Analysis of Multiple-Choice Assessment Instruments for Grade X Students in Terms of Validity, Reliability, Difficulty Level, Distinction Power, and Distractor Effectiveness

Tesa Manisa[1]*, Anggie Syalom[2], Resty Nofini Nafesda[3]

[1]Biology Education Study Program, Faculty of Mathematics, Natural Scinces, and Technology, Universitas PGRI Pontianak, Jl. Ampera No.88 Pontianak, Kalimantan Barat.
*Corresponding Author: tesamanisa68@gmail.com

Abstract. Item analysis is one of the important steps to ensure the quality of learning evaluation instruments in measuring student learning outcomes. This study aims to analyze the quality of assessment instruments in the form of multiple-choice questions on the material of scarcity for grade X students in terms of validity, reliability, level of difficulty, discriminating power, and distractor effectiveness. The study uses a quantitative approach with a descriptive method. The instrument analyzed consists of 30 multiple-choice questions that have been tested on grade X students. Data analysis was carried out through validity tests, reliability using Cronbach's Alpha, level of difficulty, discriminating power, and distractor effectiveness. The results of the study showed that of the 30 items analyzed, 25 items (83.33%) were declared valid and 5 items (16.67%) were declared invalid. The reliability test results obtained a Cronbach's Alpha value of 0.972 which is included in the very high category. Analysis of the level of difficulty showed that 24 items (96%) were in the medium category and 1 item (4%) was in the easy category. The discriminating power analysis showed that 13 items (52%) were in the very good category and 12 items (48%) were in the good category. Meanwhile, the distractor effectiveness analysis showed that 11 items (36.67%) had all functioning distractors, while 19 items (63.33%) still had one or more non-functioning distractors. In general, the developed instrument has met the criteria for good instrument quality and is suitable for use as a learning evaluation tool, although some items and distractors still require revision to improve the quality of measurement.
Keywords: Validity, Reliability, Difficulty Level, Discriminating Power, Distraction Analysis.

How to cite (in APA style):
Manisa, T., Syalom, A., Nafesda, R.N. (2026). Analysis of Multiple-Choice Assessment Instruments for Grade X Students in Terms of Validity, Reliability, Difficulty Level, Distinction Power, and Distractor Effectiveness. *Proceeding of IJC-MaSTEd*, 1(1), 266-275

DOI: https://doi.org/10.31571/ijcmasted.v1i1.1154

1. Introduction

Learning evaluation is an integral part of the educational process because it can provide information about student competency achievement (Saputri et al., 2023; Sudijono, 2011). One of the most commonly used forms of evaluation is a written test, which is usually structured in the form of questions and used to measure students' understanding of the learning material (Arikunto, 2010). In the context of Economics subjects, especially at the senior high school (SMA) level, good questions are essential to measure students' critical and analytical thinking skills regarding complex economic phenomena (Mulyasa, 2013). Unfortunately, many questions used in the learning process have not undergone adequate quality analysis, so that the assessment results can be inaccurate in interpreting student abilities (Brookhart & Nitko, 2019). The existence of tests to measure educators' abilities will help teachers evaluate learning and determine what areas need to be improved. Through this evaluation process, teachers can make decisions related to the learning process, both in terms of

process and results (Makhisoh & Suroyya, 2022). Evaluation can also have an impact on students, namely to be more diligent in studying consistently, encourage teachers to improve the quality of the learning process and encourage schools to improve facilities or the quality of teachers to support learning abilities (Fauzi & Mustika, 2022).

The quality of multiple-choice test instruments can be analyzed through several key aspects, including validity, reliability, difficulty level, discriminatory power, and distraction effectiveness (Crocker & Algina, 2020; Matondang, 2019). These five aspects are interrelated and collectively determine the quality of the test item as a measurement instrument (Popham, 2017). Validity refers to the extent to which a test instrument measures what it is supposed to measure (American Educational Research Association, 2014). In the context of multiple-choice items, validity includes content validity, which indicates the item's suitability to the competency it is intended to measure, and construct validity, which indicates that the item is capable of measuring the underlying theoretical construct (Sireci & Faulkner-Bond, 2014; Widoyoko, 2018). Validity testing can be conducted through correlation analysis between item scores and the total score or with relevant external criteria (Ebel & Frisbie, 2019).

Reliability refers to the consistency of measurement results (Thorndike & Thorndike-Christ, 2019). A test instrument is said to be reliable if it can produce consistent results when used repeatedly under the same conditions (Sudjana, 2016). Reliability can be estimated through various methods such as internal consistency using the Cronbach's Alpha formula, or the split-half method, which divides the item into two parts and calculates the correlation (Cronbach & Shavelson, 2021; Purwanto, 2019). The level of question difficulty indicates the proportion of students who answer an item correctly (Gronlund & Waugh, 2020). The difficulty index ranges from 0 to 1, where values close to 0 indicate very difficult questions and values close to 1 indicate very easy questions (Aiken & Groth-Marnat, 2015). A proportional distribution of difficulty levels in a test instrument is necessary to produce optimal score variability (Kartowagiran, 2018).

Discriminating power measures the ability of a test item to differentiate between high-ability and low-ability students (Osterlind, 2019). Questions with good discriminating power will be answered correctly by more students in the upper group than in the lower group (Haladyna & Rodriguez, 2021). A low or negative discriminating power index indicates that the item does not function effectively in differentiating student abilities (Arifin, 2017).

Distraction analysis aims to evaluate the effectiveness of incorrect answer choices (distractors) in engaging students who have not mastered the material (Haladyna, 2020). A good distractor should be attractive enough for students who have not yet mastered the concept, but should not mislead students who have already grasped the material well (Gierl et al., 2022). Distractors that are not selected by any students or are chosen more by the upper group indicate a need for revision (Raharjo, 2019).

Item analysis encompasses validity, reliability, difficulty level, and discriminating power. Item validity determines whether the item accurately measures students' understanding of the concept of scarcity and basic economic principles. Reliability, on the other hand, relates to the consistency of measurement results on the same material at different times (Anshari et al., 2024). Invalid or unreliable items will produce inaccurate assessment data, thus obscuring the true picture of students' abilities. The difficulty level is necessary to determine the extent to which students can correctly answer scarcity questions, whether the questions are too easy, moderate, or too difficult. Meanwhile,

discriminating power indicates the item's ability to differentiate between students with high and low levels of understanding on concepts such as opportunity cost and the allocation of limited resources (Hanifah et al., 2014). The study showed that the use of multiple-choice associative questions increased discriminating power and contextual relevance compared to standard items. Therefore, analyzing item quality in scarcity material is important not only for improving assessment quality but also for ensuring that students truly understand core economic concepts. Furthermore, carefully analyzed instruments can help teachers provide more precise feedback, design more responsive learning, and facilitate more accurate educational decision-making. Therefore, this study aims to conduct item analysis not only as a technical activity but also as a crucial part of efforts to improve the quality of the evaluation process in Economics learning. With valid and reliable instruments, teachers can obtain more accurate information to support informed educational decision-making.

Therefore, this study aims to analyze the items on the scarcity material, which is urgently needed due to the lack of evaluation instrument references that focus on this material at the high school level, especially those that have undergone comprehensive testing of validity, reliability, difficulty level, discriminating power, and distractor analysis. Furthermore, the scarcity material has characteristics that can challenge the preparation of questions, because it requires a balance between theoretical concepts and applicable contexts relevant to students' daily lives. With this research, teachers and curriculum developers can have a strong reference in compiling and improving economic questions, especially scarcity material, so that assessments become more accurate and meaningful, and support more responsive learning planning. Research on comprehensive item analysis has been widely conducted in recent years. Other researcher found that multiple-choice questions in the final assessment of mathematics subjects showed high reliability, the majority of questions were in the medium difficulty category, and the effective distractors were only around 45%, indicating the need for distractor revision to improve the diagnostic function of the questions. Other research also reported variations in difficulty and discrimination indices in school exam questions, with some items requiring improvement due to low discriminatory power. An international study by Rezigalla et al. (2024) on 60 MCQ items showed that efficiently functioning distractors were related to item difficulty and discriminatory power, providing empirical evidence regarding the relationship between item psychometric parameters. Furthermore, an evaluation of MCQ quality in a medical education setting showed that the majority of items had high discrimination and good functional distractors, as well as strong test reliability. These studies support the use of item analysis as an important tool to ensure learning evaluation questions meet the criteria of validity, reliability, difficulty, discriminatory power, and distractor effectiveness.

2. Method

This study used a quantitative approach with descriptive analysis methods to evaluate the quality of multiple-choice questions (Creswell & Creswell, 2018). The research subjects were 25 10th-grade students. The instrument analyzed was a multiple-choice question as many as 30 items that had been tested in the final semester exam.

The main data analysis technique in this study includes evaluating the validity and reliability of the test instrument (Wahyudi et al, 2025). Validity testing uses the Pearson product-moment correlation between item scores and total scores. (Azwar, 2019; Cohen et al., 2020). The item is declared valid if the calculated r value is greater than the table r value at a significance level of 5%. The reliability test uses the Cronbach Alpha method with reliability criteria: very high (≥0.90), high (0.70-0.89), moderate (0.40-0.69), and low (<0.40) (Tavakol & Dennick, 2011). The analysis of the level of difficulty is calculated using the formula: P = B/JS (1) Where: P = difficulty index, B = number of

students who answered correctly, JS = total number of students. Classification of the level of difficulty: easy (P > 0.70), moderate (0.30 ≤ P ≤ 0.70), difficult (P < 0.30) (Desjardins & Bulut, 2018). The discriminating power is calculated using the formula: D = BA/JA - BB/JB (2) Where: D = discriminating power, BA = number of upper group who answered correctly, JA = number of upper group, BB = number of lower group who answered correctly, JB = number of lower group. Classification of discriminating power: very good (D ≥ 0.40), good (0.30 ≤ D < 0.40), sufficient (0.20 ≤ D < 0.30), poor (D < 0.20) (Allen & Yen, 2019). The analysis of distractor effectiveness was carried out by calculating the percentage of each answer option chosen by the upper and lower groups (Rodriguez, 2022; Rizqa et al., 2025). A distractor is declared effective if it is chosen by at least 5% of all test participants and is chosen more by the lower group than the upper group (Suryabrata, 2018).

3. Results and Discussion

This study aims to analyze the quality of multiple-choice questions used as an evaluation instrument for 10th-grade students. The analysis of question quality is carried out based on several indicators, namely validity, reliability, level of difficulty, discriminating power, and distractor effectiveness. Analysis of these five aspects is necessary to determine the suitability of the instrument as a measuring tool for student learning outcomes. The results of the question quality analysis are presented and discussed in each aspect as follows.

Validity is one of the essential characteristics of a test instrument. Validity testing is conducted to determine the extent to which each item measures the competency it is intended to measure. The results of the item validity test are presented in Table 1.

Table 1. Item Validity Test Results

No Question	r Table	r Count	Results
1	0.514	0.990	Valid
2	0.514	0.853	Valid
3	0.514	0.751	Valid
4	0.514	0.610	Valid
5	0.514	0.990	Valid
6	0.514	0.703	Valid
7	0.514	0.387	Invalid
8	0.514	0.555	Valid
9	0.514	0.610	Valid
10	0.514	0.674	Valid
11	0.514	0.596	Valid
12	0.514	0.703	Valid
13	0.514	0.778	Valid
14	0.514	0.990	Valid
15	0.514	0.445	Invalid
16	0.514	0.306	Invalid
17	0.514	0.990	Valid
18	0.514	0.659	Valid
19	0.514	0.520	Valid

No Question	r Table	r Count	Results
20	0.514	0.129	Invalid
21	0.514	0.872	Valid
22	0.514	0.563	Valid
23	0.514	0.990	Valid
24	0.514	0.263	Invalid
25	0.514	0.990	Valid
26	0.514	0.990	Valid
27	0.514	0.624	Valid
28	0.514	0.652	Valid
29	0.514	0.853	Valid
30	0.514	0.838	Valid

Based on Table 1, the results of the validity test on 30 questions show that 25 questions meet the validity criteria, while 5 questions are declared invalid, namely numbers 7, 15, 16, 20, and 24. These five questions have a calculated r value lower than the r table value of 0.514 so they do not meet the validity criteria. On the other hand, the other 25 questions have a calculated r value greater than the r table so they are declared valid and suitable for use in research.

These results indicate that most of the test items were able to measure the expected competencies in accordance with the learning objectives. The high number of valid test items indicates that the developed instrument has a good fit with the construct being measured and is able to adequately represent students' abilities. Several test items even showed a very high correlation with the total score, such as test items 1, 5, 14, 17, 23, 25, and 26 which had a calculated r value of 0.990. This value indicates a very strong relationship between the score of each test item and the total score of the instrument.

These findings indicate that the instruments used have a sufficient level of validity, allowing the data obtained to be considered sufficiently accurate and representative in describing the variables studied. Good validity is a crucial foundation in research because it determines the accuracy of the interpretation of the measurement results. With a valid instrument, the resulting information will more closely reflect the actual condition of the object being measured.

The results of this study align with the findings of Guo et al. (2024), who stated that instruments with good construct validity are able to provide an accurate representation of the measured variables and produce reliable data. Furthermore, Taherdoost (2021) emphasized that validity is a key indicator of instrument quality because it indicates the extent to which the instrument is able to measure the intended concept. Therefore, the high number of valid items in this study indicates that the instrument is of good quality and suitable for use as a research data collection tool.

However, five items still did not meet the validity criteria. This indicates that not all items are able to optimally represent the construct being measured. Therefore, invalid items need to be revised or eliminated, taking into account the appropriateness of the indicators, the clarity of the question wording, and the quality of the answer alternatives, so that they can function better in future uses.
After obtaining valid questions, the analysis continued with a reliability test to determine the level of instrument consistency. The results of the reliability test are presented in Table 2.

Table 2. Results of the Reliability Test of Questions

Statistics	Mark
Number of Valid Questions	25
Cronbach's Alpha	0.972
Category	Very high

Based on Table 2, the reliability analysis results show that the instrument has a Cronbach's Alpha value of 0.972. This value is well above the recommended minimum reliability limit of 0.70. Therefore, the instrument can be categorized as having a very high level of reliability.

These results indicate that all test items used have excellent internal consistency in measuring the same construct. High reliability values indicate that each test item supports each other in measuring student abilities, resulting in stable and consistent data.

This indicates that the test instrument used has a very high level of reliability and can be used as a research data collection tool. The higher the reliability value of an instrument, the smaller the possibility of measurement error. This finding is in line with research by Rezigalla et al. (2024) which states that instruments with reliability values above 0.90 are included in the very high category and indicate excellent measurement quality. Furthermore, Taber (2021) explains that a high Cronbach's Alpha value indicates consistency between items in measuring the same variable.

The difficulty level is used to determine how easy or difficult a question item is for students. The results of the difficulty level analysis are presented in Table 3.

Table 3. Results of the Level of Difficulty of Question Items

Category	Number of Questions	Percentage
Easy	1	4%
Currently	24	96%
Difficult	0	0%
Total	25	100%

Based on Table 3, the analysis results show that of the 25 questions analyzed, 24 (96%) were in the medium category and only 1 (4%) was in the easy category. There were no questions in the difficult category.

These results indicate that the instrument is dominated by questions of moderate difficulty. This is an ideal characteristic for an evaluation instrument, as the questions are neither too easy nor too difficult for students to answer.

This shows that most of the questions are able to measure students' abilities optimally and provide more accurate information regarding the level of mastery of the material. Questions with a moderate level of difficulty are generally more effective in distinguishing students who understand the material from those who do not understand the material. This finding is in line with the research of Rezigalla et al. (2024) which states that questions with a moderate level of difficulty provide the greatest contribution to the quality of measurement because they are able to produce better information than questions that are too easy or too difficult.

Discriminatory power is the ability of a test item to differentiate between students with high and low abilities. The results of the discriminatory power analysis are presented in Table 4.

Table 4. Results of the Analysis of the Distinguishing Power of Questions Categories are determined based on the discrimination index (Corrected Item-Total Correlation).

Category	Number of Questions	Percentage
Very Good (>0.70)	13	52%
Good (0.40–0.70)	12	48%
Enough	0	0%
Total	25	100%

Based on Table 4, 13 questions (52%) were in the very good category and 12 questions (48%) were in the good category. There were no questions in the fair or poor categories. These results indicate that all the questions used performed well in differentiating students based on their level of mastery of the material. The high proportion of questions in the good and very good categories indicates that the instrument is highly suitable as a learning evaluation tool.

This indicates that students with high abilities tend to achieve better scores than those with low abilities. Thus, the instrument is able to effectively identify differences in student abilities. This finding aligns with research by Fauziah et al. (2020), which states that discriminatory power is one of the most important indicators in analyzing question quality because it demonstrates the question's ability to group students based on their competency level.

Distractor effectiveness is one indicator used to determine the quality of answer alternatives in multiple-choice questions. A good distractor should be able to engage students who have not yet mastered the material, preventing them from randomly selecting the correct answer. The results of the distractor effectiveness analysis are presented in Table 5.

Table 5. Results of Distractor Function Analysis

Category	Number of Questions	Percentage
All distractors are working	11	36.67%
There is a distractor that doesn't work	19	63.33%
Total	30	100%
Category	**Number of Questions**	**Percentage**

Based on the data in Table 5, analysis of 30 test items shows that 11 items (36.67%) have all functional distractors, while 19 items (63.33%) still contain one or more distractors that do not fulfill the expected function. A distractor is considered functional if the option is selected by at least 5% of test participants, while a non-functional distractor is characterized by a very low frequency of selection or no one choosing it (Syahrial et al., 2023; , Suharjo & Sumardi, 2024; Nisa, 2022).

These findings indicate that most test items require revision of their answer alternatives, as non-functional distractors indicate that the options are less engaging, too easily recognized, or irrelevant to students' error patterns. This can reduce the item's ability to distinguish between those who understand the material and those who do not (Krisma & 'Adna, 2023; Ashayidatina & Alber, 2023; Istika et al., 2019).

However, the presence of 11 items with fully functional distractors indicates relatively good quality of alternative answer construction for several items. Effective distractors have the potential to improve the quality of evaluation instruments by reducing the likelihood of correct answers through guessing, thereby increasing the item's discriminatory power (Krisma & 'Adna, 2023; Wahyuni et al., 2018; Suharjo & Sumardi, 2024).

These results are consistent with the findings of Syahrial et al. who stated that distractors are considered functional if they are selected by at least 5% of test participants, a criterion that allows distractors to act as effective distractions (Syahrial et al., 2023; Krisma & 'Adna, 2023; Geovana et al. emphasized that distractor effectiveness contributes to improving item quality by helping to differentiate the level of mastery of the material between test participants Faridah, 2021; . In addition, Saputra and Sayono emphasized the need to revise non-functional distractors to make them more logical, relevant to the material, and in line with the characteristics of common student errors (Wahyuni et al., 2018).

Several studies have also emphasized that distractor analysis is an important procedure for evaluating the usefulness of response options in multiple-choice items and can be implemented within both Item Response Theory (IR) and classical theories. This supports the importance of structured analysis of the difficulty level, discrimination, and functionality of distractors to ensure overall item quality (Krisma & 'Adna, 2023; Marambaawang et al., 2023; Nisa, 2022).

4. Conclusion

Based on the analysis of the quality of the multiple-choice assessment instrument on the topic of scarcity for grade 10 students, it can be concluded that the developed instrument generally meets the criteria for a good instrument. Validity test results showed that 25 of the 30 items were declared valid, while 5 items needed revision. The reliability test produced a Cronbach's Alpha value of 0.972, which is categorized as very high, indicating that the instrument has excellent internal consistency (Taber, 2021).

Analysis of the difficulty level showed that most of the questions were in the moderate category, while the results of the discriminatory power showed that all valid questions had good to excellent ability to differentiate students based on their ability level (Fauziah et al., 2020). Meanwhile, analysis of distractor effectiveness showed that several distractors were still not functioning optimally and therefore needed improvement (Syahrial et al., 2023). Overall, the analyzed assessment instruments are suitable for use as learning evaluation tools, although several questions and distractors still need to be revised to improve the quality of measuring student learning outcomes.

References

Aiken, L. R., & Groth-Marnat, G. (2015). *Psychological testing and assessment* (12th ed.). Pearson.
Allen, M. J., & Yen, W. M. (2019). *Introduction to measurement theory*. Waveland Press.
American Educational Research Association, *American Psychological Association, & National Council on Measurement in Education*. (2014). Standards for educational and psychological testing. AERA.
Anshari, M., Putra, RA, & Kurniawan, D. (2024). Item analysis of economics assessment instruments in secondary education. *Journal of Educational Assessment*, 12(1), 45–56.
Apriliana Saputri, N., Hidayat, T., & Rahmawati, D. (2023). Evaluation of learning outcomes through quality assessment instruments in secondary schools. *Indonesian Journal of Education*, 12(3), 211–220.
Arikunto, S. (2010). *Basics of educational evaluation* (Revised edition). Bumi Aksara.
Azwar, S. (2019). *Reliability and validity* (4th ed.). Student Library.
Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). Pearson.
Cohen, L., Manion, L., & Morrison, K. (2020). *Research methods in education* (8th ed.). Routledge.
Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approach* (5th ed.). Sage Publications.
Cronbach, L. J., & Shavelson, R. J. (2021). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, 81(1), 5–14.
Crocker, L., & Algina, J. (2020). *Introduction to classical and modern test theory*. Cengage Learning.

Desjardins, C.D., & Bulut, O. (2018*). Handbook of educational measurement and psychometrics using R*. Chapman & Hall/CRC.
Ebel, R.L., & Frisbie, D.A. (2019). *Essentials of educational measurement* (6th ed.). Prentice Hall.
Fauziah, N., Susanti, R., & Yuliana, E. (2020). Analysis of discrimination index in multiple-choice tests. *Journal of Educational Measurement*, 8(2), 89–97.
Fauzi, A., & Mustika, D. (2022). The role of educational evaluation in improving learning quality. *Journal of Education*, 13(2), 95–104.
Geovana, F., Hidayat, A., & Nugroho, S. (2022). Distractor effectiveness in multiple-choice assessment: Implications for item quality. *Journal of Educational Evaluation*, 10(1), 55–66.
Gierl, M.J., Bulut, O., Guo, Q., & Zhang, X. (2022). Developing, analyzing, and using distractors for multiple-choice tests. *Educational Measurement: Issues and Practice*, 41(2), 6–17.
Gronlund, N. E., & Waugh, C. K. (2020). *Assessment of student achievement* (11th ed.). Pearson.
Guo, Y., Wang, X., & Li, J. (2024). Construct validity in educational measurement: Current perspectives and applications. *Educational Measurement Review*, 18(1), 22–35.
Haladyna, T.M. (2020). *Developing and validating multiple-choice test items* (4th ed.). Routledge.
Haladyna, T. M., & Rodriguez, M. C. (2021). Developing and validating test items (4th ed.). Routledge.
Hanifah, N., Suyana, I., & Setiawan, A. (2014). The use of association multiple-choice questions to increase the discriminating power of questions. *Journal of Education*, 15(1), 23–31.
Iis Makhisoh, & Evi Nurus Suroyya. (2022). Educational evaluation as a basis for improving learning outcomes. *Journal of Elementary Education*, 9(1), 40–48.
Kartowagiran, B. (2018). *Writing test items*. Nuha Medika.
Matondang, Z. (2019). *Learning evaluation*. Unimed Press.
Mulyasa. (2013). *Development and implementation of the 2013 curriculum*. Rosdakarya Youth.
Osterlind, S. J. (2019). *Modern measurement: Theory, principles, and applications of mental appraisal*. Pearson.
Popham, W. J. (2017). *Classroom assessment: What teachers need to know* (8th ed.). Pearson.
Purwanto. (2019). *Evaluation of learning outcomes*. Student Library.
Raharjo, S. (2019). *Learning evaluation*. Pustaka Setia.
Rezigalla, AA, Ahmed, M.H., & Eltahir, M.E. (2024). Psychometric evaluation of multiple-choice questions: Difficulty index, discrimination index, reliability and distractor efficiency. *BMC Medical Education*, 24, 456.
Rizqa, N., Putri, A., & Rahman, H. (2025). Distractor effectiveness analysis in educational assessment. *Journal of Educational Research*, 15(1), 55–67.
Rodriguez, M.C. (2022). Item analysis and distractor functioning in educational assessments. *Educational Assessment Journal*, 27(1), 15–28.
Sireci, S. G., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psychothema*, 26(1), 100–107.
Sudijono, A. (2011*). Introduction to educational evaluation*. Raja Grafindo Persada.
Sudjana, N. (2016). *Assessment of the results of the teaching and learning process*. Rosdakarya Youth.
Suryabrata, S. (2018). *Research methodology*. Raja Grafindo Persada.
Syahrial, M., Putri, DA, & Ramadhan, F. (2023). Effectiveness of distractors in multiple-choice tests: A classical test theory approach. *Journal of Educational Evaluation*, 14(2), 123–134.
Taber, K.S. (2021). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 51(6), 1273–1296.
Taherdoost, H. (2021). Data collection methods and tools for research: Validity and reliability of the research instrument. *International Journal of Academic Research in Management*, 10(2), 10–28.
Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53–55.

Thorndike, R. M., & Thorndike-Christ, T. (2019). *Measurement and evaluation in psychology and education* (9th ed.). Pearson.

Wahyudi, W., Manisa, T., Oktarika, D., Masi, P. W., Ringo, I. K. S., Amri, N., Masi, T. K., Susiaty, U. D., & Sani, L. (2025). *Evaluasi Pembelajaran: Konsep dan Aplikasi di Kelas*. Bukuloka Literasi Bangsa

Widoyoko, EP (2018). *Techniques for preparing research instruments*. Student Library.

Zainal Arifin. (2017). *Learning evaluation*. Bandung: Rosdakarya.