Library Progress International
Vol.44 No. 3, Jul-Dec 2024: P. 19352-19363.

Print version ISSN 0970 1052
Online version ISSN 2320 317X

Original Article

Available online at www.bpasjournals.com

Analysis of Item Characteristics of Mathematics Test Instruments with ITEMAN 4.3 Application

Sa'adatul Ulwiyah[1*], Annisa Fitriani[2], Raden Rosnawati[3], Moch. Arifudin[4], Jumriani Sultan[5]

[1]Student, Department of Educational Research and Evaluation, Yogyakarta State University, Yogyakarta, Indonesia
[2]Student, Department of Educational Research and Evaluation, Yogyakarta State University, Yogyakarta, Indonesia
[3]Doctor, Department of Mathematics Education, Yogyakarta State University, Yogyakarta, Indonesia
[4]Student, Department of Guidance and Counseling, Yogyakarta State University, Yogyakarta, Indonesia
[5]Student, Department of Educational Research and Evaluation, Yogyakarta State University, Yogyakarta, Indonesia
[1]saadatul.2022@student.uny.ac.id,[2]annisafitriani.2022@student.uny.ac.id,
[3]rosnawati@uny.ac.id,[4]mocharifudin.2022@student.uny.ac.id, [5]jumrianisultan@student.uny.ac.id
[1]0009-0004-0496-2204, [2]0009-0006-9646-8776, [3]0000-0002-8841-0412, [4]0009-0002-7058-1747, [5]0009-0008-0690-1535

How to cite this article: Sa'adatul Ulwiyah, Annisa Fitriani, Raden Rosnawati, Moch. Arifudin, Jumriani Sultan (2024) Analysis of Item Characteristics of Mathematics Test Instruments with ITEMAN 4.3 Application. *Library Progress International,* 44(3), 19352-19363..

ABSTRACT

The quality of test instruments in education is very important to accurately measure students' abilities and achievements. This study aims to evaluate the quality of mathematics tests and describe the results of item analysis which includes reliability, difficulty level, discrimination, and the effectiveness of distractors. This research uses quantitative descriptive method with the help of ITEMAN 4.3 application. The tested test instrument consisted of 20 multiple choice questions with four answer choices, which were given to 70 seventh grade students. The results showed an Alpha reliability value of 0.610, indicating a high level of reliability for the instrument. In terms of difficulty level, 2 items (10%) were classified as difficult, 15 items (75%) as medium, and 3 items (15%) as easy. In terms of discrimination, 1 item (5%) was classified as very good, 6 items (30%) as good, 4 items (20%) as fair, and 9 items (45%) as not good. Regarding the effectiveness of distractors, 18 items (90%) have effective distractors, while 2 items (10%) have ineffective distractors.Overall, considering the reliability, difficulty level, discrimination, and effectiveness of disctractors, 10 items (50%) were in the good quality category, 1 item (5%) in the good enough category, and 9 items (45%) in the not good quality category. The results of this study indicate that although most of the items are of good quality, there are some items that need to be improved to improve the overall quality of the test. This analysis is important to ensure that the test instrument used is able to provide an accurate and fair assessment of students' abilities.

KEYWORDS: Analysis of Item Characteristics, Iteman 4.3, Classical Test Theory, Mathematics Test Instrument, Evaluation of Education.

.

Introduction:

The quality of education can be assessed through the evaluation of student learning outcomes. This evaluation process involves measurements designed to gauge the attainment of learning objectives across various scientific disciplines, in alignment with the established curriculum [1]. To make useful educational decisions, the information used must be accurate, reliable, and relevant to the problem at hand [2]. Research is of good quality when the risk of bias is low [3]. Thus, measurement results provide important information for education providers in making decisions related to students.

The evaluation process in education relies heavily on measurement. Measurement is the process of providing values or numbers that reflect the ability of students in a subject. According to [1], educational measurement involves the activity of quantifying symptoms or objects, such as motivation, achievement, and confidence, which are then expressed in the form of numbers. Measuring instruments are used to provide information about a person's position in the measured attribute. To get accurate measurement results, a measuring instrument is needed that has a high level of validity and reliability.

According to [2], measuring instruments or tests can be interpreted as devices used to obtain samples of individual behavior. A similar view is also conveyed by [3], which states that a test is one type of instrument used to make measurements with the aim of collecting information about the characteristics of an object. Tests can be understood as a set of questions designed to be answered with the purpose of assessing an individual's ability level or uncovering specific characteristics of the person being tested [4].

A measurement tool or test instrument that is often used to evaluate learners' learning outcomes is a collection of questions. It is important to ensure that the set of questions used is of good quality in order to accurately measure learners' abilities. A good instrument is an instrument that is able to produce accurate data and provide accurate information as well, so that the information obtained from the measurement results can accurately describe the ability of students [5], [6], [7].

Evaluating the quality of test instruments is crucial for assessing both the overall quality of a test and the quality of each individual item [8]. The main purpose of item analysis is to obtain detailed information about the characteristics of each test item and to conduct empirical evaluations [9]. This process aims to enhance the quality of test instruments by revising or removing ineffective items and provides valuable insights for teachers regarding students' understanding of the subject matter [10]. The findings from the analysis serve as a basis for evaluating the quality of the questions, assessing student learning outcomes, and indicating the success of the educational institution or unit [9].

The evaluation of item quality can be conducted by examining three key aspects: difficulty level, discrimination, and distractor effectiveness [6]. Difficulty level is used to categorize questions as easy, medium, or difficult. Discrimination assesses the ability of a question to differentiate between students with high and low abilities. Meanwhile, distractor effectiveness measures how well the incorrect answer options perform in terms of their functionality [11].

One application that can assist in analyzing test instrument items is the ITEMAN 4.3 Application. This application can automatically analyze the difficulty level, discrimination and distractor effectiveness on each test item [12]. The results of the analysis can help identify the weaknesses and strengths of each item, making it easier to make the necessary improvements.

This study seeks to assess the quality of mathematics test instruments and analyze their performance in terms of reliability, difficulty level, discrimination, and distractor effectiveness. The instrument under review is a mathematics test administered to seventh-grade students at a junior high school in Yogyakarta, Indonesia. The purpose of this item analysis is to enhance the quality of evaluation tools used for measuring students' mathematical skills.

2) Research Methods:

The research method used is a quantitative descriptive approach. The study aims to assess the quality of the instrument and to describe the results of the analysis of the test items in terms of reliability, difficulty, discrimination and distractor effectiveness. The research subjects were 70 seventh grade students. The data collection technique used was documentation. The data analysed consisted of the students' responses to a mathematics test instrument. This test instrument contained 20 multiple-choice items, each with four response options (A, B, C, and D).

The characteristics of the items, including reliability, difficulty, discrimination and distractor effectiveness, were analysed using the classical test theory approach. ITEMAN 4.3 was used for item analysis. The data obtained from the students' responses were processed using this application. The results of the analysis were presented as numerical indices, which were then compared with criteria established on the basis of the classical test theory approach. The results of the analysis were then interpreted to determine the characteristics of each test item, such as reliability, difficulty, discrimination and distractor effectiveness.

(a) Reliability

A reliable instrument is one that produces consistent results over time [6]. In classical test theory, reliability is linked to the concept of measurement precision [13], [14]. Reliability level categories are based on the interpretation of the reliability index according to [15], as shown in Table 1.

Table 1. Reliability Coefficient

Reliability Coefficient	Level of Reliability
0.80 - 1.00	Very High
0.60 - 0.80	High
0.40 - 0.60	Fair
0.20 - 0.40	Low
0.00 - 0.20	Very Low

An instrument is considered to be of good quality if it falls within at least the high category.

(b) Difficulty Level

The item difficulty index is defined as the proportion or percentage of examinees who respond correctly to an item [2], [16], [17], [18] If a high percentage of students answer correctly, the item is considered easy, and vice versa [19], [20]. The difficulty level, denoted by p (proportion), ranges from 0 to 1 [19]. A higher difficulty level indicates an easier item, whereas an item with $p = 0.00$ means that no students answered it correctly, and $p = 1.00$ means that all students answered it correctly [18], [21]. The criteria for difficulty levels are shown in Table 2.

Table 2. Level of Difficulty

Level of Difficulty	Criteria
> 0.70	Easy
0.30 - 0.70	Medium
< 0.30	Difficult

An instrument is considered to be of good quality if the items fall into the medium difficulty category.

(c) Discrimination

The discrimination of a test item refers to its capability to differentiate between test takers with high abilities and those with low abilities [6], [22]. The discrimination index can range from -1.0 to 1.0; however, a discrimination value below 0.0 suggests an issue with the item [11]. In this study, the discriminant criteria were determined based on the point biserial correlation (Rpbis). The criteria for the discrimination index according to [23] are outlined in Table 3.

Table 3. Discrimination

Discrimination	Criteria
> 0.40	Very good
0.30 - 0.39	Good (little or no revision required)
0.20 - 0.29	Fair (item requires revision)
< 0.19	Not good (item should be discarded)

A good item is one that has a discrimination index in the "Very Good" or "Good" category.

(d) Distractor Effectiveness

Distractor effectiveness refers to how well the incorrect answer choices deceive test takers who do not know the correct answer. The more test takers who select the distractor, the better it functions [6]. According to [3], distractors are considered good if at least 5% of the test participants choose them. On the other hand, [24] states that distractors are effective if chosen by at least 2% of the respondents. The criteria for distractor effectiveness are outlined in Table 4.

Table 4. Distractor Effectiveness

Distractor	Criteria
≥ 2%	Effective
< 2%	Not Effective

A good item is one that has distractors falling into the "Good" or "Effective" category.

(e) Item Quality

The quality of test instrument items is carried out by taking into account reliability, difficulty level, discrimination and distractor effectiveness. Test instrument items are considered to be of good quality if they exhibit high reliability, a moderate difficulty level ($0.30 < p < 0.70$), good discrimination (> 0.30), and effective distractors. [25], interpreted the quality criteria for test instrument items as detailed in Table 5.

Table 5. Quality of Instrument Items

Criteria	Description
Good	Medium level of difficulty, good/fair discrimination, all distractors are effective
Revised Alternative Answer	Medium level of difficulty, good/fair discrimination, there are ineffective distractors
Good Enough	Easy/difficult difficulty level, good/fair differentiation,
Not Good	Not good discrimination

A good item is one that meets all these criteria, ensuring it has high reliability, a moderate difficulty level, good/fair discrimination, and effective distractors.

3] Results:

This research begins with collecting data on students' responses to mathematics test instruments. The instrument analyzed consisted of mathematics test questions on whole number material. The test instrument comprises 20 multiple-choice items with four answer choices. It was administered to 70 seventh-grade students. Data on students' responses to the test instrument were then analyzed using the classical approach with the help of the ITEMAN 4.3 application.

(a) Reliability

The analysis results using the ITEMAN 4.3 application on the mathematics test instrument for seventh-grade students revealed an Alpha reliability coefficient of 0.610. According to the criteria set by [15] this indicates that the math test instrument has a reliability above 0.60, signifying high reliability. The detailed output of the reliability analysis is presented in Table 6.

Table 6. Reliability Coefficient Score With ITEMAN 4.3

Score	Scored items
Alpha	0.610
SEM	2.059
Split-Half (Random)	0.447
Split-Half (First-Last)	0.363
Split-Half (Odd-Even)	0.497
S-B Random	0.618
S-B First-Last	0.533
S-B Odd-Even	0.664

These results demonstrate that the test instrument is consistent and reliable for measuring the mathematical abilities of the students.

(b) Difficulty Level

Information about the difficulty level of each item in the mathematics test instrument for seventh-grade students was obtained through analysis using the ITEMAN 4.3 application. This analysis offers a detailed view of the characteristics of the instrument items, particularly the difficulty level of each item in the test.

Table 7. Analysis of Difficulty Level

Difficulty Level	Criteria	Item Number
> 0.70	Easy	2, 4, 17
0.30-0.70	Medium	1, 3, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 18, 19, 20
< 0.30	Difficult	9, 16

The results, as shown in Table 7, indicate that out of the 20 items in the instrument, 3 items (15%) are classified as easy, 15 items (75%) are classified as medium, and 2 items (10%) are classified as difficult. The diagram illustrating the results of the difficulty level analysis for the mathematics test instrument can be seen in Figure 1.

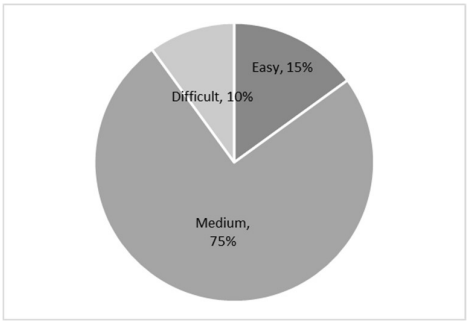

Figure 1. Analysis of Difficulty Level

(c) Discrimination

Based on the analysis results using the ITEMAN 4.3 application for the mathematics test instrument of seventh-grade students, the discrimination of the items varies. This analysis assesses how well the items can differentiate between test takers with high and low ability levels.

Table 8. Analysis of Discrimination

Discrimination	Criteria	Item Number
> 0.40	Very good	14
0.30 - 0.39	Good	1, 3, 5, 7, 19, 20
0.20 - 0.29	Fair	6, 8, 13, 17
< 0.19	Not good	2, 4, 9, 10, 11, 12, 15, 16, 18

According to the results in Table 8, there is 1 item (5%) classified as very good, 6 items (30%) classified as good, 4 items (20%) classified as fair, and 9 items (45%) classified as not good. Items with good or fair criteria need revision, while items classified as not good should be discarded or cannot be used. The diagram depicting the results of the analysis of item discrimination in the mathematics test instrument is shown in Figure 2.

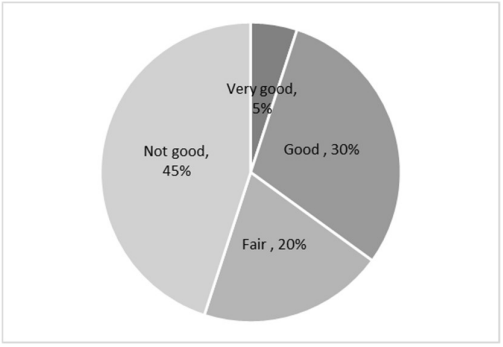

Figure 2. Analysis of Discrimination

(d) Distractor Effectiveness

Using the ITEMAN 4.3 application, an analysis of distractor effectiveness for the mathematics test instrument of seventh-grade students was conducted. This analysis provides insight into how well the incorrect answer choices can mislead test takers who do not know the correct answer.

Table 9. Analysis of Distractor Effectiveness

Distractor	Criteria	Item Number
0.00	Not Effective	2 (B)
0.00	Not Effective	18 (E)

The results indicate that out of the 20 items in the mathematics test instrument, According to the results in Table 9, 2 items (10%) have ineffective distractors. Specifically, item 2 has an ineffective alternative answer choice B, and item 18 has an ineffective alternative answer choice E. The diagram illustrating the results of the distractor effectiveness analysis is shown in Figure 3.

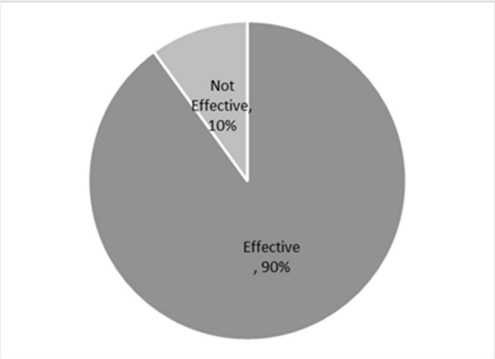

Figure 3. Analysis of Distractor Effectiveness

(e) Item Quality

Incorporating information from the previous analyses, including reliability coefficients, difficulty levels, discrimination, and distractor effectiveness, the quality of the instrument items was assessed using the criteria outlined by [25] The results are as follows:

Table 10. Analysis of Item Quality

Criteria	Item Number
Good	1, 3, 5, 6, 7, 8, 13, 14, 19, 20
Revised Alternative Answer	-
Good Enough	17
Not Good	2, 4, 9, 10, 11, 12, 15, 16, 18

Table 10 shows that out of the 20 items in the instrument, 10 items (50%) are categorized as having good quality, 1 item (5%) is categorized as having fair quality, and 9 items (45%) are categorized as having poor quality. The diagram illustrating the results of the item quality analysis for the mathematics test instrument is presented in Figure 4.

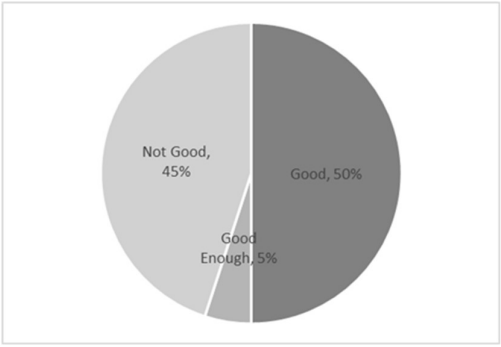

Figure 4. Analysis of Item Quality

4] Discussion:

The analysis using the ITEMAN 4.3 application for the seventh grade mathematics test instrument showed an alpha reliability coefficient of 0.610. This measure assesses the consistency and stability of the test in assessing students' abilities [26]. According to the criteria established by [15], a reliability value above 0.60 is considered to indicate high reliability. Thus, the high reliability of the mathematics test instrument for seventh-grade students indicates a high degree of consistency. This means that the instrument effectively and reliably measures students' mathematical skills and provides a stable representation of their abilities.

In this study, the high reliability of mathematics test instruments is crucial, as it ensures that assessment results are both accurate and consistent, and provide valid insights into students' mathematical abilities. Reliable test instruments allow teachers and policy makers to use these results to design more effective teaching strategies and provide appropriate feedback to students [27], [28] However, it's important to remember that reliability is not the only criterion for assessing the quality of a test instrument. Factors such as validity, difficulty, discrimination and the effectiveness of distractors must also be assessed to fully evaluate the quality of the instrument.

The analysis of the level of difficulty using the ITEMAN 4.3 application revealed a range of difficulty among the test items. Specifically, 3 items (15%) were classified as easy, 15 items (75%) were classified as medium and 2 items (10%) were classified as difficult. Difficult items may discourage students due to their complexity, while easy items may not effectively challenge students' abilities [29] Ideally, test items should be of moderate difficulty, with an index between 0.30 and 0.70 [2] This range provides the best opportunity to accurately assess students' abilities [30], [31]. Therefore, items of medium difficulty are considered optimal as they strike a balance between being too easy and too hard [32]. Consequently, out of the 20 items in the seventh grade mathematics test, 15 items were considered to have an appropriate level of difficulty, while the remaining 5 items were considered to be in need of improvement. Steps should be taken to increase the difficulty of these less effective items.

Following the analysis of item difficulty, several actions can be taken, as outlined by [33]. First, items classified as medium difficulty, which are considered to be good, should be kept in the question bank for future use. Second, for items rated as difficult, there are three possible actions: a) discard the item and refrain from using it in future tests; b) investigate the reasons for the difficulty of the item and, if possible, revise the question to make it clearer and less prone to multiple interpretations. Revised items can be used in subsequent tests; c) Retain the item for use in selection tests where a high level of difficulty is appropriate, as most test takers may not pass. Thirdly, for items that are considered easy, three approaches can be taken: a) remove the item and do not use it in future assessments; b) analyse why the item is too easy to answer and modify the answer choices and question complexity. The improved item can then be used in future tests; c) Retain the item for use in less formal assessments or tests.

Based on the discrimination analysis conducted with the ITEMAN 4.3 application, variations in item effectiveness were observed. The analysis showed that 1 item (5%) met the very good criteria, demonstrating its ability to effectively discriminate between high and low ability students. Items rated as very good are considered to be of high quality and do not require revision. Six items (30%) were rated good, indicating their ability to differentiate between students of different abilities. However, these items could be improved to achieve higher quality. Four items (20%) were judged to be fair, meaning that they partially met the criteria but needed adjustments to improve their discriminatory power. Nine items (45%) were not good, showing significant problems in differentiating

students' abilities. These items are ineffective and should be discarded. To improve these items, efforts should be made to clarify questions to avoid confusion, especially for high ability students, and to ensure that items effectively reflect differences in students' understanding of the material [34].

Based on the results of the discrimination analysis, it is clear that several items rated good or fair need to be revised to improve their quality. Items rated as not good should either be improved or removed from the test instrument. In particular, items with negative discrimination, such as those with index values of -0.029 and -0.087 for items 4 and 16 respectively, should not be used in future assessments as they are of very poor quality [34]. A negative discrimination index suggests that knowledgeable test takers tend to answer these items incorrectly, while less knowledgeable test takers answer them correctly. This suggests that the items may assess content outside the intended scope of the test or have incorrect answer keys [11]. Improving these items will help the mathematics test instrument to produce more accurate results and better differentiate between students' abilities.

The effectiveness of distractors is assessed by counting the number of students who select each answer choice (a, b, c or d) or leave it blank. Effective distractors are those selected by at least 2% of respondents [24]. The analysis of distractor effectiveness showed that out of 20 items, 2 items (10%) had ineffective distractors. Specifically, choice B was ineffective in item 2 and choice E was also ineffective in item 18. This suggests that these choices failed to engage students who did not understand the concept being tested [11], [35]. The ineffectiveness of these distractors means that they did not effectively challenge students who were unfamiliar with the material, potentially leading to incorrect answers from those who guessed. Ineffective distractors are often too obvious or not varied enough, making them less likely to be selected by those who do not understand the material [34].

According to Grounlund in [36], ineffective distractors or response options can affect both the validity and reliability of mathematics testing instruments. Ineffective distractors can reduce the instrument's ability to accurately assess students' understanding and ability. Therefore, it's important to revise items with ineffective distractors. It is important to ensure that all alternative answers are challenging for students who lack understanding, while the correct answer is clear for those who know the material. Improving distractors can increase the reliability of mathematics test instruments, making them better tools for assessing students' understanding of the material.

The analysis of item quality, taking into account difficulty levels, discrimination indices, and distractor effectiveness, reveals several variations in the quality of the 20 items evaluated. According to [25], 10 items (50%) were classified as having good quality. These items exhibit a moderate difficulty level and demonstrate effective or adequate discrimination, meaning they can successfully differentiate between students of varying abilities. Additionally, the distractors in these items effectively challenge students who lack understanding of the material. One item (5%), specifically item number 17, is rated as having fair quality. Although this item is usable, it requires improvements, particularly regarding its difficulty level, which may be either too easy or too challenging, affecting its alignment with the learners' ability levels. Furthermore, the distractors for this item need enhancement. On the other hand, 9 items (45%) were found to be of poor quality. These items face issues with difficulty levels, discrimination, or distractor effectiveness. They may be either too difficult or too easy, leading to inaccurate assessments of students' abilities. The discrimination index for these items might be low, reducing their effectiveness in distinguishing between students of different abilities. Additionally, the distractors may fail to effectively challenge students who do not grasp the material.

The results of this analysis highlight the need to revise and improve the poor quality items. Improvements should focus on adjusting the difficulty index, the discrimination index and the effectiveness of distractors. By addressing these issues, test instruments can achieve greater validity and reliability in accurately assessing students' abilities and understanding of the material.

5] Conclusion:

The research and analysis using the ITEMAN 4.3 application on the mathematics test instrument for seventh grade students revealed several key findings. Firstly, the reliability of the mathematics test instrument was found to have an alpha value of 0.610, indicating high reliability as it exceeds the minimum threshold of 0.60. Secondly, the analysis of item discrimination showed that 1 item (5%) was rated as very good, 6 items (30%) as good, 4 items (20%) as fair and 9 items (45%) as poor. Items rated as good or fair should be revised, while those rated as poor should be discarded. Thirdly, in terms of difficulty, 3 items (15%) were rated as easy, 15 items (75%) as moderate and 2 items (10%) as difficult.

Finally, 2 items (10%) were identified as having ineffective distractors. Combining these findings - reliability, difficulty, discrimination and distractor effectiveness - the quality analysis revealed that out of 20 items, 10 (50%) were of good quality, 1 (5%) was of fair quality and 9 (45%) were of poor quality. Based on these results, it is recommended that the mathematics test instrument for seventh-grade students be revised and improved to increase its overall quality.

6] Acknowledgement:

Gratitude is extended to Lembaga Pengelola Dana Pendidikan (LPDP) for funding the research and publication, as well as to the supervisors for their guidance and evaluation of this publication. Appreciation is also due to Yogyakarta State University and the junior high school that facilitated and supported this research. Additionally, thanks are given to everyone who has provided assistance and support throughout the completion of this research.

7] References:

[1] D. Mardapi, *Pengukuran, Penilaian, dan Evaluasi Pendidikan*. Yogyakarta: Prama Publishing, 2017.

[2] M. J. Allen and W. M. Yen, *Introduction to Measurement Theory*. Monterey: CA: Brooks/Cole Publishing Company, 1979.

[3] D. Mardapi, *Teknik Penyusunan Instrument Tes dan Nontes*. Yogyakarta: Mitra Cendikia Press, 2006.

[4] Widoyoko, *Teknik Penyusunan Instrumen Penelitian*. Yogyakarta: Pustaka Pelajar, 2012.

[5] S. Azwar, *Reliabilitas dan Validitas*. Yogyakarta: Pustaka Pelajar, 2012.

[6] E. Istiyono, *Pengembangan Instrumen Penilaian dan Analisis Hasil Belajar Fisika Dengan Teori Tes Klasik dan Modern*. Yogyakarts: UNY Press, 2020.

[7] H. Retnawati, *Analisis Kuantitatif Instumen Penelitian (Panduan Peneliti, Mahasiswa, dan Psikometrian*. Yogyakarta: Prama Publishing, 2016.

[8] R. Irawati, E. Yusliana, and S. Budiawanti, "Analisis Butir Soal Ujian Akhir Semester Gasal Menggunakan Program Anbuso di SMA Negeri 1 Boyolali," *Jurnal Materi dan Pembelajaran Fisika*, vol. 10, no. 1, pp. 11–20, May 2020, doi: https://doi.org/10.20961/jmpf.v10i1.42084.

[9] M. S. Sarea and S. Hadi, "Analisis Kualitas Soal Ujian Akhir Semester Mata Pelajaran Kimia SMA di Kabupaten Gowa," *Jurnal Evaluasi Pendidikan*, vol. 3, no. 1, pp. 35–43, Mar. 2015.

[10] C. Boopathiraj and K. Chellamani, "Analysis of Test Items on Difficulty Level and Discrimination Index The Test For Reseach in Education," *International Journal of Social Science & Interdisciplinary Research*, pp. 189–193, 2013.

[11] H. Djiju *et al.*, *Analisis Instrumen Penelitian dengan Teori Tes Klasik dan Modern Menggunakan Program R*, 1st ed. Yogyakarta: UNY Press, 2022.

[12] N. Huda, "Item and test Analysis (ITEMAN) 4.3," UIN Malang. Accessed: Jun. 02, 2023. [Online]. Available: http://repository.uin-malang.ac.id/8322/1/8322.pdf

[13] L. W. Anderson and D. R. Krathwohl, *Taxonomy for Learning Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longmann, 2001.

[14] L. Crocker and J. Algina, *Introduction to Clasical and Modern Test Theory*. New York: Holt, Rinehart and Wiston, Inc, 1986.

[15] J. P. Guilford, *Fundamental Statistic in Psychology and Education*, 3rd ed. New York: McGraw-Hill Book Compani, Inc, 1956.

[16] S. Zainudin, K. Ahmad, N. M. Ali, and N. F. A. Zainal, "Determining Course Outcomes Achievement Through Examination Difficulty Index Measurement," *Procedia Soc Behav Sci*, vol. 59,

pp. 270–276, Oct. 2012, doi: https://doi.org/10.1016/j.sbspro.2012.09.275.
[17] S. S. Pande, S. R. Pande, V. R. Parate, A. P. Nikam, and S. H. Agrekar, "Correlation between difficulty & discrimination indices of MCQs in formative exam in Physiology," *South-East Asian Journal of Medical Education*, vol. 7, no. 1, p. 45, Jun. 2013, doi: https://doi.org/10.4038/seajme.v7i1.149.
[18] J. Johari *et al.*, "Difficulty Index of Examinations and Their Relation to the Achievement of Programme Outcomes," *Procedia Soc Behav Sci*, vol. 18, pp. 71–80, 2011, doi: https://doi.org/10.1016/j.sbspro.2011.05.011.
[19] S. M. J. Arokia Marie and S. Edannur, "Relevance of Item Analysis in Standardizing an Achievement Test in Teaching of Physical Science in B.Ed Syllabus," *i-manager's Journal of Educational Technology*, vol. 12, no. 3, pp. 30–36, Dec. 2015, doi: https://doi.org/10.26634/jet.12.3.3743.
[20] F. Taib and M. S. B. Yusoff, "Difficulty index, discrimination index, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance," *J Taibah Univ Med Sci*, vol. 9, no. 2, pp. 110–114, Jun. 2014, doi: https://doi.org/10.1016/j.jtumed.2013.12.002.
[21] Z. Arifin, "Kriteria Instrumen dalam suatu Penelitian," *Jurnal THEOREMS (The Original Research of Mathematics)*, vol. 2, no. 1, pp. 28–36, 2017.
[22] R. L. Brennan, *Educational Measurement*, 4th ed. Westport: Greenwood Publishing Group, 2006.
[23] Ebel and Frisbie, *Essential of Education Measurument*, 5th ed. New Jersey: Prentice-Hall, 2009.
[24] P. W. Miller, *Measurement and Teaching*, 1st ed. USA: Patrick W. Miller and Associates, 2008.
[25] A. Muhson, *Panduan Penggunaan AnBuso (Analisis Butir Soal) Versi 8.0*. Yogyakarta: Universitas Negeri Yogyakarta, 2017. Accessed: Feb. 03, 2024. [Online]. Available: https://staffnew.uny.ac.id/upload/132232818/pendidikan/Panduan%20Penggunaan%20AnBuso.pdf
[26] H. S. Setyaedhi, A. Ardianik, and M. Hanif, "Comparison of the Reliability Test of Semester Final Exam Scores for Graphic Media Courses Using Various Reliability Test Methods," in *Proceedings of the International Joint Conference on Arts and Humanities 2023 (IJCAH 2023), Advances in Social Science, Education and Humanities Research 785*, Atlantis Press, Dec. 2023, pp. 1206–1227. doi: https://doi.org/10.2991/978-2-38476-152-4_123.
[27] G. Usman and M. R. Yunus, "Pengembangan Instrumen Penilaian Autentik Pada Pembelajaran Fisika di SMA Negeri 4 Halmahera Utara," *Jurnal Hibualamo: Seri Ilmu-ilmu Sosial dan Kependidikan*, vol. 3, no. 2, pp. 23–29, Dec. 2019, Accessed: Jan. 21, 2024. [Online]. Available: https://journal.unhena.ac.id/index.php/sosialkependidikan/article/view/47
[28] H. Maulana, "Analisis Kualitas Instrumen Evaluasi Pembelajaran Menggunakan Media Digitalisasi Untuk Memotivasi Hasil Belajar Peserta Didik," *Bersatu: Jurnal Pendidikan Bhinneka Tunggal Ika*, vol. 1, no. 4, pp. 9–20, Jul. 2023, doi: https://doi.org/10.51903/bersatu.v1i4.255.
[29] Y. F. Ambarwati and I. Ismiyati, "Analisis Butir Soal Pilihan Ganda Ulangan Akhir Semester Genap Mata Pelajaran Kearsipan," *Measurement In Educational Research (Meter)*, vol. 1, no. 2, pp. 64–75, Mar. 2022, doi: https://doi.org/10.33292/meter.v1i2.144.
[30] K. Eldin. M. A. Salih *et al.*, "Psychometric Analysis of Multiple-Choice Questions in an Innovative Curriculum in Kingdom of Saudi Arabia," *J Family Med Prim Care*, vol. 9, no. 7, p. 3663, 2020, doi: https://doi.org/10.4103/jfmpc.jfmpc_358_20.
[31] D. N. Purnama and F. Alfarisa, "Karakteristik Butir Soal Try Out Teori Kejuruan Akuntansi SMK Berdasarkan Teori Tes Klasik dan Teori Respons Butir," *Jurnal Pendidikan Akuntansi Indonesia*, vol. 18, no. 1, pp. 36–46, Jun. 2020, doi: https://doi.org/10.21831/jpai.v18i1.31457.

[32] M. Fauzie, A. U. T. Pada, and S. Supriatno, "Analysis of The Difficulty Index of Item Bank According to Cognitive Aspects During The Covid-19 Pandemic," *Jurnal Penelitian dan Evaluasi Pendidikan*, vol. 25, no. 2, Dec. 2021, doi: https://doi.org/10.21831/pep.v25i2.42603.
[33] A. Sudijono, *Pengantar Evaluasi Pendidikan*, 15th ed. Jakarta: PT. Rajagrafindo, 2017.
[34] A. Muluki, P. Bundu, and I. Sukmawati, "Analisis Kualitas Butir Tes Semester Ganjil Mata Pelajaran IPA Kelas IV MI Radhiatul Adawiyah," *Jurnal Ilmiah Sekolah Dasar*, vol. 4, no. 1, p. 86, Apr. 2020, doi: https://doi.org/10.23887/jisd.v4i1.23335.
[35] H. Retnawati, *Validitas Reliabilitas dan Karakteristik Butir*, 2nd ed. Yogyakarta: Parama, 2017.
[36] Z. Arifin, *Evaluasi Pembelajaran*, 10th ed. Bandung: PT. Remaja Rosdakarya, 2017.