Library Progress International *Print version ISSN 0970 1052*
Vol.44 No.3, Jul-Dec 2024: P.8737-8744 *Online version ISSN 2320 317X*

Original Article *Available online at www.bpasjournals.com*

Item Quality Analysis of IPA Test Instruments Using R Studio

Annisa Fitriani[1], Sa'adatul Ulwiyah[2], Widihastuti[3] and Harun[4]

[1,2]Student, Educational Research and Evaluation, Yogyakarta State University, Yogyakarta, Indonesia
[3]Doctor, Educational Research and Evaluation, Yogyakarta State University, Yogyakarta, Indonesia
[4]Professor, Early Childhood Education,Yogyakarta State University, Yogyakarta, Indonesia
[1]annisafitriani.2022@student.uny.ac.id,[2]saadatul.2022@student.uny.ac.id, [3]widihastuti@uny.ac.id and [4]harun@uny.ac.id
Orchid Id number: 10009-0006-9646-8776
Corresponding Author*: Annisa Fitriani.

How to cite this article: Annisa Fitriani, Sa'adatul Ulwiyah, Widihastuti , Harun (2024). Item Quality Analysis of Ipa Test Instruments Using R Studio. *Library Progress International,* 44(3), 132-140.

ABSTRACT
A test is a statement, task or set of tasks planned to obtain information about educational and psychological traits or attributes. The purpose of this research is to see the quality of the science test instrument items that have been tested on 8th grade students of SMPN 8 Yogyakarta and describe the results of item analysis related to the Differentiation Index (IDB), Difficulty Level Index (ITK), and distractor effectiveness. The type of research used in this study is descriptive quantitative with the help of the R program. The test instrument analyzed was in the form of 30 multiple choice questions with 4 answer choices (A, B, C and D). The subjects used were 32 students of class VIII SMPN 8 Yogyakarta. The results showed that the Differentiation Index (IDB) was 5 items in the "Good" category, and there were 25 items in the "Not Good" category. The Index of Level of Difficulty (ITK) there are 8 items in the "Medium" category, and there are 22 items in the "Easy" category. The effectiveness of distractors as many as 3 items have "Good" distractors, as many as 2 items have "Good Enough" distractors and as many as 25 items have "Not Good" distractors. Overall, it can be seen that the differential power is more dominant in the "Not Good" category as many as 25 items. For the results of the level of difficulty is at 0.709 to 1.000 with the category "Easy" as many as 25 items. For the results of the effectiveness of distractors, 25 items are dominant in the "Not Good" criteria.

Keywords: Differentiation Index (IDB), Difficulty Level Index (ITK), Distractor Effectiveness, Descriptive Quantitative Research, IPA Test Instrument Analysis.

1) INTRODUCTION

Assessment of student learning outcomes is a reflection of the quality of education. This evaluation process is carried out by measuring the extent to which learning objectives in various disciplines are achieved in accordance with the established curriculum. Basically, this evaluation uses measurement tools to assess the achievement of learning objectives [1]. For educational decisions to have useful value, the information used must be accurate, reliable and appropriate to the problem at hand [2]. When the risk of bias is low, the research is considered to be of good quality [3]. Therefore, measurement data is important information for education organisers to make decisions related to learners..

The importance of evaluation in education is reflected through measurement results that produce grades or scores, reflecting students' abilities in a subject. In educational measurement activities, there is a process of quantifying phenomena or objects, such as motivation, achievement, and self-view, which are then expressed in the form of numbers. The use of measurement tools provides information about a person's position in the measured attribute. Therefore, it is important to choose a measurement instrument that has a high level of validity and reliability to ensure accurate measurement results.

One test instrument that is often used to assess student learning outcomes is a question bank. It is important to ensure that the question bank used is of good quality so that it can accurately measure students' abilities. An effective instrument is one that is able to provide accurate data and precise information, so that the measurement results can reflect students' abilities accurately [4].

An evaluation method involves a statement, task, or set of tasks planned to elicit information related to educational and psychological traits or attributes. Each element of the question or task in this evaluation method has an answer or condition that is considered correct. Grouping of evaluation methods can be done based on the form, type, and variation applied [5]. Measurement involves assigning numbers to attributes or characteristics possessed by certain people, things or objects according to clearly defined rules or formulations.

The characteristics of measurement lie in the use of certain numbers or scales, as well as the application of special rules or formulas [6]. The decision-making process involves using information obtained through the measurement of learning outcomes, using both test and non-test tools. In other words, giving value to the quality of something is the essence of evaluation. The link between test, measurement and evaluation is that evaluation of learning outcomes can only be done properly and correctly when using information obtained through measurement of learning outcomes using test as a measuring tool.

The various roles of testing, measurement, and assessment in education involve aspects such as selection, placement, diagnosis, remedial, feedback, motivation and guidance, curriculum improvement, educational programmes, and scientific advancement. Test planning is very important because the meaning of a test can only be achieved if it consists of items that test crucial objectives and reflect various domains of knowledge, skills, and abilities equally. There are six things to consider in test planning, including sample and item selection, type of test to be used, aspects to be tested, item format, number of items, and item difficulty distribution [7].

The quality of an item does not depend solely on its form or type, but is rather determined by how well or poorly constructed each item is. Whether objective or descriptive, both can be effective instruments for measuring learning success, depending on the level of quality of their respective item construction [8]. In fact, in some circumstances, description items carry a higher level of risk than objective items. The difference lies in the fact that the quality of description items depends not only on the student's ability to answer the question, but is more influenced by the skill and objectivity of the question maker in assessing the exam results. Meanwhile, objective question items can be analysed with more accuracy and responsibility, allowing for more precise identification of weaknesses.

Objective questions can be used repeatedly, provided they are not in the same test set. Therefore, there is an advantage or usefulness of item analysis, which is then revised so that poorly constructed items can be corrected [9]. Finally, the items that have been tested will be obtained, and they will provide an accurate measurement of the learning outcomes to be measured.

There are several reasons why item analysis is needed. [10] say these reasons include: a. To find out the strengths and Item analysis serves several key purposes in the context of test development and evaluation. Firstly, through the identification of weaknesses in test items, the selection and revision process can be conducted more effectively. Secondly, the provision of comprehensive item specifications provides valuable guidance to question makers in constructing sets of questions that meet the needs of examinations in different areas and levels. Thirdly, quickly identifying problems in items, such as ambiguities, misplaced answer keys, questions that are too difficult or too easy, or questions with low differentiation, allows question makers to make timely decisions regarding the removal or revision of problematic items as well as determining student scores. Fourth, item analysis is used as a tool to assess the difficulty or ease of questions. Fifth, it is used to evaluate items to be stored in the question bank. Sixth, by obtaining information about items, it enables the construction of multiple sets of parallel questions, which is very beneficial for retesting or assessing the ability of different groups of examinees at different times.

Analysis of instrument quality can be done through three main factors, namely the index of Level of Difficulty (ITK), the Index of Distinction (IDB), and the effectiveness of distractors [11].The Index of Level of Difficulty (ITK) helps in grouping questions into medium, easy, or difficult categories. The Index of Distinction (IDB) or discrimination is used to distinguish between participants who have high abilities and low, while the effectiveness of distractors is used to assess the extent to which the available answer choices function properly [12].

One application that can help in analyzing test instrument items is R Studio. This application can automatically analyze the Index of Level of Difficulty (ITK), Index of Distinction (IDB), and the effectiveness of distractors on each test item [13]. The results of the analysis can help identify the weaknesses and strengths of each item, making it easier to make the necessary improvements. In its use, this program has many advantages, some of which are effective data handling and deviation facilities, graphical facilities for analysis and data display, and the programming language is well developed and simple. However, most programs written in R are basically temporary, and are written for one data analysis [14].

In light of the preceding explanation, the aim of this research is to evaluate the effectiveness of mathematics test instruments and provide a comprehensive overview of the outcomes of item analysis concerning reliability, Differentiation Index (IDB), Level of Difficulty Index (ITK), and the effectiveness of distractors. The subject of analysis is a science test instrument previously administered to eighth-grade students at SMPN 8 Yogyakarta. The evaluation of the test instrument's quality is conducted as a strategic initiative to enhance the overall quality of evaluation tools utilized for assessing students' proficiency in science.

2) METHODS

This study applied a quantitative descriptive method with the main objective of assessing the quality of the instrument and explaining the results of item analysis of the test used, especially related to the Differentiation Index (IDB), the Index of Difficulty Level (ITK), and the effectiveness of distractors. The research participants were 32 grade VIII students from SMPN 8 Yogyakarta. The data collection technique used was the documentation technique, in which the data collected were students' responses to the science test instrument. The mathematics test instrument consisted of 30 items in the form of multiple choice without labelling A, B, C, or D at the end of each answer.

A classical test theory framework was employed to analyze the characteristics of items, encompassing the Differentiation Index (IDB), Index of Difficulty Level (ITK), and the effectiveness of distractors. R Studio served as the analytical tool for scrutinizing the instrument items. Student responses were processed using the application to manage the acquired data. The results of the analysis were translated into numerical indices and subsequently compared with predefined standards derived from the classical approach in test theory. These analysis outcomes were then interpreted to elucidate the distinctive features of each question within the instrument, including the Differentiation Index (IDB), the Index of Difficulty Level (ITK), and the efficiency of distractors.

A. Index of Distinction (IDB)

The discriminating power (discrimination) of a test item is the ability of an item to distinguish between high-ability and low-ability test takers [15], providing the criteria for the Discriminating Power Index (IDB) as follows:

Table 1: Index of Distinction (IDB)

Distinguishing Power Index	Criteria
>0.40	**Very Good**
0.30-0.39	**Good, little or no revision required**
0.20-0.29	**Fairly good, items require revision**
<0.19	**Not good, item should be discarded**

The discrimination of a test item is the ability of the item to distinguish between high and low ability test takers, providing a criterion for assessing the effectiveness of the item in measuring differences in participant ability.

B. Index of Difficulty (ITK)

The difficulty index is the proportion of test takers who answer correctly [16]. One of the most prevalent and uncomplicated indicators of question difficulty often utilized to evaluate the difficulty index of a question is the ratio of correct answers, also known as the proportion correct [17]. The following are the criteria for the Index of Difficulty (ITK).

Table 2: Index of Level of Difficulty (ITK)

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

The difficulty index is the proportion of test takers who answer correctly. The most common and simple indicator of question difficulty used to assess the difficulty index is the correct answer ratio, also known as the proportion of correct answers.

C. Distractor Effectiveness

Effectiveness of distractors assesses how well the incorrect options succeed in misleading individuals taking the test without knowledge of the correct answer. The quality of a distractor in concealing the correct choice improves as more test takers opt for it [18]. [19] states that distractors are said to be good if at least 5% of the test participants choose them. Meanwhile, [20] states that distractors are said to be effective if selected by at least 2% of the respondents.

Table 3: Distractor Effectiveness

Distractor	Criteria
≥ 2%	**Effective**
< 2%	**Not Effective**

Exposure effectiveness assesses how well the wrong answer option manages to mislead test takers who do not know the correct answer. the quality of an exception increases when more test takers select it. an exception is said to be good if it is selected by at least 5% of test takers and effective if it is selected by at least 2% of respondents.

3) RESULTS

The study initiated by collecting the responses given by students to a science test instrument. The test consisted of 30 items in a multiple-choice format, offering four answer options labeled A, B, C, and D. This test instrument was distributed to 32 eighth-grade students at SMPN 8 Yogyakarta. Subsequently, the data pertaining to students' responses to the test instrument underwent analysis using the classical approach, with the assistance of R Studio.

Index of Distinction (IDB)

The outcomes of the examination of the discriminatory capability of questions utilizing R Studio employed the classical test theory (CTT) methodology, specifically through the analysis of biserial point scores. The findings from the analysis of differentiation parameters using R Studio are presented in Table 4.

Table 4: Distinguishing Power Parameters with Rstudio

Item No.	Distinguishing Power		Item No.	Distinguishing Power	
	pbis	Description		*pbis*	Description
1	NA	Not good	16	0.521	Good
2	NA	Not good	17	0.537	Good
3	NA	Not good	18	0.224	Not good
4	-0.372	Not good	19	-0.163	Not good
5	-0.199	Not good	20	0.122	Not good
6	-0.044	Not good	21	-0.117	Not good
7	0.166	Not good	22	-0.090	Not good
8	0.042	Not good	23	0.408	Good
9	0.065	Not good	24	-0.252	Not good
10	-0.032	Not good	25	0.028	Not good
11	0.204	Not good	26	0.388	Good
12	-0.061	Not good	27	-0.034	Not good
13	0.028	Not good	28	0.156	Not good
14	0.083	Not good	29	0.368	Good
15	-0.090	Not good	30	NA	Not good

The R Studio software was utilized to compute the differential power for each item in this instrument, resulting in a summary of the differential power values presented in Table 4. The analysis of 30 items using R Studio software yielded the differentiation parameter values for each item. According to Table 4, five items fall into the "Good" category, specifically item numbers (16, 17, 23, 26, and 29). Conversely, there are 25 items categorized as "Not Good," identified by item numbers (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 18, 19, 29, 21, 22, 24, 25, 27, 28, and 30).

Index of Difficulty (ITK)

RStudio's application facilitates a more in-depth exploration of information associated with the difficulty level, including the proportion of participants capable of answering correctly and the number of participants encountering challenges in responding to a question. Within this software, the difficulty level is observable through rspP. This examination offers valuable perspectives for comprehending the efficacy and precision of the questions posed in the test. The outcomes of the analysis regarding the difficulty level of questions using RStudio are presented in Table 5.

Table 5. Difficulty Level Parameters with Rstudio

Item No.	Difficulty Level		Item No.	Difficulty Level	
	Coefficient	Description		Coefficient	Description
1	1.000	Easy	16	0.838	Easy
2	1.000	Easy	17	0.709	Easy
3	1.000	Easy	18	0.645	Medium
4	0.870	Easy	19	0.516	Medium
5	0.903	Easy	20	0.709	Easy

6	0.967	Easy	21	0.967	Easy
7	0.838	Easy	22	0.774	Easy
8	0.806	Easy	23	0.580	Medium
9	0.774	Easy	24	0.580	Medium
10	0.645	Medium	25	0.967	Easy
11	0.709	Easy	26	0.516	Medium
12	0.935	Easy	27	0.709	Easy
13	0.709	Easy	28	0.870	Easy
14	0.354	Medium	29	0.483	Medium
15	0.709	Easy	30	1.000	Easy

The RStudio software was employed for the computation of difficulty level parameters. Out of the 30 items scrutinized using RStudio, the resulting parameter values for the difficulty level of each item indicate that eight items fall into the "Medium" category, identified by item numbers (10, 14, 18, 19, 23, 24, 26, and 29). These particular items exhibit coefficients ranging from 0.354 to 0.645. Additionally, there are 22 items categorized as "Easy," corresponding to item numbers (1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 15, 16, 17, 20, 21, 22, 24, 27, 28, and 30). These items display coefficients ranging from 0.709 to 1.000.

Distractor Effectiveness

The analysis of distractor effectiveness in RStudio closely mirrors the approach used in Iteman, entailing the scrutiny of available information for each item and consideration of pertinent metrics. The term rspP, denoting "relative selected proportion P," signifies the proportion of learners opting for a specific answer choice, encompassing the correct option. A higher rspP value suggests an elevated preference for that answer choice among learners. The outcomes of the analysis on the effectiveness of distractors using RStudio are presented in Table 6.

Table 6. Results of Answer Distribution with Rstudio

Question No.	**Distribution of Student Answers**				**Question No.**	**Distribution of Student Answers**			
	A	**B**	**C**	**D**		**A**	**B**	**C**	**D**
1	**NA**	**NA**	**NA***	**NA**	**16**	**0.000**	**0.838***	**0.000**	**0.161**
2	**NA**	**NA***	**NA**	**NA**	**17**	**0.290**	**0.290***	**0.000**	**0.000**
3	**NA**	**NA**	**NA***	**NA**	**18**	**0.129**	**0.000**	**0.645***	**0.225**
4	**0.967**	**0.870***	**0.000**	**0.322**	**19**	**0419**	**0.516***	**0.032**	**0.032**
5	**0.096**	**0.903***	**0.000**	**0.000**	**20**	**0.129**	**0.709***	**0.161**	**0.000**
6	**0.032**	**0.000**	**0.000**	**0.967***	**21**	**0.032**	**0.000**	**0.967***	**0.000**
7	**0.096**	**0.000**	**0.838***	**0.645**	**22**	**0.225**	**0.000**	**0.774***	**0.000**
8	**0.806***	**0.193**	**0.000**	**0.000**	**23**	**0.064**	**0.129**	**0.225**	**0.580***
9	**0.225**	**0.000**	**0.000**	**0.774***	**24**	**0.960**	**0.580***	**0.161**	**0.161**
10	**0.645***	**0.032**	**0.258**	**0.064**	**25**	**0.000**	**0.000**	**0.967***	**0.032**
11	**0.709***	**0.129**	**0.000**	**0.161**	**26**	**0.161**	**0.129**	**0.516***	**0.193**
12	**0.322**	**0.935***	**0.000**	**0.322**	**27**	**0.000**	**0.000**	**0.709***	**0.258**
13	**0.322**	**0.967***	**0.000**	**0.000**	**28**	**0.129**	**0.000**	**0.000**	**0.870***
14	**0.654**	**0.000**	**0.354***	**0.000**	**29**	**0.194**	**0.064**	**0.483***	**0.258**
15	**0.709***	**0.000**	**0.129**	**0.161**	**30**	**NA***	**NA**	**NA**	**NA**

Utilizing RStudio software for the analysis of distractor effectiveness in each item of this instrument yields a summary indicating that a majority of the questions exhibit a quality classified as "Not Good," totaling 25 items, specifically item numbers (1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 19, 20, 21, 22, 24, 25, 27, 28, and 30). For questions categorized as "Good Enough," there are 2 items, specifically question numbers (16 and 17). Furthermore, for questions categorized as "Good," there are 3 items, namely question numbers (23, 26, and 29). This data highlights that the distractors or incorrect answer choices do not effectively serve their purpose, leading to the classification of many questions as "Not Good" based on the applied assessment criteria.

4) DISCUSSION

The examination of the Index of Difficulty (ITK) using RStudio revealed variations in the difficulty levels of the assessed questions. It was observed that 8 questions fell within the "Medium" category, indicating a well-balanced level of difficulty—neither excessively challenging nor overly straightforward. This balance is crucial as it provides learners with diverse abilities the opportunity to assess their comprehension of the tested material. Additionally, there were 22 questions categorized as "Easy," implying that these questions are perceived as more manageable for learners. In this context, these easily answered questions can serve as a confidence boost for learners when confronting the science exam.

Derived from the outcomes of the ITK analysis, it can be inferred that the science test exhibits diversity in the difficulty levels of questions, encompassing both easy and medium categories. The presence of questions with varying difficulty levels contributes to a holistic assessment of students' comprehension and proficiency in the science material under examination. In the process of test development, it becomes crucial to reassess the difficulty levels of questions to achieve a more balanced distribution. The objective is to ascertain that the challenge aligns with the learners' capabilities, fostering enhanced understanding of the science material.

The analysis of the Differentiation Index (IDB) using RStudio reveals diverse assessments of item quality. Five items were classified as "Good," signifying their efficacy in effectively distinguishing between learners with varying levels of ability. Nonetheless, there is potential for enhancing the quality of these items. Conversely, 25 items were categorized as "Not Good," indicating limitations in their ability to differentiate learners' abilities. Items falling into this category are deemed ineffective and merit consideration for potential removal from the assessment.

Conclusively, the outcomes of the Differentiation Index (IDB) analysis indicate the necessity for revising several items rated as "good" to attain elevated quality standards. Simultaneously, items receiving unfavorable ratings should undergo corrections or be excluded from the test instrument. The aim is to enhance the accuracy of results and foster a more effective differentiation of students' abilities through improvements or removal of these items in the science test instrument.

The analysis outcomes pertaining to the efficacy of distractors revealed that among the 30 items assessed, 25 items had answer choices categorized as "Not Good." Two items featured answer options in the "Good Enough" classification, while three items had answer options falling into the "Good" category. This data underscores that the distractors or incorrect answer choices do not perform effectively, leading to the classification of a significant number of items as "Not Good" based on the applied assessment criteria.

The inefficacy of distractors or answer options in items may adversely affect the validity and reliability of science test instruments. Inadequate answer options can diminish the instrument's capacity to precisely gauge learners' comprehension and skills. Hence, there is a need to modify items featuring subpar answer options. It is crucial to give careful consideration and ensure that each answer choice effectively challenges learners lacking understanding of the material, while the correct answer remains clear and easily discernible for those who comprehend the material. Through the enhancement of distractors or answer options, science test instruments can evolve into more dependable tools for evaluating learners' comprehension of the tested science material.

The findings from this analysis underscore the necessity to revise and enhance items exhibiting suboptimal quality. Enhancements should be guided by careful consideration of criteria within the difficulty index, the differentiation index, and the effectiveness of distractors. Through suitable improvements, test instruments have the potential to achieve heightened validity and reliability in assessing students' abilities and comprehension of the material under examination.

5) CONCLUSION

The research and analysis conducted using rstudio on the science test instrument for 8th-grade students at smp negeri 8 yogyakarta yielded several significant findings. Firstly, in the examination of differential power, there were 5 items categorized as "good" and 25 items in the "not good" classification. Items meeting the criteria for "good" require revision, while those falling under "not good" cannot be utilized and should be discarded. Secondly, concerning the analysis of difficulty levels, there were 8 items classified as "medium" and 22 items in the "easy" category. Thirdly, 25 items in the science test instrument had alternative answers categorized as "not good," with 2 items in the "good enough" category and 3 items in the "good" category. Based on these findings, it can be concluded that the science test instrument items for 8th-grade students at smp negeri 8 yogyakarta necessitate revision and improvement to enhance their quality.

6) ACKNOWLEDGMENTS

We would like to thank to Sa’adatul Ulwiyah, Dr. Widihastuti, and Prof. Harun for their contributions to this work. Special thanks to BPPT and Lembaga Pengelola Dana Pendidikan (LPDP) for their assistance and financial support our research and publication.

7) RECOMMENDATION

From the results of the research and analysis using RStudio on the science test instrument of class VIII students of SMP Negeri 8 Yogyakarta, there are several recommendations that can be considered. First, it is important to make in-depth revisions to the items that are rated "Good" in order to improve their quality, while the items categorized as "Not Good" need to be thoroughly repaired or even deleted if they cannot be repaired. Second, in setting the level of difficulty of the questions, it is recommended to adjust the distribution of difficulty between 8 items in the "Medium" category and 22 items in the "Easy" category to achieve a better balance. Third, the items that have "Not Good" answer alternatives as many as 25 items, with only 2 items in the "Good Enough" category and 3 items in the "Good" category, indicate the need for in-depth improvement of these answer choices to improve the validity and quality of the science test instrument as a whole. By making careful revisions and comprehensive improvements according to these findings, it is hoped that the science test instrument for class VIII students of SMP Negeri 8 Yogyakarta can provide a more accurate measure of students' understanding of the science material being tested.

After improvements are made, it is important to conduct revalidation by teachers or experts in the science subject area to ensure the expected quality improvement. In addition, before widespread use, a pilot test of the improved instrument on a small sample of students needs to be conducted to evaluate the effectiveness of the implemented improvements. Finally, periodic maintenance of the test instrument is highly recommended to maintain its relevance and quality, including monitoring student responses and periodic item updates. By taking these steps, it is expected that the science test instrument for grade VIII students of SMP Negeri 8 Yogyakarta can significantly improve its quality.

8) REFERENCES

[1] Idrus, "Evaluasi Dalam Proses Pembelajaran Idrus L 1," *Eval. Dalam Proses Pembelajaran*, no. 2, pp. 920–935, 2019.

[2] S. A. Nurfatimah, S. Hasna, and D. Rostika, "Membangun Kualitas Pendidikan di Indonesia dalam Mewujudkan Program Sustainable Development Goals (SDGs)," *J. Basicedu*, vol. 6, no. 4, pp. 6145–6154, 2022, doi: 10.31004/basicedu.v6i4.3183.

[3] F. N. A. Kurniawati, "Meninjau Permasalahan Rendahnya Kualitas Pendidikan Di Indonesia Dan Solusi," *Acad. Educ. J.*, vol. 13, no. 1, pp. 1–13, 2022, doi: 10.47200/aoej.v13i1.765.

[4] S. Ermawati and T. Hidayat, "Penilaian Autentik Dan Relevansinya Dengan Kualitas Hasil Pembelajaran," *J. Pendidik. Ilmu Sos.*, vol. 27, no. 1, pp. 1412–3835, 2017.

[5] S. Munadi, "Pneilaian Hasil Belajar," *Angew. Chemie Int. Ed. 6(11), 951–952.*, 2018.

[6] J. Indrastoeti, "Pengembangan asesmen pembelajaran sekolah dasar," p. 23, 2012, [Online]. Available: https://digilib.uns.ac.id/dokumen/detail/29050/Pengembangan-asesmen-pembelajaran-sekolah-dasar

[7] Vinta A. Tiarani, "Teknik Pengembangan Soal Objektif," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2013.

[8] 2019 Goleman et al., "Analisis Butir Soal," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2019, doi: 10.13140/RG.2.2.26498.71360.

[9] M. Fitrianawati, "Peran Analisis Butir Soal Guna Meningkatkan Kualitas Butir Soal, Kompetensi Guru Dan Hasil Belajar Peserta Didik," *Pros. Semin. Nas. Pendidik. PGSD UMS HDPGSDI Wil. Jawa*, vol. 5, no. 3, pp. 282–295, 2015.

[10] T. P. Siregar, E. Surya, and E. Syahputra, "Quality Analysis of Multiple Choice Test and Classical Test at X Grade Students of Senior High School," *Int. J. Adv. Res. Innov. Ideas Educ.*, vol. 3, no. 2, pp. 2153–2159, 2017.

[11] H. Maulidah, S. Sukarno, and B. Syefrinando, "Analisis Kualitas Instrumen Tes Fisika Kelas X Menggunakan Software Anates," *Phys. Sci. Educ. J.*, vol. 2, no. April 2021, pp. 153–162, 2023, doi: 10.30631/psej.v2i3.1660.

[12] R. S. Diyah Kiki Widiyaningrum, Nurul Syamsiah, "Analisis Kualitas Butir Soal Multiple Choice Pada Tes Akademik Matematika," vol. 2507, no. February, pp. 1–9, 2020.

[13] F. A. Setiawati, R. E. Izzaty, and V. Hidayat, "Evaluasi Karakteristik Psikometrik Tes Bakat Differensial dengan Teori Klasik," *Humanitas (Monterey. N. L).*, vol. 15, no. 1, p. 46, 2018, doi: 10.26555/humanitas.v15i1.7249.

[14] D. U. Wustqa, E. Listyani, R. Subekti, R. Kusumawati, M. Susanti, and K. Kismiantini, "Analisis Data Multivariat Dengan Program R," *J. Pengabdi. Masy. MIPA dan Pendidik. MIPA*, vol. 2, no. 2, pp. 83–86, 2018, doi: 10.21831/jpmmp.v2i2.21913.

[15] M. Solichin, "Analisis Daya Beda Soal, Taraf Kesukaran, Validitas Butir Tes, Interpretasi Hasil Tes dan Validitas Ramalan dalam Evaluasi Pendidikan," *Dirasat J. Manaj. Pendidik. Islam*, vol. 2, no. 2, pp. 192–213, 2017, [Online]. Available: www.depdiknas.go.id/evaluasi-proses-

[16] R. Damayanti, Wi. D., Halidjah, S., & Pranata, "Analisis tingkat kesukaran butir soal pilihan gandapada penilaian tengah semester kelas iv," *J. Pendidik. dan Pengajaran Khatulistiwa*, vol. 10, no. 11, pp. 1–10,

2021, [Online]. Available: https://jurnal.untan.ac.id/index.php/jpdpb/article/view/50458/75676591120

[17] A. Iskandar and M. Rizal, “Analisis kualitas soal di perguruan tinggi berbasis aplikasi TAP,” *J. Penelit. dan Eval. Pendidik.*, vol. 22, no. 1, pp. 12–23, 2018, doi: 10.21831/pep.v22i1.15609.

[18] V. O. Bano, D. N. Marambaawang, and Y. Njoeroemana, “Analisis Kriteria Butir Soal Ujian Sekolah Mata Pelajaran IPA di SMP Negeri 1 Waingapu,” *Ideas J. Pendidikan, Sos. dan Budaya*, vol. 8, no. 1, p. 145, 2022, doi: 10.32884/ideas.v8i1.660.

[19] M. Yani, “Efektivitas Distractor Pada Tes Pilihan Ganda Untuk Mendeteksi Kesalahan Siswa Dalam Menyelesaikan Soal Matematika,” *Al Khawarizmi J. Pendidik. dan Pembelajaran Mat.*, vol. 2, no. 2, p. 125, 2019, doi: 10.22373/jppm.v2i2.4502.

[20] W. Budiaji, “The Measurement Scale and The Number of Responses in Likert Scale,” *J. Ilmu Pertan. dan Perikan. Desember*, vol. 2, no. 2, pp. 127–133, 2013, doi: 10.31227/osf.io/k7bgy.