 2021
Accepted: 18 May 2021

▪ INTRODUCTION

The test is a collection of items consisting of many questions that students must answer. The purpose of cognitive tests is to measure cognitive abilities with the level of understanding and mastery of the material coverage required and following specific learning objectives. According to Mardapi (2017: 94), a test is a form of instrument used to make a measurement. Measurement is the third stage in the development of the test instrument after designing and testing the test.

Researchers such as Oriondo & Antonio (1998) and Mardapi (2008) developed the procedure for developing test instruments. According to Zimmerman & Risembarg in Istiyono (2018), the test device development stage consists of test design, test trials, and measurements. The design of the test includes setting test objectives, determining the competence and material to be tested, compiling the instrument matrix, preparing and preparing the test grid (blueprint), writing and assembling test items, compiling scoring rubrics, validating the contents of the items, revising to improve the test items, and instrument assembly. Then, the test trial consists of determining the test subject, implementing the test trial, scoring and analyzing the test items, revising the items that do not meet the criteria for the desired test item parameters. The next stage, the measurements, includes assembling the test based on the trial results, determining the measurement subject, implementing the measurement, scoring and analyzing the measurement data, and interpreting the results.

The development of a mathematics problem-solving ability test instrument aims to measure students' mathematical problem-solving abilities in class XI and XII of Senior High School or equivalent. Competency and material testing based on the selection of core competencies and basic competencies to measure mathematical problem-solving abilities. Then, the preparation of the test instrument matrix consists of the material and substance column, while the row consists of aspects or sub aspects of ability. Next is the practice of the test grid, scoring guidelines, and the validation of the contents of the test instruments. The next stage is testing the test instrument to see the suitability of the test instrument with the ability to be measured, and the final step, namely the measurement, produces response data to the test instrument. The stages of instrument validation in this study consisted of content validation using qualitative and quantitative analysis. Another validation is quantitative validity. Then, the student response data were obtained from the results of the test instrument trial to analyze the characteristics of the test instrument item of the mathematical problem-solving ability.

Validity is essential in the development of test and non-test instruments (Mardapi, 2016). Also, Subali (2014) emphasizes that validity and reliability analysis are fundamental and need to be done in developing an instrument. Valid instruments must have internal and external validity (Sugiyono, 2015). The internal validity is rational, while internal validity consists of construct and content validity. According to Alifa et al. (2018), content validity consists of material, construct, and language aspects in an instrument. Logical or empirical analysis of the validation of the test content is evidence of the validity of interpreting test scores. Content validity is fundamental and proven to support the accuracy of the measurement results of a test. On the other hand, according to Hinkin (1995), several techniques for determining the validity of the content that researchers have applied have not been able to ensure that the scale or test has valid content. However, the information will provide evidence that the item is a viable construct in testing and reduce the need for future repairs.

Analysis of the validity of the contents of an instrument is the initial stage of developing a test instrument to obtain the quality of the instrument in the form of the accuracy of the student's ability assessment data. This research focuses specifically on the analysis of content validity as the initial stage of developing an instrument for assessing the mathematical problem-solving abilities of high school students. The technique of determining the validity of the content is based on qualitative expert reviews and quantitative analysis based on expert judgment. A quantitative analysis approach based on expert judgment uses the Aiken index (Aiken, 1985) and Lawshe's CVR formula (Lawshe, 1975).

Several studies related to content validity analysis in developing test instruments. Research by Ikhsanudin & Subali (2018) conducted an analysis of the validity of the contents of the formative test instruments for high school students in Biology subjects for the first semester. The results showed that the instrument contained one item from 35 items that did not agree on the proportion between the assessors regarding the essence of the test instrument. Besides, based on the quantitative analysis review using the Aiken index, three items were obtained in low validity. Then, a qualitative study suggests paying attention to the main language of the questions for the nine items on the test instrument.

The development of another instrument in learning volleyball is an information system-based assessment instrument (Yudiana et al., 2017). The results of the content validation analysis using Lawshe CVR showed that the information system-based early-stage assessment instrument proved valid in measuring the performance of high school students' volleyball game. The stages required in conducting a content validation analysis consist of defining the content domain to be assessed, item construction, and expert judgment of the constructed items (Delgado-Rico et al., 2012).

In addition to testing the validity of obtaining good quality instruments, it is also essential in developing the test instrument to analyze the characteristics of the items. The use of test instruments must have excellent and representative item characteristics in measuring every aspect of actual student achievement. The characteristics of the items consist of difficulty level, difference power index, and distractor. The difficulty level of a suitable item has a composition of easy, medium, and difficult questions that spread proportionally according to the test material testing. A good item discrimination index can also differentiate between groups of high and low ability students, and the effectiveness of the distractor is functioning correctly. The approach to analyzing item characteristics consists of classical test theory (CTT) and item response theory (IRT) models.

Researchers have carried out an analysis of test measurement data using the classical test theory approach (Mardapi, 2018; Suhariyono et al., 2014; Sarea & Ruslan, 2019; Setiawati et al. 2018; Sunarmi et al. 2016; Fernanda & Hidayah, 2020; Awopeju & Afolabi, 2016; Bichi et al. 2019; Bichi et al. 2015, Adegoke, 2013; Guler, et al. 2014; Pido, 2012; Jabrayilov et al. 2016; Hambleton & Jones, 1993; Royce, 2009; Eleje et al. 2018), the Rasch model, and the item response theory. Mardapi (1998) research shows that the number of items rejected according to the classical test theory is more than according to the item response theory. The consistency index of the item analysis results is low. There is a high relationship between the results of the estimated ability in the two approaches of 0.99 and 0.98 for the test instrument. First and second, the test reliability index according to classical test theory and item response theory are high in terms of the standard error measurement error.

Subsequent research, the analysis of the accuracy of the final semester test equipment for Physics subjects using the CTT and IRT approaches showed that the reliability coefficient for the Education Unit Level Curriculum and 2013 Curriculum questions were 0.806 and 0.576, respectively. Then the level of difficulty for the question models was 0.374 and 0.364, respectively. Simultaneously, the difference power index of the two models is 0.370 and 0.270 (Suhariyono et al., 2014). Then, the analysis of the classical theory approach and the item response theory was also carried out by Sarea & Ruslan (2019). It showed that the difficulty level of the items categorized as well based on the classical test theory was 27 items. In comparison, the difference power index was 15 items which were categorized as good.

Another relevant research by Setiawati et al. (2018) by evaluating the psychomotor characteristics of differential gifted tests with classical theory shows that most of the different power index is good, and some items are low or need improvement. The abstract thinking subtest items have many tricksters that are classified as ineffective, and all the different gifted subtests are classified as reliable. Analysis of the quality of the items on the quadratic equation test instrument was also carried out using the classical test theory approach and the Rasch model. It shows the quality of the measurement instrument for concept understanding through the classical test theory approach was of good quality compared to the Rasch model. The test reliability index of the test kits through the classical test theory approach and the Rasch model is moderate (sufficient). Meanwhile, the difficulty level index through the classical test theory approach does not have good quality. Simultaneously, the Rasch model shows various levels of difficulty, namely easy, difficult, and very difficult. The distinguishing power of the pre-test instrument question understanding of the concept through the two approaches was categorized as evil.

Analysis of final test item items using the classical theory approach was also carried out by Sunarmi et al. (2016) to show that the reliability of the class X Final Semester Examination test is sufficient, while for class XI, it is low. The validity of the item content for both tests was 87.5% and 85.83%, and the construct validity was 95.71% and 84.29%. The percentage of the validity of the second language test was 89.16% and 91.67%. The rate of difficult items in class X was more significant than the medium and easy items. In contrast, for class XI, the moderate items were more significant than the complex and easy items. Next, the item difference between the two tests has terrible and destructive criteria. Meanwhile, the effectiveness of distractor for class X items with excellent and good standards was 70%, while for class XI items were 20%. Next, analyze the quality of the statistical exam questions by comparing the CTT and the Rasch Model (Fernanda & Hidayah, 2020). The results showed that 21 items met the difficulty level and different power index, while the Rasch model consisted of 42 qualified items, then the other items needed to be evaluated. So the Rasch model is better than CTT.

Analysis of the items in the statistical and psychometric test instruments used the classical test theory approach and the item response theory based on the difficulty level and the item difference index (Awopeju & Afolabi, 2016). The results showed that the two methods had a comparable number of item characteristics based on these two parameters—the use of the two instruments in the development of the national exam. A comparison of the two theoretical approaches was also carried out by Bichi et al. (2019) in analyzing the item's characteristics in Chemistry subjects. The results showed that the CTT and IRT approaches were effective and reliable in analyzing test items that gave similar results. Other research results indicate that there is no standardization or content validation process for the instruments used.

Based on the explanation above, the problem in this study is the analysis of the characteristics of the items on the mathematical problem-solving ability test instrument using the classical test theory approach. Item analysis studies for mathematics problem-solving ability test instruments are rarely carried out using the classical theory approach. Also, the use of the R and QUEST programs is still lacking in analyzing the characteristics of the instrument items. Subsequent studies in this study provide information and serve as a reference in applying the technique of determining content validity in the development of test instruments. The preliminary analysis in this study is the content validity of the test instrument for the ability to solve mathematical problems using qualitative validity by experts. Then, the quantitative validity uses the biserial point coefficient. The reliability of the test using an internal consistency approach is the Alpha-Cronbach method.

▪ METHOD

This research uses descriptive quantitative analysis to analyze the item's characteristics on the test instrument of mathematical problem-solving abilities. The research stages consisted of (1) determining the purpose of instrument preparation, (2) developing instrument item indicators, (3) compiling instrument items, (4) content validation, (5) revision based on input from expert validators, and (6) making the final assessment instrument. The next stage, (7) data collection of trial results, and (8) analysis of test results using the classical test theory analysis approach. The collecting data in research with the questionnaire method and online test using google form. The questionnaire method was used to collect validator data on the test instrument, while the test method was to collect data on high school students' mathematics problem-solving abilities. Student responses to the test instrument were dichotomous data as many as 359 respondents. The test instrument consists of 10 items consisting of linear program material, probability, permutations and combinations, functions, straight line equations, circles, and trigonometry. The data analysis technique used the item characteristic analysis based on the classical test theory approach with the R and QUEST programs. The characteristics of the items consist of the level of difficulty and differentiation parameters as well as the effectiveness of the distractor.

▪ RESULTS AND DISCUSSION

Validity

Validity is a condition in which the instrument can measure precisely based on what is to be measured. According to Miller et al. (2009), a good test instrument fulfills three characteristics: validity, reliability, and reusability. Empirical validity testing consists of content validity, construct validity, and criterion validity (Azwar, 2017). The instrument consists of 10 items to measure the problem-solving ability test of middle school students. Validity testing is carried out by reviewing the quality of the items before testing the party who is the subject of the study.

Polit and Beck (2006) state content validity as the extent to which the evaluation instrument consists of sufficient items to construct an assessment. In line with that, Wynd et al. (2003) stated that content validity refers to the evidence needed to determine the extent of the instrument adequately in sampling the corresponding research domains. In this sense, content validity is generally understood as the degree to which a sample of items represents an adequate operational definition of a good instrument construct (Polit and Beck, 2006). In this study, the content validity was based on qualitative expert reviews and quantitative analysis of the expert judgment. Also, quantitative validity is to see the level of validity of each item on the test instrument. The qualitative content validity is based on expert reviews and quantitative analysis using the Aiken index (Aiken, 1985), while quantitative validity uses biserial point correlation.

Qualitative validity

The validity of the content is qualitatively based on the expert judgment of as many as four validators. The validators each have a background in mathematics, each of whom has a doctorate in research and evaluation of education and mathematics education, and 2 of them are students of a doctoral program in educational research and evaluation—collecting data on the validation of the problem-solving ability test instrument using google form. The validation technique uses the panel method by conducting a study for each item based on the suitability of the instrument aspect with the indicator, the usefulness of the indicator with the item, the suitability of the item substance, the clarity of the sentence, the sentence is not confusing. The format of writing, symbols, and pictures is quite clear. The results of the qualitative review by several experts showed that each item had conformity with the mathematical problem-solving indicator. However, technically, the writing of a few items still needed improvement.

The quantitative content validation analysis uses the Aiken index based on scoring the relevance of items with indicators using a score of 1 = irrelevant, 2 = less relevant, 3 = quite relevant, 4 = relevant, and a score of 5 = very relevant. They were categorizing the validity of the instrument items based on the index. If the index is less than or equal to 0.4, then the validity is less; 0.4 - 0.8 medium validity category; and if it is more significant than 0.8, it is said to be valid (Retnawati, 2018). The results of the content validation analysis showed that for each item, items 6 and 8 were obtained with moderate validity, each of which had a value of 0.75, while the other items had a validity level. Then, based on the Aiken validity table with a total number of evaluations of 4 validators on a scale of 5 having an index at intervals of 0 and 1.0 (Aiken, 1985).

The subsequent content validity using Lawshe's CVR formula uses essential criteria, is useful but not essential, and is not required. Analysis of calculating the value of the Content Validity Ratio (CVR) using the formula $CVR = 2\left(\frac{ne}{N}\right) - 1$, where ne and N each state the number of experts who provide validation with essential criteria and the number of experts who participated in conducting the validation. Content Validity Ratio (CVR) is a statistical item that is useful for rejection or acceptance of items in a questionnaire and is internationally recognized as an assessment technique for confirming the validity of content (Polit et al. 2006).

According to Polit et al. (2007), items with a CVR value of 0.78 or higher with three or more experts can be considered evidence of good content validity. However, if an item does not meet that category, the item will be removed from the instrument for instrument testing purposes. The research results show that each item has a CVR value greater than 0.78, so that the item has good content validity. Thus, based on the content validity analysis using the Aiken index and the Lawsche CVR formula, it shows that each item of the test instrument for the ability to solve mathematics problems can reach the measurement stage of the test instrument.

Quantitative validity

Quantitative testing of instrument items can be done using statistical analysis techniques, namely biserial point correlation. Biserial point correlation is a unidirectional relationship between variables where the magnitude of the first variable score occurs together with the importance of the score for other variables. The low score of the first variable appears together with the low score of the other variables. According to Ebel & Frisbie (1986), the biserial point correlation shows the correlation between the test item score and the total score for each test taker. Biserial with positive and high scores indicates the tendency of those with high scores to answer correctly and those with low scores to give wrong answers. On the other hand, negative biserial scores provide information about the ability of testees with high scores to give wrong answers in answering items. In contrast, testes with low scores are correct in answering these items. In analyzing items to select good items, items with negative biserial should be excluded from the model.

The item analysis technique uses the correlation coefficient between the instrument item scores. Items are said to be valid if the items on the instrument have a score with a significant correlation coefficient with the total score of the instrument. The following shows the results of the calculation analysis using the R software in Table 1, which is the biserial point correlation value for each item on the test of mathematical problem-solving abilities.

Table 1. Correlation value of question points biserial points

Item	Biserial Points Correlation
B1	0.51
B2	0.62
B3	0.60
B4	0.64
B5	0.54
B6	0.30
B7	0.55
B8	0.26
B9	0.46
B10	0.68

Table 1. above provides information related to the correlation value of biserial points for each item. Each item has a positive value biserial point correlation coefficient, which indicates that the relationship between variables is unidirectional. The value of the first variable occurs together with the significant score of the other variables, and the low score of the first variable occurs together with the low score of the other variables. The biserial point correlation coefficient for each item above shows the validity coefficient where item 10 has the highest validity coefficient of 0.68 while the lowest validity coefficient is item 8, which is 0.26. It indicates that based on the content validity quantitatively by identifying the biserial point correlation coefficient, it shows that each item of the test kit can precisely measure mathematical problem solving ability to be used for the measurement stage of the test instrument.

Reliability

Reliability or reliability is a coefficient indicating the level of consistency or consistency of the measurement results of a test (Mardapi, 2017). The consistency of these measurement results can be identified by carrying out several measurements of the same group of subjects, and then the results are relatively the same where the aspects measured in the issue have not changed. The test reliability index can be calculated using several approaches: the retest (test-retest), the parallel form approach, and the internal consistency approach.

In this study, the test reliability index using the internal consistency approach was the Alpha-Cronbach method. The assumptions underlying the classical test theory using this method are that there is no correlation between the actual score and the error score. The mean random error of measurement is zero (Allen & Yen, 1979). According to Crocker & Algina (2008), the reliability of the reliability index of the test is obtained using the Alpha-Cronbach formula. The reliability index of a good test is a minimum of 0.70 (Linn, 1989; Allen & Yen, 1979). Analysis of the calculation of test reliability using the Alpha-Cronbach formula, namely $r = \frac{k}{k-1}\left(1 - \frac{\sum S_b^2}{S_t^2}\right)$ (Mehrens & Lehmann, 1984), obtained the reliability index of the mathematics problem-solving ability test of 0.83. It means that the test instrument is suitable. According to Guilford (1956), the same thing shows that the test instrument of mathematical problem-solving ability has a very high level of reliability, which means that repetition of the test will produce stable results. So, the test instrument for the ability to solve mathematical problems has a very high level of reliability, which means that repetition of the test will produce stable results. Thus, the instrument is reliable (Crocker & Algina, 2008) to measure high school students' problem-solving abilities.

Analyze the characteristics of the instrument items

Analysis of the characteristics of the items used in this study was the classical test theory model using the QUEST software. Characteristics of items consist of parameters of difficulty level and the different power of the questions, and the effectiveness of the distractor. The index of difficulty level and different strength for each item was obtained by calculating analysis using QUEST software. In contrast, the effectiveness of the distractor was obtained by identifying the ratio of the number of test-takers answering with a specific answer choice and the number of test-takers.

Difficulty level

The difficulty level of the item is the opportunity to answer a question correctly at a certain ability level. Analysis of calculating the level of difficulty can be done using proportion correct, linear difficulty index, Davis index, and Bivariate scale. For proportion correct, it is the ratio of the correct answer to the number of answerers (Azwar, 2002). According to Sumintono & Widhiarso (2015), categorizing the difficulty level of items using a logit scale. If the difficulty level index is more significant than 1.0, then the category is very difficult; index 0 - 1.0 difficult category; index -1.0 - 0 easy categories; and the difficulty index is less than -1.0 with straightforward category. The following shows the results of the calculation analysis using the QUEST software in Table 2, which is the level of difficulty for every ten items on the mathematics problem-solving ability test.

Table 2. Difficulty levels of question items

Item	***Threshold***
B1	-0.87
B2	-0.36

B3	-0.96
B4	-0.19
B5	0.18
B6	1.84
B7	-0.64
B8	1.27
B9	-0.26
B10	-0.01

Table 2. above provides information regarding the level of difficulty for each item. Each item has a threshold value where items 6 and 8 are with a very difficult level and item 5 with a difficult level. Other items with an easy level are 1, 2, 3, 4, 5, 7, 9, and 10, with their respective difficulty level values at intervals of -1.0 and 0. The items with the highest difficulty level are point 6 with a difficulty value of 1.84, while point 3 is the item with the lowest or easy difficulty level of -0.96.

Distinguishing power

The difference power of an item is the ability of an item to distinguish test participants who have high and low abilities. Also, the function of distinguishing power is to identify the smallest individual differences among the test takers. Determination of item difference power can be done using correlation techniques, namely biserial, biserial, phi, and tetrachoric points techniques (Ebel & Frisbie, 1986). In this study, the calculation of the power difference index used biserial correlation. The following shows the results of the calculation analysis using R software, the biserial correlation value for every ten items in the mathematics problem-solving ability test.

Table 3. Correlation value of question points biserial points

Item	Biserial Correlation
B1	0.65
B2	0.78
B3	0.78
B4	0.81
B5	0.68
B6	0.41
B7	0.69
B8	0.35
B9	0.58
B10	0.85

Based on Table 3 above provides information related to the biserial correlation value for each item. Each item has a biserial correlation value greater than 0.30, which indicates that each item received functions well to distinguish the ability of high and low-ability test takers. Item 10 is the best item to determine the test taker's ability with a different power index of 0.85. In contrast, the lowest item to distinguish the test taker's ability with an index of 0.35 is item 8.

Then, the distinguishing power for each item provides information related to the item that can differentiate the ability of the test taker. By definition, the distinguishing power is the correlation between a test item score and the total score and is called the biserial point correlation. The technique for determining the item difference index uses biserial correlation. According to Mardapi in Istiyono (2018), determining the criteria for item power where items can be accepted with a more significant difference index 0.30, indexes between 0.10 to 0.30 are items with less difference, while items with an index smaller than 0.10 are items that cannot be used. In the measurement stage for data collection. Meanwhile, according to Rao et al. (2016), the item power index is more significant than 0.4 with suitable criteria, the index is between 0.3 to 0.39 in the good category, the index is between 0.2 to 0.29 in the sufficient category and needs to be corrected, while the index is smaller by 0.2 for the bad category and the questions.

The results showed that each item had a biserial correlation value greater than 0.30, which indicated that the problem-solving ability test items functioned well to distinguish the abilities of high and low-ability test takers. The item that was best for distinguishing the test taker's ability with a different power index of 0.85 was item 10. At the same time, the lowest item in distinguishing the test taker's ability was item 8 with an index of 0.35.

Distractor effectiveness

The characteristics of multiple-choice items are one question and several answer choices with only one correct answer. Another answer option is called a distractor. Distractors are the answer choices for each item that distract the test-takers from answering the questions. Multiple-choice items consist of one question with three or four answer choices where one answer is the most correct. Analysis of the distractor function can be determined based on the analysis of the distribution pattern of the test participants' responses for each item by choosing the choice of answers to the possible answers that have been paired on each item (Sudjiono, 2005). Analysis of the calculation of the effectiveness of the distractor by comparing the number of students who choose a particular answer option with the number of test-takers. According to Fernandes in Istiyono (2018), the distractor can be said to be functioning well where at least 2% or 0.02 of all test participants gave incorrect responses to the items. The results of calculating the effectiveness of the distractor for each item are given in Table 4 below.

Table 4. Item distractor effectiveness ratios

Item	Effectiveness Ratio
B1	0.36
B2	0.44
B3	0.35
B4	0.47
B5	0.53
B6	0.76
B7	0.40
B8	0.69
B9	0.46
B10	0.50

Table 4 shows that each item has a distractor effectiveness ratio, namely the number of test-takers choosing a particular answer compared to the number of test-takers. The results of the calculation analysis show that the distractor for each item of the mathematics problem-solving ability test is functioning well with the distractor effectiveness ratio greater than 0.02 or 2%. The item with the distractor level that is functioning properly is point 6, with an effective ratio of 0.76, while point 3 is the item with the lowest distractor effectiveness ratio of 0.35.

▪ CONCLUSION

The reliability of the test of mathematical problem-solving abilities using an internal consistency approach indicates that the test instrument is suitable. The test instrument has a very high level of reliability where repetition of the test will produce stable results. The qualitative validity of the content by the expert states that for each item, it corresponds to an indicator of solving mathematical problems. It's just that technically writing a few items still needs improvement. Then, quantitative validity by expert judgment using the Aiken index obtained 6 and 8 with moderate content validity while other items had very valid validity levels. Also, the content validity using Lawshe's CVR formula also shows good content validity.

Quantitative validity using biserial point correlation was obtained for each valid item. This shows that for each item, the test instrument can measure precisely the ability to solve mathematical problems so that the use of the instrument reaches the measurement stage. Analysis of item characteristics using the classical test theory approach shows that the test instrument has a hard level item consisting of 2 items and eight other items with an easy level. Then, for each item, the test taker was able to distinguish the test taker's ability and the effectiveness of the distractor to function correctly.

The contribution of this research is the use of classical test theory in analyzing the characteristics of the items consisting of the difficulty level parameters, the distinguishing power of the questions, and the distractor effectiveness. Also, this study provides reference information in conducting content validity based on qualitative reviews and quantitative analysis by expert judgment. Another contribution is to contribute information to teachers or researchers regarding the use of good test instruments for mathematical problem-solving abilities. While the limitations in this study are based on data collection techniques using google Forms with unsupervised quality and the student processing time is not well controlled, the data obtained does not reflect the students' abilities. Therefore, the suggestions in this study are for data collection techniques using an online test application using the duration of the processing time to obtain accurate data regarding the ability to solve mathematical problems.

▪ REFERENCES

Adegoke, B.A. (2013). Comparison of item statistics of Physics achievement test using classical test and item response theory frameworks. *Journal of Education and Practice*, *4*(22), 87-96.

Aiken, L.R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement, 45*, 131-134.

Alifa, T.F., Ramalis, T.R., & Purwana, U. (2018). Karakter Tes Penalaran Ilmiah Siswa SMA Materi Mekanika Berdasarkan Analisis Tes Teori Respon Butir. *Jurnal Inovasi dan Pembelajaran Fisika*, *5*(1), 80-89.

Allen, M.J., & Yen, W. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Publishing Company.

Awopeju, O.A., & Afolabi, E.R.I. (2016). Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination. *European Scientific Journal*, *12*(28), 263-284.

Azwar, S. (2009). *Tes, prestasi, fungsi, dan pengembangan pengukuran prestasi belajar* [Test, achievement, function, and development of learning achievement measurements]. Yogyakarta: Pustaka Pelajar Offset.

Azwar, S. (2013). *Metode Penelitian* [Research methods]. Yogyakarta: Pustaka Pelajar Offset.

Azwar, S. (2017). *Reliabilitas dan Validitas* [Reliability and Validity]. Yogyakarta: Pustaka Pelajar Offset.

Bichi, A.A., Embong, R.B., Mamat, M., & Maiwada, D.A. (2015). Comparison of Classical Test Theory and Item Response Theory: A Review of Empirical Studies. *Australian Journal of Basic and Applied Sciences*, *9*, 549-556.

Bichi, A.A., Embong, R., Talib, R., Salleh, S., & Ibrahim, A. (2019). Comparative Analysis of Classical Test Theory and Item Response Theory using Chemistry Test. *International Journal of Engineering and Advanced Technology (IJEAT)*, *8*(5), 1260-1266.

Crocker, L., & Algina, J. (2008). *lntroduction to classical & modern test theory*. USA: Cengage Learning.

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*, *12*(3), 449-460.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement*. New Jersey: Prentice Hall, Inc.

Eleje, L.I., Onah, F.E., & Abanobi, C.C. (2018). Comparative Study of Classical Test Theory and Item Response Theory Using Diagnostic Quantitative Economics Skill Test Item Analysis Results. *European Journal of Educational & Social Sciences*, *3*(1), 57-75.

Fernanda, J.W., & Hidayah, N. (2020). Analisis Kualitas Soal Ujian Statistika Menggunakan Classical Test Theory dan Rasch Model [Analysis of the Quality of Statistics Exam Questions Using Classical Test Theory and the Rasch Model]. *SQUARE: Journal of Mathemtics and Mathematics Education*, *2*(1), 49-60. http://dx.doi.org/10.21580/square.2020.2.1.5363

Gilbert, G.E., & Prion S. (2016). Making Sense of Methods and Measurement: Lawshe's Content Validity Index. *Clinical Simulation in Nursing*, *12*(12), 530-531.

Guilford, J. P. (1956). *Fundamental Statistic in Psychology and Education* (3rd ed.). New York: McGraw-Hill Book Company, Inc.

Guler N., Uyanik, G.K., & Teker, G.T. (2014). Comparison of classical test theory and item response theory in terms of item parameters. *European Journal of Educational Research*, *2*(1), 1-6.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Education Measurement: Issues and Practice*, *12*(3), 38-47.

Hinkin, T. R., Tracey, J.B., & Enz, C.A. (1997). Scale construction: developing reliable and valid measurement instruments. *Journal of Hospitality & Tourism Research*, *21*, 001, 100-120.

Ikhsanudin & Subali. (2018). Content validity analysis of first semester formative test on biology subject for senior high school. *J. Phys.: Conf. Ser.* 1097 (012039). 1-9. doi :10.1088/1742-6596/1097/1/012039

Istiyono, E. (2018). *Pengembangan Instrumen Penilaian dan Analisis Hasil Belajar Fisika dengan Teori Tes Klasik dan Modern* [Development of Assessment Instruments and Analysis of Physics Learning Outcomes with Classical and Modern Test Theory]. Yogyakarta: UNY Press.

Jabrayilov, R., Emons, W.H.M., & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Applied Psychological Measurement, 40*(8), 559-572.

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 63-575.

Linn, R. L. (1989). *Educational measurement*. New York: Mac Millan Publishing.

Mardapi, D. (1998). *Analisis butir dengan teori klasik dan teori respon butir* [Item Analysis with Classical Theory and Item Response Theory]. *Jurnal Kependidikan, 28*(1), 25–34.

Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan non tes* [The technique of preparing test and non-test instruments]. Yogyakarta: Mitra Cendekia Offset.

Mardapi, D. (2016). *Pengukuran, Penilaian, dan Evaluasi Pendidikan* [Measurement, assessment and evaluation of education]. Yogyakarta: Pustaka Pelajar.

Mardapi, D. (2017). *Pengukuran, penilaian dan evaluasi pendidikan* [Measurement, assessment and evaluation of education]. Yogyakarta: Parama Publishing.

Mehrens, W. A. & Lehmann, I. J. (1984). *Measurement and evaluation in educational and psychology*. New York: Rinehart and Winston.

Miller, M.D., Lin, R.L., & Gronlund, N.E. (2009). *Measurement and Assessment in Teaching*. New Jersey: Pearson Educational, Inc.

Oriondo, L.L. & Dallo-Antonio, E. (1998). *Evaluation educational outcomes*. Florentino St: Rex Printing Compagny, Inc.

Pido, S. (2012). Comparison of item analysis results obtained using item response theory and classical test theory approaches. *Journal of Educational Assessment in Africa, 7*, 192–207.

Polit, D.F., & Beck, C.T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, 29, 489497.

Polit, D., Beck, C., & Owen, S. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing and Health, 30*(4), 459-467.

Rao, C., L, K. P. H., Sajitha, K., Permi, H., & Shetty, J. (2016). Item analysis of multiple choice questions : Assessing an assessment tool in medical students. *International Journal of Educational and Psychological Researches*, 2–5.

Retnawati, H. (2016). *Validitas, Reliabilitas dan Karakteristik Butir* [Validity, Reliability and Item Characteristics]. Yogyakarta: Nuha Medika.

Royce, H. (2009). Comparison of the item discrimination and item difficulty of the quick-mental aptitude test using CTT and IRT methods. *The International Journal of Educational and Psychological Assessment, 1*(1), 12-18.

Sarea, M. S., & Ruslan, R. (2019). *Karakteristik butir soal: Classical test theory vs item response theory* [Item characteristics: Classical test theory vs item response theory]. *Jurnal Kependidikan, 13*(1), 1-13.

Setiawati, F. A., Izzaty, R. E., & Hidayat, V. (2018). *Evaluasi Karakteristik Psikometrik Tes Bakat Differensial dengan Teori Klasik* [Evaluation of Psychometric Characteristics of Differential Aptitude Test with Classical Theory]. *Humanitas Indonesian Psychological Journal*, *15*(1), 46-61.

Subali, B. (2016). *Pengembangan Tes: Beserta Penyelidikan Validitas dan Reliabititas secara Empiris* [Test Development: Along with Empirical Validity and Reliability Investigation]. Yogyakarta: UNY Press.

Sudjiono, A. (2005). *Pengantar Evaluasi Pendidikan* [Introduction to Educational Evaluation]. Jakarta: Paja Grafindo Persada.

Suhariyono, Sriyono, & Ngazizah, N. (2014). *Akurasi pendekatan classical test theory dan pendekatan item response theory dalam menganalisis soal UAS Fisika semester genap kelas X SMA Negeri di Purworejo tahun pelajaran 2013/2014* [The accuracy of the classical test theory approach and the item response theory approach in analyzing the UAS Physics questions in the even semester of class X SMA Negeri in Purworejo academic year 2013/2014]. *Radiasi*: *Jurnal Berkala Pendidikan Fisika*, *5*(2), 75-79.

Sunarmi, Prasetyo, T. I., & Ramadhiana, C. B. (2016). *Analisis butir soal ulangan akhir semester gasal Biologi kelas X dan XI tahun pelajaran 2016/2017 di SMAN 1 Kampak berdasarkan teori tes klasik* [Analysis of test items at the end of odd semester of Biology class X and XI for the 2016/2017 academic year at SMAN 1 Kampak based on classical test theory]. *Jurnal Pendidikan Biologi*, *8*(1), 27-31.

Vakili, M.M., & Jahangiri, N. (2018). Content Validity and Reliability of the Measurement Tools in Educational, Behavioral, and Health Sciences Research. *Journal of Medical Education Development*, *10*(28), 105-117.

Wynd, C.A., Schmidt, B., and Schaefer, M.A. (2003). Two quantitative approaches for estimating instrument content validity. *Western Journal of Nursing Research*, *25*, 508-518.

Yudiana, Y., Hidayat, Y., Hambali, B., & Slamat, S. (2017). Content Validity Estimation of Assessment Instrument Based on Volleyball Information System of Volleyball Learning: Field Research. *J. Phys.: Conf. Ser.: Mater. Sci. Eng., 180* (012230). 1-6.