Asatiza : Jurnal Pendidikan
P-ISSN: 2721-0723 | E-ISSN: 2716-3202
https://ejournal.stai-tbh.ac.id/index.php/asatiza

Evaluating Test Item Quality: A Comprehensive Analysis of Economics Multiple-Choice Questions in Indonesian High Schools (Case Study at SMA Negeri 1 Gedangan)

***Arsharil Novan Setiawan[1,a] , Putri Ulfa Kamalia[2,b]**
[1,2]Department of Economics Education, Faculty of Economics and Business, State University of Surabaya, East Java, Indonesia
[a]arsharil.22058@mhs.unesa.ac.id, [b]putrikamalia@unesa.ac.id

ARTICLE INFORMATION

Article History:
Received : 12/08/2025
Revised : 21/08/2025
Accepted : 25/08/2025
Published : 30/09/2025

***Keywords*:**
Question Item Analysis; Daily Assessment; Multiple-Choice Questions

DOI:
https://doi.org/10.46963/asatiza.v6i3.3168

***Correspondence Author:**
arsharil.22058@mhs.unesa.ac.id

Abstract
This study aims to evaluate the psychometric characteristics of multiple-choice question items in the daily assessment instrument of Economics subjects in phase F at SMA Negeri 1 Gedangan. The instrument was developed through the first three stages of a 4D model (Define, Design, Develop) and analyzed using the Classical Test Theory (CTT) approach. Content validity was tested using the Aiken's V index, while empirical validity, difficulty level, differentiation, trick effectiveness, and reliability were analyzed quantitatively. The validation results showed that nine out of ten questions had sufficient content validity ($V \geq 0.50$), but one question was declared invalid and removed. Quantitative analysis showed that 90% of the questions were relatively easy ($P > 0.80$), 35% had good differentiation ($D \geq 0.40$), and the reliability of the instrument was in the medium category (KR-20 = 0.692). Some tricksters do not function optimally, indicating the need for improvements in the design of the answer choice. These findings affirm the importance of question item analysis in improving the quality of economic learning evaluation, especially in supporting formative assessments in accordance with the principles of the Independent Curriculum.

How to cite this article:
Setiawan, A. N., & Kamalia, P. U. (2025). Evaluating test item quality: A comprehensive analysis of economics multiple-choice questions in Indonesian high schools (Case study at SMA Negeri 1 Gedangan). *Asatiza: Jurnal Pendidikan, 6*(3), 287-297. https://doi.org/10.46963/asatiza.v6i3.3168

INTRODUCTION

Education is a multidimensional process characterized by complex interactions among philosophical, psychological, and social factors. These interconnected dimensions create a learning dynamic oriented not only toward knowledge transfer but also toward character formation and the development of human values. As Habsy et al. (2024) suggest, education functions as a vehicle for shaping identity and moral values through the relationships between educators, students, and their social environment. Consequently, the educational process is inherently tied to foundational values such as justice, freedom of thought, and social responsibility.

This process of learning occurs continuously throughout life and is significantly influenced by the quality of social relationships and a supportive environment. Rohanah et al. (2020) emphasize that effective learning arises from dynamic interactions between individuals and their environment, resulting in measurable changes in knowledge, skills, and attitudes. Within this framework, the evaluation of learning outcomes is a critical component of any education system. It serves both as a measure of competency achievement and as a source of feedback to enhance learning effectiveness (Khuzaemah Allaely, 2024; Meguellati et al., 2024).

Evaluation can be conducted through various techniques, including both test and non-test methods. Among these, the multiple-choice test remains prevalent in educational practice due to its advantages in ease of administration, time efficiency, and a high degree of scoring objectivity (Stuart & Helen, 2022; Zuriyati, 2016). However, the overall quality of any evaluation is contingent upon the quality of its individual items. Thus, item analysis is a crucial step in establishing the validity and reliability of an assessment instrument. This process allows for the systematic measurement of key item characteristics, including difficulty index, discrimination power, and distractor effectiveness (Eleragi et al., 2025; Gebremichael et al., 2025).

The application of item analysis spans numerous educational domains, including medical education, teacher training, and competency-based assessment. Research by Rahmi and Friyatmi (2022) demonstrates that item quality significantly influences assessment outcomes, particularly in measuring higher-order thinking skills. Prasetyo (2019) highlights the role of item analysis in ensuring that assessment items align with curriculum standards and learning objectives. Furthermore, multiple studies affirm that item analysis provides educators with valuable, ongoing feedback for refining evaluation instruments. This systematic practice is essential for teachers to identify flawed items, implement revisions, and develop more valid and reliable questions that accurately reflect learning goals (Savika et al., 2025; Siregar et al., 2024).

Within the specific context of economic education, evaluation presents distinct challenges. The discipline requires students to develop a deep conceptual understanding and the ability to apply knowledge to real-world situations. Therefore, effective assessments must be capable of measuring higher-order thinking skills, such as analysis, synthesis, and evaluation, which are vital for fostering the critical and creative thinking necessary to address complex problems (Akhiralimi et al., 2022; Mytra et al., 2021). Despite this need, high school economics teachers often encounter difficulties constructing items that align with learning outcomes and curriculum standards. As Kurniawati (2021) notes, limited resources and a lack of specialized training in item construction are primary obstacles to implementing high-quality evaluations.

The recent implementation of Indonesia's Merdeka Curriculum

(Independent Curriculum) has further emphasized the integral role of formative assessment within the learning process. At the senior high school level, Phase F is designed to promote diagnostic and continuous assessment, wherein daily assessments aim to monitor student progress consistently and provide constructive feedback (Kemendikbudristek, 2022). However, a conspicuous gap exists in the literature: there is no comprehensive study that specifically evaluates the quality of multiple-choice questions used in daily economic assessments within Phase F. This gap underscores the need for empirical research to support the implementation of more effective and accurate formative assessments in this critical subject area.

To provide a robust foundation for this analysis, the present study adopts the framework of Classical Test Theory (CTT). CTT is a psychometric approach that enables the quantitative evaluation of key item characteristics, including validity, reliability, difficulty, and discrimination (Arifin, 2013; Meguellati et al., 2024; Pratama, 2019). Hartono et al. (2024) demonstrate that CTT can be effectively employed to improve the quality of evaluation instruments at both tertiary and secondary education levels. Its flexibility in data analysis and interpretation of results makes it particularly suitable for application in the context of daily formative assessments.

Based on this background, this study aims to evaluate the psychometric characteristics of multiple-choice questions used in the daily assessment of Phase F Economics at SMA Negeri 1 Gedangan. It is expected that this evaluation will yield actionable recommendations to improve the quality of economic learning evaluation instruments, strengthen formative assessment practices, and ultimately support the more effective implementation of the Merdeka Curriculum at the secondary school level.

METHOD

This study employs a developmental research design aimed at producing a daily assessment instrument comprising multiple-choice questions for Phase F Economics. The development process adhered to the 4D model (Define, Design, Develop, Disseminate) proposed by Thiagarajan (1974), with a focus on the first three stages: Define, Design, and Develop. The primary activities for each of these stages are detailed in Table 1.

Table 1. Stages of the 4D Model in the Research

Phase	Main Activities
Define	Conducting a literature review, performing a needs analysis, and identifying basic competencies.
Design	Preparing question matrices, defining indicators, and formulating the assessment format.
Develop	Conducting expert validation, performing question testing, and executing item analysis.

The research participants consisted of 33 Phase F students from SMA Negeri 1 Gedangan, Sidoarjo. A purposive sampling technique was utilized to select the sample based on the following criteria: having completed the National Income instructional material, willingness to participate in the trial, and the absence of

cognitive barriers that could impair comprehension of the test items. Prior to full implementation, a pilot test was conducted with five students to ensure the clarity of the questions and the appropriateness of their difficulty level. This research received formal approval from the school administration and was conducted in accordance with established ethical principles for educational research.

Data collection yielded both qualitative and quantitative data. Qualitative data were derived from expert evaluations of the question items using validation sheets. Quantitative data were obtained from student test results, which were subsequently analyzed to determine the psychometric characteristics of each item. The validation process was conducted in two stages. The first stage involved internal validation by two content experts, who assessed each question item based on three aspects: material substance, construction, and language clarity. Assessments were made using a four-category scale: Poor, Sufficient, Good, and Excellent. Content validity was quantitatively assessed using Aiken's V formula (Aiken, 1985), with a score of ≥ 0.70 considered to indicate acceptable validity.

The second stage involved external validation through field testing with the student participants. The answer responses were analyzed to measure each item's difficulty index, discriminating power, empirical validity, reliability, and distractor effectiveness. The difficulty index (P) was categorized as easy ($P > 0.70$), medium ($0.30 \leq P \leq 0.70$), or difficult ($P < 0.30$). Discriminating power (D) was classified as good ($D \geq 0.40$), fair ($0.20 \leq D < 0.40$), or poor ($D < 0.20$). Empirical validity was calculated by determining the point-biserial correlation between the item score and the total test score. Reliability for the entire test was analyzed using the Kuder-Richardson 20 (KR-20) formula. Distractor effectiveness was analyzed using the Anbuso V.8 software. These classification guidelines for difficulty and discrimination are consistent with the standards outlined by Haladyna (2004).

Data analysis incorporated both qualitative and quantitative techniques. Qualitative analysis was applied to interpret the expert judgments regarding item feasibility. Quantitative analysis was performed on the student test data using Classical Test Theory (CTT) formulas, facilitated by Microsoft Excel and Anbuso V.8 software. This mixed-methods approach was employed to obtain a comprehensive evaluation of the quality of the developed assessment instrument.

RESULT AND DISCUSSION

This study aimed to develop a daily assessment instrument comprising multiple-choice questions aligned with the characteristics of Phase F Economics students. The development process followed the first three stages of the 4D model (Define, Design, and Develop), executed systematically to ensure the resulting instrument possessed adequate validity and reliability.

Stages of Instrument Development

The Define stage involved identifying learning needs, analyzing core competencies, and examining learning

objectives alongside student characteristics. The outcomes of this analysis served as the foundation for designing relevant and contextual question indicators.

During the Design stage, a test blueprint was constructed based on the established indicators. The item writing process carefully considered cognitive, affective, and psychomotor dimensions to ensure alignment with the intended learning outcomes.

In the Develop stage, the drafted questions underwent validation by two experts in the field of educational evaluation. The validation assessed material accuracy, construction quality, and language clarity. Subsequently, the questions were tested on students to facilitate quantitative analysis.

Content Validity

Content validity was assessed quantitatively by two validators using the Aiken's V index (Aiken, 1985), calculated with the formula:

$V = \Sigma s / [n (c - 1)]$

where:

s : r - lo
n : number of raters
r : rating given by the reviewer
Lo : lowest validity rating number
c : highest validity rating number

The results of the Aiken's V calculations are presented in Figure 1.

Figure 1 *Results of Aiken's V Calculations for Item Validity*

Aiken's V calculation

Reviewer	1		2		3		4		5		6		7		8		9		10	
	r	s	r	s	r	s	r	s	r	s	r	s	r	s	r	s	r	s	r	s
R1	2	1	3	2	3	2	2	1	3	2	2	1	4	3	2	1	3	2	4	3
R2	3	2	3	2	2	1	3	2	3	2	3	2	1	0	2	1	2	1	2	1
∑s	3		4		3		3		4		3		3		2		3		4	
V	0.5		0.6666666667		0.5		0.5		0.6666666667		0.5		0.5		0.3333333333		0.5		0.6666666667	
Category	Currently		Currently		Currently		Currently		Currently		Currently		Currently		Low		Currently		Currently	

Of the ten questions developed, nine obtained scores between 0.50 and 0.667, indicating a sufficient level of validity warranting minor revisions. One question, item 8, obtained a value of 0.333 and was deemed invalid; it was consequently excluded from subsequent trial stages. An Aiken's V value above 0.70 is generally considered to indicate high validity, while values between 0.50 and 0.69 suggest a need for improvement in specific aspects (Aiken, 1985).

Question Item Analysis

Following content validation, the nine questions that passed were subjected to quantitative analysis based on three

primary parameters: difficulty level, discriminating power, and distractor effectiveness.

Difficulty Level: The analysis revealed that eight questions had a difficulty index (P) above 0.80, classifying them as relatively easy. One question was in the medium category with an index of 0.45. No questions were classified as difficult. The distribution of difficulty levels is presented in Table 2.

Table 2. *Difficulty Level of Multiple-Choice Question Items*

Category (P)	Item Numbers	Count	%
Difficult (P < 0.30)	-	0	
Medium (0.30 ≤ P ≤ 0.70)	2	1	10%
Easy (P > 0.70)	1 3 4 5 6 7 9 10	8	90%

This distribution indicates that the majority of the items were insufficiently challenging to optimally gauge students' higher-order thinking abilities.

Discriminating Power: Three questions exhibited a discriminating power (D) above 0.40, categorizing them as good due to their ability to differentiate between high- and low-achieving students. One question was in the acceptable category without revision (0.30 ≤ D < 0.40), two required revision (0.20 ≤ D < 0.30), and three were classified as poor (D < 0.20) and require significant revision or replacement. Item 1 demonstrated very low discriminating power and was declared empirically invalid. The results are summarized in Table 3.

Table 3. *Discriminating Power of Multiple-Choice Question Items*

Category (D)	Item Numbers	Count	%
Good (D ≥ 0.40)	4 5 9	3	35%
Acceptable (0.30 ≤ D < 0.40)	2	1	10%
0.20<DP<0.29 Revision Needed (0.20 ≤ D < 0.30)	3 6	2	20%
Poor (D < 0.20)	1 7 10	3	35%

Distractor Effectiveness: Analysis of answer choices indicated that several distractors failed to function as intended. As shown in Figure 2, some incorrect options were selected by no students or only a handful, suggesting they were either non-plausible or easily identifiable as incorrect choices.

Figure 2. *Distractor Effectiveness Analysis for a Sample Item*

Item No	Different Power		Difficulty Level		Alternative Answers Ineffective	Information
	Coefficient	Information	Coefficient	Information		
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	0.109	Not good	0.939	Easy	BE	Not good
2	0.407	Good	0.667	Currently	-	Good
3	0.714	Good	0.909	Easy	BE	Distractor Revision
4	0.422	Good	0.818	Easy	BE	Distractor Revision
5	0.591	Good	0.848	Easy	AE	Distractor Revision
6	0.714	Good	0.909	Easy	A	Distractor Revision
7	0.357	Good	0.848	Easy	A	Distractor Revision
8	0.474	Good	0.848	Easy	B	Distractor Revision
9	0.542	Good	0.939	Easy	AB	Distractor Revision

Instrument Reliability

The instrument's reliability, calculated using the Kuder-Richardson 20 (KR-20) formula, yielded a coefficient of 0.692. This value indicates moderate internal consistency, sufficient for use in low-stakes daily formative assessment. While approaching the common educational research threshold of 0.70, this result suggests room for improvement to achieve high reliability (Nitko & Brookhart, 2011).

Student Learning Outcomes

All 33 students participated in the instrument trial. The mean score was 86.94, with a maximum of 100 and a minimum of 13. The standard deviation of 15.14 indicates a notable variation in student ability levels. Individual mastery was achieved by 97% of students, and classical mastery reached 80%. These results suggest that most students met the established learning targets, though the high average and low difficulty index should be considered when interpreting this success. The summary is presented in Table 4.

Table 4. *Descriptive Statistics of Student Scores*

Metric	Value
Total Values	2869
Mean Score	86,94
Highest Score	100
Lowest Score	13
Standard Deviation	15,14
Number of Students	33 People
Number of Students Mastered	32 People
Individual Mastery Rate	1 Person
Individual absorbency	97%
Classical Mastery Rate	80%

Discussion

Instrument Validity and Internal Consistency

The content validation results indicated that nine of the ten items possessed sufficient validity (Aiken's V = 0.50–0.667) for use with minor revisions. This aligns with research by Eleragi et al. (2025) and Gebremichael et al. (2025), who emphasize the critical role of content validity in ensuring items align with learning objectives. However, the empirical analysis revealed that not all items passing expert validation functioned optimally. The very low discriminating power of item 1, rendering it empirically invalid, underscores the perspective of Haladyna (2004) that a comprehensive validation process must integrate both qualitative (expert judgment) and quantitative (empirical) analyses to ensure overall instrument quality.

Difficulty Level and Learning Implications

The finding that 90% of the items were categorized as easy (P > 0.80) suggests the instrument in its current form may lack the rigor to effectively identify variations in student ability. This finding is consistent with Gebremichael et al. (2025), who note that overly easy items can limit an assessment's diagnostic utility. Within the framework of the Merdeka Curriculum, which prioritizes formative assessment and the development of critical thinking, it is essential that assessments include a range of item difficulties. Subsequent revisions should therefore focus on incorporating more items of medium difficulty and developing challenging items that effectively measure higher-order cognitive skills.

Differentiation and Quality of Evaluation

While 33.3% of items demonstrated good discriminating power (D ≥ 0.40), an equal proportion (33.3%) were classified as poor (D < 0.20). Low discrimination indicates an item's failure to reliably differentiate between high- and low-achieving students, a core principle of effective diagnostic assessment (Haladyna, 2004). Items with poor discrimination provide little information for learning improvement and can result in a homogenized score distribution that fails to reflect the true spectrum of student understanding.

Distractor Effectiveness and Answer Choice Design

The analysis revealed that several distractors were non-functional, meaning they were not selected by any lower-achieving students. According to Eleragi et al. (2025), effective distractors should reflect common student misconceptions and be contextually plausible. The ineffectiveness of the distractors in this study highlights a need for targeted teacher professional development focused on the principles of effective multiple-choice item writing, particularly in designing attractive and plausible incorrect options.

Reliability and Practical Implications

The obtained KR-20 reliability coefficient of 0.692 indicates moderate internal consistency. This level is generally acceptable for formative assessments intended to provide feedback for learning (Nitko & Brookhart, 2011). However, to enhance the instrument's utility for more formal or summative purposes, improvements in item quality—specifically, by increasing the proportion of items with good discrimination and a balanced range of difficulties—are necessary to boost reliability above the 0.70 threshold.

Linkage with the Independent Curriculum

The findings of this study resonate with the formative assessment principles of the Merdeka Curriculum for Phase F senior high school. The developed instrument provides a foundational snapshot of student achievement. However, to fully realize the curriculum's goal of diagnostic assessment that drives learning reflection (Kemendikbudristek, 2022; Savika et al., 2025), the instrument must be refined. Future development must focus on enhancing item discrimination and distractor effectiveness to create a more powerful tool for identifying specific student needs and supporting the curriculum's objective of fostering critical and analytical thinking.

CONCLUSION

This study successfully developed a prototype daily assessment instrument comprising ten multiple-choice questions for Phase F Economics. Expert validation, quantified via Aiken's V index, confirmed that nine questions met the threshold for sufficient content validity (V ≥ 0.50), while one item (number 8) was deemed invalid (V = 0.333) and was excluded from subsequent testing.

Quantitative analysis of student responses (N=33) yielded the following key findings:

1. Difficulty Level: Ninety percent of the questions were classified as easy (P > 0.80), with the remaining 10% in the moderate range (P = 0.45).

2. Discriminating Power: The items demonstrated varied discriminatory ability: 35% were categorized as good ($D \geq 0.40$), 10% as acceptable ($0.30 \leq D < 0.40$), 20% required revision ($0.20 \leq D < 0.29$), and 35% were poor ($D < 0.19$).
3. Distractor Effectiveness: On average, two distractors per question were non-functional, indicating a failure to effectively attract students from the correct answer.
4. Reliability: The instrument demonstrated moderate internal consistency, with a KR-20 coefficient of 0.692.

Item 1 was declared empirically invalid due to its very low discriminating power and non-functional distractors. Consequently, only eight of the original ten items are recommended for use, and these require targeted revisions.

These findings underscore that a systematic development process incorporating item analysis is crucial for enhancing the quality of learning evaluation. However, significant attention must be paid to crafting a varied distribution of item difficulty and designing plausible distractors to ensure the instrument is both diagnostic and fair. While this instrument shows potential for formative assessment purposes, its current psychometric properties preclude its recommendation for high-stakes summative assessment without substantial improvement.

This research contributes to the implementation of the Merdeka Curriculum by highlighting the gap between intention and practice in formative assessment. The prevalence of easy, low-discrimination items suggests a continued tendency among teachers to construct factual, recall-based questions rather than those that measure higher-order thinking skills like analysis, synthesis, and evaluation. This indicates a pressing need for comprehensive teacher professional development focused on item-writing techniques, including the design of plausible distractors that reflect common student misconceptions (Haladyna, 2004; Eleragi et al., 2025).

Theoretically, this study emphasizes the critical role of psychometric analysis in economic education at the secondary level, an area that has received limited scholarly attention. The developed instrument serves as a preliminary model for competency-based assessment; however, its limitations necessitate further refinement before broader application.

The generalizability of these conclusions is constrained by the study's limitations, including its focus on a single topic (National Income) and a relatively small, homogeneous sample size (N=33). Future research should aim to:

1. Develop instruments encompassing a broader range of economic topics.
2. Validate findings with larger, more diverse student populations to enhance external validity.
3. Employ modern test theory approaches, such as Item Response Theory (IRT), to provide more robust and nuanced psychometric insights.
4. Conduct longitudinal studies to evaluate the impact of improved assessment instruments on sustained student learning outcomes.

REFERENCES

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Akhiralimi, N., Fitriani, A., Sari, I. P., & Maulidah, R. (2022). Analisis keterampilan berpikir tingkat tinggi siswa SMA pada pembelajaran fisika. *Jurnal Eksakta Pendidikan (Jep)*, *6*(2), 204–213. https://doi.org/10.24036/jep/vol6-iss2/696

Arifin, Z. (2013). *Evaluasi pembelajaran*. Remaja Rosdakarya.

Eleragi, A. M. S., Miskeen, E., Hussein, K., Rezigalla, A. A., Adam, M. I. E., Al-Faifi, J. A., Alhalafi, A., Al Ameer, A. Y., & Mohammed, O. A. (2025). Evaluating the multiple-choice questions quality at the College of Medicine, University of Bisha, Saudi Arabia: a three-year experience. *BMC Medical Education*, *25*(1), 2–9. https://doi.org/10.1186/s12909-025-06700-2

Gebremichael, M. W., Baraki, B., Mehari, M. A., & Assalfew, B. (2025). Item analysis of multiple choice questions from assessment of health sciences students, Tigray, Ethiopia. *BMC Medical Education*, *25*(1). https://doi.org/10.1186/s12909-025-06904-6

Habsy, B. A., Satsabhila, A., Syakilah, N. J. F., & Sanallah, A. K. (2024). Hakikat pendidikan dan pembelajaran, serta tanggung jawab dan kompetensi guru. *Tsaqofah*, *4*(6), 4189–4203. https://doi.org/10.58578/tsaqofah.v4i6.4158

Haladyna, T. M. (2004). *Developing and Validating Multiple-choice Test Items*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203825945

Hartono, I. D. I., Tenriawaru, A. B., & Ningsih, K. (2024). Analisis butir soal penilaian sumatif IPA kelas vii SMP Negeri 3 Pontianak menggunakan anates. *Jurnal Kajian Pembelajaran Dan Keilmuan*, *8*(2), 162–171. https://doi.org/10.26418/jurnalkpk.v8i2.78282

Kemendikbudristek. (2022). *Panduan Pembelajaran dan Asesmen*. Badan Standar, Kurikulum, Dan Asesmen Pendidikan Kementerian Pendidikan, Kebudayaan, Riset, Dan Teknologi Republik Indonesia, 123.

Khuzaemah Allaely, N. S. (2024). Pentingnya proses evaluasi dalam pembelajaran di sekolah menengah pertama. *Jurnal Bahasa Dan Sastra Indonesia Serta Pengajarannya*, *2*(2), 139–148. https://journal.uinjkt.ac.id/index.php/bestari/article/view/46246

Kurniawati, F. (2021). Exploring teachers' inclusive education strategies in rural Indonesian primary schools. *Educational Research*, *63*(2), 198–211. https://doi.org/10.1080/00131881.2021.1915698

Meguellati, S., Samia, A., Ferhat, A., Djelloul, A., & Khalifa, Z. A. (2024). A critical analysis of the use of classical test theory (CTT) in psychological testing: A comparison with item response theory (IRT). *Pakistan Journal of Life and Social Sciences*, *22*(2), 9442–9449. https://doi.org/10.57239/PJLSS-

2024-22.2.00715

Mitra Prawiki Suci, & Helendra. (2022). Analisis kualitas butir soal ujian akhir semester ganjil tahun pelajaran 2020/2021 mata pelajaran biologi kelas x SMA Negeri 1 Teluk Sebong. *Biodidaktika: Jurnal Biologi Dan Pembelajarannya*, *17*(2), 13–23.

Mytra, P., Wardawaty, A., & Kusnadi, R. (2021, September). Society 5.0 in education: Higher order thinking skills. In *BIS-HSS 2020: Proceedings of the 2nd Borobudur International Symposium on Humanities and Social Sciences*, *BIS-HSS 2020, 18 November 2020, Magelang, Central Java, Indonesia (Vol. 242)*. European Alliance for Innovation. http://dx.doi.org/10.4108/eai.18-11-2020.2311812

Nitko, A. J., & Brookhart, S. M. (2011). *Educational assessment of students* (6th ed.). Pearson.

Pratama, D. (2019). Analysis of Clasical Test Theory (Ctt) Approch on academic ability test instrument. *Jisae: Journal of Indonesian Student Assesment and Evaluation*, *5*(2), 43–54. https://doi.org/10.21009/jisae.052.05

Rahmi, E., & Friyatmi, F. (2022). Financial Management Behavior of Student during the Covid 19 Pandemic *BT - Proceedings of the Eighth Padang International Conference on Economics Education, Economics, Business and Management, Accounting and Entrepreneurship (PICEEBA-8 2021)*. 663–668. https://doi.org/10.2991/aebmr.k.220702.099

Rohanah, L., Mirawati, M., & Anwar, W. S. (2020). Pengaruh interaksi sosial terhadap aktivitas belajar peserta didik. *Jurnal Pendidikan Dan Pengajaran Guru Sekolah Dasar (JPPGuseda)*, *03*(September), 139–143. http://journal.unpak.ac.id/index.php/jppguseda

Savika, H. I., Zuhriyah, I. A., & Susilawati, S. (2025). Peran guru dalam analisis butir soal di sekolah dasar. *JIIP - Jurnal Ilmiah Ilmu Pendidikan*, *8*(3), 3313–3319. https://doi.org/10.54371/jiip.v8i3.7534

Sholikhah, S., Sugiharto, B., & Raharjo, S. B. (2023). Analisis kemampuan berpikir tingkat tinggi (HOTS) peserta didik di SMA Negeri 1 Ngemplak dalam menyelesaikan soal asam basa. *Prosiding SNPS (Seminar Nasional Pendidikan Sains)*, *September*, 267–275. https://proceeding.uns.ac.id/snps/issue/view/21

Siregar, N. H., Remiswal, R., & Khadijah, K. (2024). Analisis butir soal ujian tengah semester pada mata pelajaran pendidikan agama Islam. *Urwatul Wutsqo: Jurnal Studi Kependidikan dan Keislaman*, *13*(2), 179–189. https://doi.org/10.54437/urwatulwutsqo.v13i2.1637

Thiagarajan, S. (1974). *Instructional development for training teachers of exceptional children: A sourcebook*. Council for Exceptional Children.

Zuriyati, Z. (2016). *Analisis butir soal*. Kencana.