Nazhruna: Jurnal Pendidikan Islam
Vol. 9 Issue 1, 2026. pp. 187-206 E-ISSN: 2614-8013
DOI: https://doi.org/10.31538/nzh.v9i1.418

Authentic Mathematics Assessment Using an Integrated Deep Learning and Adiwiyata Testlet Model for Elementary Schools and Madrasah Ibtidaiyah

Syukrul Hamdi[1], Lia Yuliana[2], Anisa Dwi Oktarina[3], Kana Hidayati[4], Elly Arliani[5], Nurul Mu'minin Mz[6]
Universitas Negeri Yogyakarta, Indonesia[1, 2, 3, 4, 5]
Sekolah Dasar Negeri Sosrowijayan, Yogyakarta, Indonesia [6]
Correspondence Email: syukrulhamdi@uny.ac.id

Received: 15-10-2025 Revised: 28-12-2025 Accepted: 07-02-2026

Abstract

This study aimed to develop an authentic mathematics assessment for elementary schools in the form of a testlet-based instrument integrated with deep learning principles and the Adiwiyata context within the framework of transformative Islamic education in the Special Region of Yogyakarta. Employing basic research with an embedded mixed-methods design, the development process adapted the Plomp model and the instrument development framework of Oreondo and Antonio. Data were collected through teacher needs surveys, expert validation, readability testing, and field trials involving 462 fifth-grade students from elementary schools and Islamic elementary schools. Quantitative analyses included Content Validity Index (CVI), Aiken's V, Cronbach's Alpha, Classical Test Theory (CTT), and Item Response Theory (IRT) using the 2-PL Graded Response Model. The results indicate that the developed instruments meet acceptable psychometric standards, with Aiken's V values ranging from 0.75 to 1.00 and high internal consistency for both the testlet instrument (Cronbach's Alpha = 0.845) and the environmental awareness questionnaire (Cronbach's Alpha = 0.850). Item analysis shows adequate discrimination and a structured progression of difficulty, although one item exhibited low discrimination in the 2-PL GRM, highlighting the importance of IRT-based diagnostics for testlet refinement. Descriptive findings reveal that students demonstrate high levels of environmental awareness, particularly in the knowledge and attitude dimensions, while mathematical achievement remains low on non-routine items. Correlation analysis shows no significant relationship between environmental awareness and mathematical ability. Methodologically, this study contributes a validated and contextually grounded assessment framework that integrates expert judgment, reliability analysis, and complementary CTT–IRT procedures. Theoretically, the findings reconceptualize authentic assessment as a diagnostic bridge rather than a direct causal link between affective values and cognitive performance, demonstrating that environmental concern functions as a potential cognitive resource only when explicitly activated within mathematical tasks.

Keywords: Adiwiyata, Authentic Mathematics Assessment, Deep Learning, Islamic Education, Testlet.

Abstrak

Penelitian ini bertujuan untuk mengembangkan penilaian matematika autentik untuk sekolah dasar dalam bentuk instrumen berbasis testlet yang terintegrasi dengan prinsip pembelajaran mendalam dan konteks Adiwiyata dalam kerangka pendidikan Islam transformatif di Daerah Istimewa Yogyakarta. Proses pengembangan menggunakan penelitian dasar dengan desain campuran tertanam, mengadopsi Model Plomp dan kerangka kerja pengembangan instrumen Oreondo dan Antonio. Data dikumpulkan melalui survei kebutuhan guru, validasi ahli, uji keterbacaan, dan uji lapangan yang melibatkan 462 siswa kelas lima dari sekolah dasar dan sekolah dasar Islam. Analisis kuantitatif mencakup Indeks Validitas Konten (CVI), Aiken's V, Cronbach's Alpha, Teori Uji Klasik (CTT), dan Teori Respons Item (IRT) menggunakan Model Respons Tergradasi 2-PL. Hasil menunjukkan bahwa instrumen yang dikembangkan memenuhi standar psikometrik yang dapat diterima, dengan nilai Aiken's V berkisar antara 0,75 hingga 1,00 dan konsistensi internal yang tinggi untuk instrumen testlet (Cronbach's Alpha = 0,845) dan kuesioner kesadaran lingkungan (Cronbach's Alpha = 0,850). Analisis item menunjukkan diskriminasi yang memadai dan kemajuan kesulitan yang terstruktur, meskipun satu item menunjukkan diskriminasi rendah dalam Model Respons Tergradasi 2-PL, menyoroti pentingnya diagnostik berbasis IRT untuk penyempurnaan testlet. Temuan deskriptif menunjukkan bahwa siswa menunjukkan tingkat kesadaran lingkungan yang tinggi, terutama dalam dimensi pengetahuan dan sikap, sementara pencapaian matematika tetap rendah pada soal-soal non-rutin. Analisis korelasi menunjukkan tidak ada hubungan yang signifikan antara kesadaran lingkungan dan kemampuan matematika. Secara metodologis, studi ini menyumbangkan kerangka penilaian yang valid dan berakar pada konteks, yang mengintegrasikan penilaian ahli, analisis reliabilitas, dan prosedur CTT–IRT yang komplementer. Secara teoretis, temuan ini merekonseptualisasi penilaian autentik sebagai jembatan diagnostik rather than hubungan kausal langsung antara nilai afektif dan kinerja kognitif, menunjukkan bahwa kepedulian lingkungan berfungsi sebagai sumber daya kognitif potensial hanya ketika secara eksplisit diaktifkan dalam tugas-tugas matematika.

***Kata Kunci**: Adiwiyata, Penilaian Matematika Autentik, Pembelajaran Mendalam, Pendidikan Islam, Testlet*

INTRODUCTION

Assessment in basic education plays a central role in shaping how student learning is defined, measured, and valued. In mathematics education at the elementary level, particularly in Indonesian and Islamic school contexts, assessment practices remain predominantly focused on final answers and procedural accuracy, providing limited insight into students' reasoning processes, problem-solving strategies, and the application of concepts in meaningful contexts (Griffin, 2017; Morris et al., 2021; Moss & Brookhart, 2019; Selvaraj et al., 2021; Sewagegn, 2020). This problem is amplified in elementary mathematics, where students often perceive mathematics as abstract and detached from daily life, leading to superficial learning outcomes (Hernandez-Martinez & Vos, 2018; Sachdeva & Eggen, 2021; Vos, 2018). Consequently, assessment models that generate both quantitative evidence of achievement and qualitative information about students' thinking are required to capture elementary students' mathematical competencies more authentically.

In response to these challenges, authentic assessment has been widely promoted as an approach that enables students to demonstrate mathematical knowledge, reasoning, and attitudes through contextual and meaningful tasks (Gravett, 2025; Koh, 2017). Empirical studies in elementary mathematics education show that context-based assessment supports students in

applying mathematical concepts to real-life situations, such as calculating daily needs, interpreting environmental data, or designing simple patterns (Ahdhianto & Santi, 2020; Wijaya et al., 2018), while reducing reliance on rote memorization (Acharya, 2017). However, most existing studies emphasize project-based or portfolio-based formats, which, although valuable, often function as isolated assessment activities rather than systematically structured instruments capable of capturing progressive understanding across related tasks. As a result, alternative formats such as testlet-based assessments remain relatively underexplored in elementary mathematics contexts.

From the perspective of Islamic education, assessment is expected to evaluate not only cognitive achievement but also the development of affective and behavioral dimensions aligned with the goal of forming insan kamil, namely learners who integrate intellectual, moral, and spiritual development (Achmad & Prastowo, 2022). Empirical studies in Islamic elementary schools indicate that mathematics assessment practices remain largely exam-oriented, limiting the assessment of values such as honesty, responsibility, and social awareness (Black & Wiliam, 2009; Brookhart, 2018; Rahman, 2025; Sabri & Retnawati, 2019; Safitri et al., 2020; Sahin, 2018). In parallel, national initiatives such as the Adiwiyata program emphasize environmental responsibility as a core educational value (Hajar, 2024; Utaya & Wafaretta, 2021). However, empirical evidence suggests that environmental values are more frequently addressed at the instructional level than systematically embedded within assessment instruments, particularly in mathematics learning, resulting in fragmented alignment between values, instruction, and assessment.

Another limitation in the existing literature concerns the depth and structure of assessment implementation in schools, particularly in relation to 21st-century learning demands. While 21st-century education emphasizes critical thinking, problem solving, and the ability to apply knowledge across contexts, classroom assessments often fail to capture these competencies, especially in elementary mathematics (Maxwell et al., 2021). Recent discussions on deep learning define it as an assessment-oriented approach that focuses on uncovering students' conceptual understanding, reasoning processes, and meaningful application of knowledge (Jiang, 2022; Torshizi & Bahraman, 2019; J.-L. Zhang, 2020). Empirical classroom-based studies report that teachers experience difficulties in translating deep learning principles into concrete assessment instruments, leading to assessments that are fragmented and procedural (Darling-Hammond & Hyler, 2020; Lin, 2018; Lukitasari et al., 2021). Testlet models, which comprise interconnected items designed to trace learning trajectories, offer a potential solution because they support adaptive and diagnostic assessment. However, their application in elementary mathematics, particularly in Islamic school settings, remains limited (Guo et al., 2024; Ma et al., 2023). Empirical studies integrating testlet-based authentic assessment with environmental contexts in Islamic elementary schools are still scarce (Mundofi, 2025; Triyandana et al., 2024), especially in the Special Region of Yogyakarta, where assessment innovations have not been systematically documented despite the growth of Islamic-based schools.

Based on these empirically grounded gaps, this study aims to develop an authentic mathematics assessment for elementary schools in the form of a testlet model integrated with deep learning principles and Adiwiyata values within the framework of transformative Islamic

education in the Special Region of Yogyakarta. In this study, transformative Islamic education is operationalized as an educational approach that integrates academic competence, ethical reasoning, spiritual awareness, and social–environmental responsibility within learning and assessment practices. This study seeks to test the argument that a testlet-based authentic assessment can more comprehensively capture students' mathematical understanding, reasoning processes, and the internalization of Islamic and environmental values compared to conventional assessments. Accordingly, the research is guided by the following research questions: (RQ1) How can an authentic mathematics assessment instrument based on a testlet model be developed by integrating deep learning principles and the Adiwiyata context for elementary and Islamic elementary schools? (RQ2) To what extent does the developed testlet-based authentic assessment instrument meet the criteria of content validity, reliability, and item quality as analyzed using Classical Test Theory (CTT) and Item Response Theory (IRT)? (RQ3) What are the profiles of students' mathematical abilities and environmental awareness as measured by the developed instruments, and how do these profiles differ between elementary school and Islamic elementary school students? (RQ4) Is there a significant relationship between students' environmental awareness and their mathematical abilities as measured through the testlet-based authentic assessment? Theoretically, this research contributes to the literature by positioning assessment at the intersection of mathematics education, deep learning, environmental education, and Islamic educational values. Practically, the findings are expected to provide a structured assessment model that supports both measurement and value formation, while maintaining a clear distinction between assessment and instructional intervention in elementary mathematics learning.

METHOD

This study employed basic research using an embedded mixed-methods design, in which qualitative data supported instrument development and refinement, while quantitative data were used to evaluate psychometric properties (Morse, 2016). The qualitative component informed needs analysis, expert validation, and readability testing, whereas the quantitative component examined validity, reliability, item quality, group differences, and correlations.

The research design adapted Plomp's (2013) educational design research model and was operationally integrated with the instrument development framework of Oreondo and Antonio (1984). As illustrated in Figure 1, Plomp's model structured the overall iterative stages of needs analysis, design, development, evaluation, and revision, while the procedures of Oreondo and Antonio were embedded within the design and evaluation phases to guide indicator formulation, expert judgment, readability testing, and psychometric refinement. This integration resulted in six development stages, from teachers' needs analysis to iterative revision prior to field implementation.

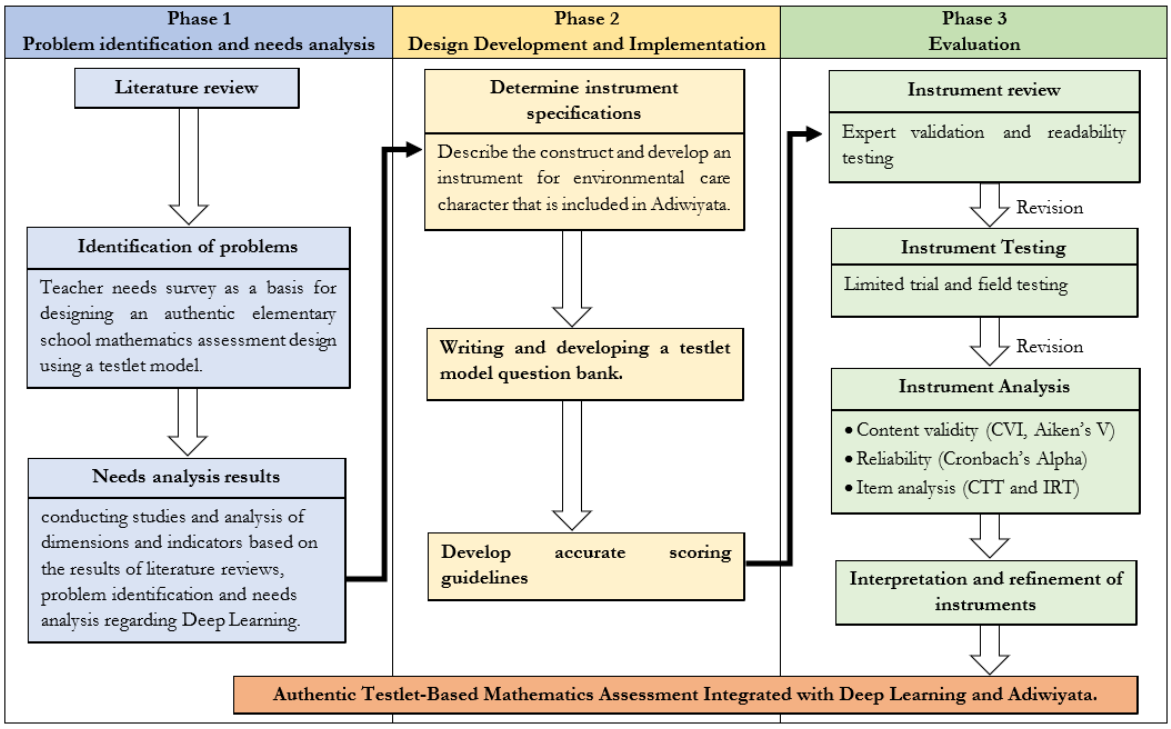

Figure 1. Development Framework of a Testlet-Based Authentic Mathematics Assessment Integrating Deep Learning and Adiwiyata

The study was conducted in the Special Region of Yogyakarta, Indonesia, from May to September 2025. Two instruments were developed: a testlet-based mathematics assessment and an environmental awareness questionnaire. Testlets consisted of clusters of interrelated items sharing a common contextual stimulus and were scored polytomously (0–3) to capture partial understanding across increasing levels of conceptual mastery. This structure produced ordinal data suitable for Classical Test Theory (CTT) and Item Response Theory (IRT) analyses using the Graded Response Model (Hambleton et al., 1991). The environmental awareness questionnaire was developed based on the Adiwiyata framework and character education theory, operationalized into three dimensions: knowledge, attitude, and action. The instrument consisted of 30 items, distributed across knowledge (10 items), attitude (10 items), and action (10 items), measured using a four-point Likert scale with balanced positive and negative statements (Lickona, 1996).

Participants included 3 experts as validators, 10 elementary school teachers involved in needs analysis and readability testing, and 462 fifth-grade students aged 11–12 years, comprising 253 elementary school (ES) students and 209 Islamic elementary school (IES) students. Students were selected using purposive sampling based on school typology comparison. The sample size met recommended thresholds for stable estimation in two-parameter IRT models (de Ayala, 2022). Qualitative data were analyzed descriptively to inform instrument revision. Quantitative data were analyzed using Content Validity Index (CVI) and Aiken's V for content validity with Aiken's $V \geq 0.80$ (Aiken, 1985), Cronbach's Alpha for reliability $\alpha \geq 0.70$ (Tavakol & Wetzel, 2020), and CTT and IRT analyses, including 1-PL and 2-PL Graded Response Models (DeMars, 2018). Because the data were not normally distributed, Mann–Whitney U tests were used for group comparisons and Spearman's rho for correlation analysis, with a

significance level of $p < 0.05$. Ethical approval was obtained from school authorities. Written parental consent and student assent were secured, participation was voluntary, and all data were anonymized to ensure confidentiality.

RESULTS AND DISCUSSION

Instrument Validity and Reliability Test

This study was conducted to develop an authentic mathematics assessment instrument based on a testlet model integrated with deep learning principles and the Adiwiyata context for elementary school students. The research process was carried out in six stages, namely (1) a teacher needs survey to map assessment needs, (2) analysis of relevant deep learning components as a basis for development, (3) description of the environmental awareness construct to be measured, (4) compiling a testlet model question bank and an environmental awareness questionnaire instrument, (5) expert validation and readability testing of the initial instrument model, and (6) revising and refining the instrument before implementing it in the field stage.

During the teacher needs assessment stage, information was obtained regarding the difficulties encountered in measuring students' higher-order thinking skills, particularly in the context of environment-based mathematics learning. Teachers tend to use conventional tests that emphasize procedural memorization, so that assessments have not been able to reveal students' mathematical reasoning and connection skills. The next stage is the analysis of deep learning components that can be integrated into the assessment. The results of the study show that the aspects of concept connection, reflective thinking skills, and the application of concepts in new situations are the main components that need to be accommodated in the development of instruments.

Next, the construct of environmental awareness was described based on the Adiwiyata program indicators. This construct covers aspects of knowledge, attitudes, and actions related to environmental awareness that are aligned with Islamic educational values. Based on this foundation, an initial instrument was developed consisting of a testlet question bank to measure mathematical cognitive abilities and an affective questionnaire to assess environmental awareness.

The expert validation and readability testing stages are important steps to ensure the quality of the instrument before it is used in the field. Validators consisting of assessment experts, mathematics education experts, and environmental education experts provide input regarding the suitability of indicators, question wording, and language readability. Teachers were also involved in assessing the level of suitability of the instruments to classroom conditions, so that the instruments developed could be more contextual and easier for students to understand. The results of the content validity of the testlet model based on expert assessment are presented in Table 1:

Table 1. Validation of Testlet Model Questions

No.	Score	Validity	No.	Score	Validity	No.	Score	Validity
1.1	0.97	High	2.3	1.00	High	4.2	1.00	High
1.2	0.75	Medium	3.1	1.00	High	4.3	0.97	High
1.3	1.00	High	3.2	0.97	High	5.1	1.00	High

2.1	0.86	High	3.3	1.00	High	5.2	1.00	High
2.2	1.00	High	4.1	1.00	High	5.3	0.94	High

Source: Researcher data analysis (2025)

Table 2. Validation of the Environmental Care Questionnaire Instrument

No.	Score	Validity	No.	Score	Validity	No.	Score	Validity	No.	Score	Validity
1	1.00	High	8	0.86	High	15	1.00	High	22	0.86	High
2	1.00	High	9	0.97	High	16	0.86	High	23	0.94	High
3	1.00	High	10	0.86	High	17	1.00	High	24	0.97	High
4	0.86	High	11	1.00	High	18	0.86	High	25	1.00	High
5	0.94	High	12	0.92	High	19	1.00	High	26	0.89	High
6	0.86	High	13	1.00	High	20	0.89	High	27	0.97	High
7	0.92	High	14	0.89	High	21	0.97	High	28	0.86	High

Source: Researcher data analysis (2025)

From the results of the testlet model item validation, a high level of validity was obtained except for item number 1.2, which had a moderate level of validity. Therefore, it can be concluded that the developed item instrument has met the content validity criteria. However, the validators also provided a number of inputs and suggestions for improvement so that the instrument would be more relevant to the research context and easier for respondents to understand. Some of the suggestions focused on the clarity of the question wording, consistency of the writing format, relevance of the context to students' daily experiences, and technical improvements to the illustrations and answer choices. All of these comments and suggestions were important references in refining the instrument before it was used in the field implementation stage. A summary of the comments and suggestions from the validators is presented in Table 3 below.

Table 3. Comments and Suggestions from Reviewers

No.	Comments and Suggestions
1	The wording of questions needs to be simplified and adjusted to the reading ability of elementary school students.
2.	Question 1.2 needs to be adjusted to match its indicator.
3.	Illustrations or images must be clear, proportional, and sized as necessary.
4.	The writing of symbols, units, and question formats must be consistent with the rules.
5.	Each question should only measure one aspect/construct in order to maintain content validity.
6.	In questionnaire instruments, statements must be relevant and realistic, and positive and negative statements should be arranged so that they are not easily guessed.

Source: Researcher data analysis (2025)

In addition, the environmental awareness questionnaire instrument was also validated. Table 2 shows that the questionnaire validation results were in the high validity category. This was also evident from the comments and suggestions provided by the reviewers, which did not require major revisions. The reviewers only provided suggestions on the editorial section, stating that the statements needed to be made more relevant to the conditions of students at school and that positive and negative statements should not be easily predictable. After the

instrument was confirmed to be valid through a validity test, it was tested on a limited sample. The test data was then used to test the reliability of the instrument to determine the internal consistency of the questions and statements.

Table 4. Reliability of Test Items and Questionnaires

No	Types of Instruments	Cronbach's Alpha Value
1	Testlet questions	0.845
2.	Questionnaire	0.850

Source: Researcher data analysis (2025)

The reliability test results in Table 4 show that both instruments have excellent internal consistency. The testlet instrument obtained a Cronbach's Alpha value of 0.845, while the environmental concern questionnaire obtained a value of 0.850. These values are above the minimum threshold of 0.70, so it can be concluded that the instruments are suitable for further research.

Item Quality Analysis

After the instruments were declared reliable, the next analysis focused on item quality. The analysis was conducted using the Classical Test Theory (DeVellis, 2006) and Item Response Theory (2021) approaches to obtain more comprehensive information about the characteristics of each item.

Table 5. CTT Analysis Results for Testlet Questions

Item	Sample SD	Item-Total	Item-Tot.woi	Difficulty	Discrimination	Item Reliab	Item Rel.woi
T1	0.850	0.705	0.546	2.066	1.394	0.599	0.464
T2	0.922	0.852	0.746	1.621	1.729	0.784	0.687
T3	0.963	0.902	0.823	1.469	1.948	0.868	0.792
T4	0.882	0.819	0.702	1.186	1.529	0.722	0.619
T5	0.810	0.631	0.457	1.084	1.116	0.510	0.370

Source: Researcher data analysis (2025)

The results of item analysis using Classical Test Theory (CTT) in Table 5 show that the variation in respondent scores is quite good, as indicated by standard deviation values ranging from 0.810 to 0.963. The item–total correlations, which represent the correlation between each item score and the total test score, are all positive and relatively high, with the highest value found in T3 (0.902), indicating strong internal consistency between items. The corrected item–total correlations (Item-Tot.woi), calculated by correlating each item with the total score excluding that item, further support the contribution of each item to the overall construct. Item difficulty values vary from 1.084 (T5) to 2.066 (T1), indicating a reasonable level of difficulty across items. The discrimination indices for all items are positive and relatively high, with the highest value observed in T3 (1.948), confirming that each item has good discriminatory power in relation to respondents' abilities. In addition, item reliability indicates the contribution of each item to the overall test reliability, while reliability without item (Item Rel.woi) represents the reliability of the test if the corresponding item is removed.

Overall, the CTT analysis results confirm that the testlet questions have good measurement quality, although some items with lower reliability values (e.g., T5) can be reviewed for improvement. However, because CTT is highly dependent on the sample, item analysis was continued using Item Response Theory (IRT) to obtain more accurate and sample-free information. In this study, the 1-PL GRM (1-Parameter Logistic Graded Response Model) and 2-PL GRM (2-Parameter Logistic Graded Response Model) were used, which are suitable for items with categorical or partial scores.

Table 6. Comparison of 1-PL GRM and 2-PL GRM Analysis Results on Testlet Questions

Item	1-PL GRM			2-PL GRM		
	Threshold (Extrmt1–3)	Dscrmn	Interpretation	Threshold (Extrmt1–3)	Dscrmn	Interpretation
T1	-2.223 ; -0.887 ; 0.466	2.16	Easy, high discrimination	-2.706 ; -0.952 ; 0.611	1.645	Easy, moderate discrimination
T2	-1.488 ; -0.205 ; 1.204	2.16	Moderate, high discrimination	-1.103 ; -0.024 ; 1.253	2.778	Moderate, discrimination is very high
T3	-1.181 ; 0.014 ; 1.350	2.16	Moderate, high discrimination	-0.779 ; 0.160 ; 1.266	4.030	Moderate, discrimination is very high (most informative)
T4	-0.921 ; 0.525 ; 1.876	2.16	Somewhat difficult, high discrimination	-0.698 ; 0.774 ; 2.150	2.048	Somewhat difficult, high discrimination
T5	-0.844 ; 0.758 ; 2.112	2.16	Difficult, high discrimination	-1.048 ; 1.435 ; 3.769	0.982	Most difficult, low discrimination

Source: Researcher data analysis (2025)

Table 6 shows that in the 1-PL GRM model, all items have the same discrimination parameter (set at 2.16), so that the quality of discrimination cannot be compared between items. In contrast, in the 2-PL GRM model, discrimination variation is clearly visible: T3 has the highest value (4.030), indicating high effectiveness in distinguishing respondents based on ability, while T5 has the lowest value (0.982), which means it is less effective in distinguishing participants' abilities. Meanwhile, the difficulty parameters (thresholds) in both models show a gradation in difficulty levels between items. T1 is relatively the easiest, while T5 is the most difficult, especially in the highest category (3.769). Thus, although the instrument as a whole is of good quality, item T5 still needs to be revised to improve measurement accuracy.

Descriptive Character Caring for the Environment

Table 7. Comparison of Descriptive Statistics from ES and IES Student Questionnaires

Statistics	ES (N = 253)	IES (N = 209)	Overall (N = 462)
Minimum Value	64	61	61
Maximum Value	100	100	100
Mean (average)	88,04	85,23	86,57
Standard Deviation	7,68	7,47	7,66

Source: Researcher data analysis (2025)

The results in Table 7 show that the average score for environmental awareness was in the high category, both in elementary school (88.04) and Islamic elementary school (85.23). This indicates that, in general, students have shown good concern for the environment. The standard deviation values for both groups are relatively moderate (7.68 for elementary school and 7.47 for Islamic elementary school), so the variation in scores between students is still within reasonable limits. Combined, the average score for environmental awareness is 86.57, which is in the high category, confirming that the majority of students have positive environmental awareness. Although the overall picture shows a high category, a more detailed analysis is needed to see how the profile of students' environmental awareness is distributed across each aspect, namely knowledge, attitude, and action. These details are presented in Table 8.

Table 8. Comparison of the Environmental Awareness Profiles of ES and IES Students by Aspects (%)

Aspect	Category	ES (%)	IES (%)
Knowledge	Very Low	1.98	1.44
	Low	7.11	3.83
	Moderate	24.90	30.62
	High	23.72	26.79
	Very High	42.29	37.32
Attitude	Very Low	1.58	3.35
	Low	9.09	13.88
	Moderate	12.65	11.96
	High	22.92	29.19
	Very High	53.75	41.63
Action	Very Low	13.04	22.01
	Low	17.00	31.58
	Moderate	17.00	13.40
	High	16.21	11.48
	Very High	36.76	21.53

Source: Researcher data analysis (2025)

Table 8 shows a comparison of the distribution of environmental awareness scores between elementary school and Islamic elementary school students. In terms of knowledge, both elementary schools and Islamic elementary schools were dominated by the high and very high categories, although the proportion of elementary school students in the very high category (42.29%) was higher than that of Islamic elementary schools (37.32%). In terms of attitude, a similar pattern is seen, with a dominance of high and very high categories, but the proportion of very high is greater in elementary school (53.75%) than in Islamic elementary school (41.63%). The most striking difference is seen in terms of action, where elementary school students are relatively more in the high and very high categories (52.97%) than Islamic elementary school students (32.01%). Conversely, the proportion of low and very low categories was greater in IES (53.59%) than in ES (30.04%).

In general, both ES and IES students have positive environmental knowledge and attitudes. However, implementation in the form of concrete actions is more consistent among ES students than IES students. These findings confirm the gap between cognitive and affective aspects and actual behavior, especially among IES students.

Descriptive Mathematical Ability

Table 9. Comparison of Descriptive Statistics of ES and IES Students' Mathematics Test Results

Statistics	ES (N = 253)	IES (N = 209)	Overall (N = 462)
Minimum Value	0	0	0
Maximum Value	67	80	80
Mean (average)	25.45	32.63	28.80
Standard Deviation	13.64	15.85	15.20

Source: Researcher data analysis (2025)

In contrast to the high level of environmental awareness, the average math test scores were still low. Elementary school students scored an average of 25.45, while junior high school students scored 32.63. The combined score for all respondents was 28.80, which is considered low. Pedagogically, this low level of achievement indicates that most Grade 5 students have not yet mastered the expected mathematical competencies, particularly in understanding concepts, applying procedures, and solving contextual problems aligned with the curriculum standards. The standard deviation in both groups (13.64 for elementary school and 15.85 for Islamic elementary school) shows that there is quite a wide range of achievement among students. This gives an early indication that even though students have good environmental awareness, their math skills are still relatively limited. To obtain a more comprehensive picture, the next analysis not only highlights the average scores but also the descriptive distribution of students' answers on each testlet item. Details of the distribution can be seen in Table 10.

Table 10. Distribution of Answer Categories for All Respondents on the Mathematics Test (Testlet)

Category Description	Score	Point 1 (f/%)	Point 2 (f/%)	Point 3 (f/%)	Point 4 (f/%)	Point 5 (f/%)	Total (f/%)
Students do not master and understand the subject matter well.	0	18 (3.9%)	107 (23.2%)	145 (31.4%)	146 (31.6%)	166 (35.9%)	582 (25.2%)
Students tend to guess.	0	14 (3.0%)	95 (20.6%)	75 (16.2%)	28 (6.1%)	60 (13.0%)	272 (11.8%)
Students get answers by guessing.	0	7 (1.5%)	33 (7.1%)	52 (11.3%)	55 (11.9%)	46 (10.0%)	193 (8.4%)
Students are not thorough/careful,	0	13 (2.8%)	23 (5.0%)	33 (7.1%)	15 (3.2%)	33 (7.1%)	117 (5.1%)

so questions are easily overlooked.							
Students understand the basic concepts of the questions.	1	73 (15.8%)	67 (14.5%)	78 (16.9%)	103 (22.3%)	77 (16.7%)	398 (17.2%)
Students understand the basic concepts but are not thorough.	1	66 (14.3%)	27 (5.9%)	14 (3.0%)	37 (8.0%)	20 (4.3%)	164 (7.1%)
Students understand basic and intermediate concepts.	2	108 (23.4%)	74 (16.1%)	52 (11.0%)	52 (11.3%)	32 (6.9%)	318 (13.8%)
Students master and understand the subject matter well.	3	163 (35.3%)	36 (7.8%)	13 (2.8%)	26 (5.6%)	28 (6.0%)	266 (11.5%)
Total (f/%)		462 (100%)	462 (100%)	462 (100%)	462 (100%)	2310 (100%)	462 (100%)

Source: Researcher data analysis (2025)

The combined results of all respondents (N = 462) show a shift in the dominance of answer descriptions between items. In the initial items, most students were still able to master the material well (35.3%), but this proportion decreased dramatically to only 6.0% in Item 5. Conversely, the category of not mastering the material increased sharply from 3.9% in Item 1 to 35.9% in Item 5. The category of understanding basic concepts was relatively stable at around 15–22%, while the categories of guessing and guessing appeared more frequently in the final questions. This pattern confirms that the level of difficulty of the questions increased as the item number increased, thereby reducing students' mastery of the material.

When analyzed by grade level, a pattern emerges that is relatively similar to the combined results, but with some important differences. Among elementary school students, the proportion of mastery of the material in Item 1 was 27.7% and decreased to 9.0% in Item 5, while the category of non-mastery increased to 28.5%. Among Islamic elementary school students, mastery of the material in Item 1 was even higher (44.5%), but also declined sharply to 4.3% in Item 5, while the category of not mastering the material increased to 30.6%. Although Islamic elementary school students initially showed better mastery, both groups experienced a significant decline as the level of difficulty of the questions increased. These descriptive findings provide an initial indication of differences in achievement between ES and IES students, particularly in the pattern of mastery of questions with higher levels of difficulty. These differences will be further analyzed statistically through difference tests, while also being reinforced with correlation analysis to see the relationship between environmental awareness and students' mathematical abilities.

Difference and Correlation Tests

Table 11. Results of Difference and Correlation Tests

Analysis	Group	Test	Statistics	p-value	Interpretation
Normality	ES	Kolmogorov–Smirnov	—	< 0.05	Not normally distributed
Normality	IES	Kolmogorov–Smirnov	—	< 0.05	Not normally distributed
Difference (Environmental Awareness)	ES vs. IES	Mann–Whitney U	U = 21,206	0.000	Significant difference
Difference (Mathematics Score)	ES vs. IES	Mann–Whitney U	U = 19,697	0.000	Significant difference
Correlation (Combined)	All students	Spearman's ρ	r = 0.012	0.796	Not significant
Correlation (ES)	ES	Spearman's ρ	r = 0.076	0.228	Not significant
Correlation (IES)	IES	Spearman's ρ	r = –0.022	0.757	Not significant

Source: Researcher data analysis (2025)

Note. Nonparametric tests were used because the data were not normally distributed ($p < 0.05$).

Table 11 presents the results of the difference and correlation analyses. Because the data were not normally distributed, Mann–Whitney U tests were used to examine group differences and Spearman's correlation was used to analyze relationships. The results show significant differences between elementary school and Islamic elementary school students in both environmental awareness and mathematics achievement. However, no significant correlation was found between environmental awareness and mathematics test scores, either in the combined data or when analyzed by school level.

Discussion

The validity and reliability test results indicate that the authentic assessment instrument based on a testlet model integrated with deep learning principles and the Adiwiyata context meets acceptable quality criteria. High content validity indices (Tables 1 and 2) suggest strong alignment between the developed indicators and the intended cognitive and affective constructs, while high internal consistency (Cronbach's Alpha > 0.80) indicates reliable measurement. Nevertheless, it is important to acknowledge the methodological limitations of these indices. Content validity measures such as CVI and Aiken's V rely heavily on expert judgment and do not capture empirical construct functioning in real testing conditions. Similarly, high reliability coefficients may partly reflect item homogeneity rather than optimal construct coverage, raising the possibility of item redundancy, particularly within clustered formats such as testlets. Therefore, the interpretation of high reliability in this study is complemented by item-level analyses using CTT and IRT, which provide more nuanced evidence regarding item discrimination, difficulty, and information. These findings are in line with the views of Olfos

and Zulantay (2007) and Prayitno and Jaedun (2018) that a good authentic assessment must meet two main requirements, namely content validity and consistency reliability. Thus, the development of this instrument can answer the needs of teachers in measuring higher-order thinking skills, which have been difficult to achieve through conventional tests.

Analysis of item quality using both CTT and IRT indicates that the testlet instrument demonstrates adequate discrimination power and a purposeful progression of item difficulty, which is essential for capturing students' learning trajectories within a testlet structure. Item T3 emerged as the most informative item due to its high discrimination parameter, indicating its effectiveness in differentiating students across varying ability levels and its role as a key diagnostic item within the testlet sequence. In contrast, Item T5 showed relatively low discrimination in the 2-PL GRM model, suggesting that although it represents a higher level of cognitive demand, its wording or scoring structure may obscure students' true understanding and introduce construct-irrelevant difficulty (Holland & Stevens, 2021; Schmucker & Moore, 2025). This pattern is critical in testlet-based assessments, as weak discrimination in later items can disrupt the intended progression from conceptual understanding to advanced reasoning (Krajcik & Shin, 2023; Seah & Horne, 2020). These findings further demonstrate that combining CTT and IRT analyses provides complementary insights for testlet refinement: while CTT confirms acceptable score consistency at the test level, IRT reveals how individual items function across ability levels, enabling targeted revisions rather than wholesale item removal (Alqarni, 2019; D. Zhang et al., 2023).

From an affective perspective, the descriptive results indicate that both elementary school and Islamic elementary school students demonstrate a high level of environmental awareness, particularly in the knowledge and attitude dimensions (Table 8). This high level may be attributed to the consistent exposure of students to environmental themes through school routines, curriculum integration, and extracurricular activities commonly associated with environmentally oriented programs such as Adiwiyata, which emphasize environmental knowledge formation and value internalization at the school level (Saadah et al., 2023). However, despite strong knowledge and positive attitudes, a noticeable gap emerges between these dimensions and actual environmental behavior, especially among Islamic elementary school students, who tend to score lower on the action dimension. This pattern aligns with the Theory of Planned Behavior, which posits that attitudes alone are insufficient to predict behavior without adequate perceived behavioral control and supportive social norms (Hagger et al., 2022). In this context, students' environmental actions are likely constrained by situational factors such as limited opportunities for practice, variations in school facilities, and differing levels of reinforcement from teachers and families. These findings highlight the importance of contextual and action-oriented learning approaches, such as school-based environmental projects, to bridge the gap between environmental awareness and actual behavior.

Meanwhile, the mathematics test results show low achievement in both groups of students. The average scores for ES (25.45) and IES (32.63) are both in the low category, with a sharp decline in mastery of the material on items with a high level of difficulty (Table 10). These findings reinforce the results of research by Kablan and Uğur (2021) that one of the main challenges in mathematics learning is the transition from routine questions to non-routine questions that require conceptual connections and reflective reasoning. The low achievement

of students in answering medium to difficult questions shows that their higher-order thinking skills are still not optimal. This condition reaffirms the urgency of deep learning-based authentic assessment to develop more meaningful mathematical competencies.

Interestingly, the correlation analysis indicates that environmental concern is not significantly associated with students' mathematical ability. Rather than merely contradicting prior studies that report positive links between affective traits and academic achievement, this finding suggests that the relationship between affective dispositions and cognitive performance is highly domain-specific. Previous research demonstrating positive associations typically involves affective constructs that are directly related to learning processes, such as academic motivation, self-discipline, or school engagement (Hagger & Hamilton, 2019; Li et al., 2021). In contrast, environmental concern represents a value-oriented and context-dependent disposition, which may not directly support mathematical reasoning unless it is explicitly activated within learning tasks. From this perspective, the absence of correlation implies that high environmental awareness does not automatically translate into improved mathematical performance when assessment tasks primarily emphasize cognitive processing without requiring the application of environmental values. This interpretation aligns with character education theory, which conceptualizes cognitive, affective, and behavioral domains as interconnected yet functionally distinct, rather than linearly causal (Araújo et al., 2020; Bates, 2021). Consequently, this finding highlights the need for intentionally designed learning and assessment models that explicitly integrate environmental values into mathematical problem solving, enabling affective dispositions to operate as cognitive resources rather than parallel outcomes.

Thus, this study contributes in two main aspects. First, methodologically, this study successfully developed a valid, reliable, and contextual authentic assessment instrument based on the principles of Adiwiyata and deep learning, which can be used as a reference for research and assessment practices in elementary schools. Second, theoretically, this study enriches the study of the relationship between students' cognitive abilities and affective character by providing evidence that the two are not always directly correlated. Therefore, further research needs to focus on the implementation of instruments in real learning and the exploration of a stronger integrative model between character education and students' cognitive achievements. In line with the framework of transformative Islamic education, the authentic assessment developed here not only functions as a tool for measuring cognitive abilities but also as a means of instilling spiritual, moral, and ecological values that strengthen the character of insan kamil (the concept of a complete human being in Islamic educational thought).

CONCLUSION

This study demonstrates the feasibility of developing a testlet-based authentic mathematics assessment integrated with deep learning principles and the Adiwiyata context. Empirically, the instrument met acceptable quality standards, as indicated by high content validity indices (Aiken's V ranging from 0.75 to 1.00) and strong internal consistency for both the cognitive testlet (Cronbach's Alpha = 0.845) and the environmental awareness questionnaire (Cronbach's Alpha = 0.850). Item-level analyses further showed that most testlet items exhibited adequate discrimination and a structured progression of difficulty, although one item

(T5) displayed low discrimination in the 2-PL GRM model, underscoring the importance of IRT-based diagnostics for testlet refinement.

Methodologically, the core contribution of this study lies in the integrated use of expert judgment–based validity indices, reliability analysis, and complementary CTT–IRT procedures to evaluate both test-level consistency and item functioning. This approach enabled targeted revision of problematic items rather than reliance on aggregate scores alone, which is particularly critical in clustered formats such as testlets.

Substantively, the findings reveal a clear contrast between students' affective and cognitive outcomes. Although students demonstrated high environmental awareness, particularly in the knowledge and attitude dimensions (mean scores above 85), no significant correlation was found between environmental awareness and mathematical achievement (Spearman's $\rho \approx 0.01$, $p > 0.05$). This result suggests that affective dispositions related to environmental concern operate as relatively independent constructs unless explicitly activated within mathematical tasks, rather than functioning as automatic predictors of achievement.

Several limitations should be acknowledged. The study was conducted within a specific regional and grade-level context using purposive sampling, which constrains generalizability. In addition, the findings reflect instrument performance rather than instructional impact. Future research should therefore examine the use of the developed assessment in instructional settings, test its sensitivity to learning interventions, and explore its applicability across broader populations. Overall, this study positions authentic assessment not as a normative ideal but as a measurable and diagnostically informative approach for examining the alignment and misalignment between cognitive performance and affective dispositions in elementary mathematics education.

ACKNOWLEDGMENT

This research was funded by the Directorate of Research, Technology, and Community Service (DRTPM), Ministry of Higher Education, Science, and Technology of the Republic of Indonesia, through the Fundamental Research – Regular scheme, based on contract number B/17.61/UN34.9/PT/2025.

REFERENCES

Acharya, B. R. (2017). Factors affecting difficulties in learning mathematics by mathematics learners. *International Journal of Elementary Education*, *6*(2), 8–15. https://doi.org/10.11648/j.ijeedu.20170602.11

Achmad, G. H., & Prastowo, A. (2022). Authentic assessment techniques on cognitive aspects in Islamic religious education learning at elementary school level. *Jurnal Ilmiah Sekolah Dasar*, *6*(1), 75–84. https://doi.org/10.23887/jisd.v6i1.43470

Ahdhianto, E., & Santi, N. N. (2020). The Effect of Metacognitive-Based Contextual Learning Model on Fifth-Grade Students' Problem-Solving and Mathematical Communication Skills. *European Journal of Educational Research*, *9*(2), 753–764. https://doi.org/10.12973/eu-jer.9.2.753

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

Alqarni, A. M. (2019). A Design for Comparing CTT and IRT in Test Assembly, Scoring, and Argumentation: Differences Among Reliability, Information, and Validation. *I-Manager's Journal on Educational Psychology*, *13*(2), 1. https://doi.org/10.26634/jpsy.13.2.16084

Araújo, D., Davids, K., & Renshaw, I. (2020). Cognition, emotion and action in sport: an ecological dynamics perspective. In *Handbook of sport psychology* (pp. 535–555). Wiley Online Library. https://doi.org/10.1002/9781119568124.ch25

Bates, A. (2021). *Moral emotions and human interdependence in character education: Beyond the one-dimensional self*. Routledge.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (Formerly: Journal of Personnel Evaluation in Education)*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Bock, R. D., & Gibbons, R. D. (2021). *Item response theory*. John Wiley & Sons.

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, *3*, 22. https://doi.org/10.3389/feduc.2018.00022

Darling-Hammond, L., & Hyler, M. E. (2020). Preparing educators for the time of COVID… and beyond. *European Journal of Teacher Education*, *43*(4), 457–465. https://doi.org/10.1080/02619768.2020.1816961

de Ayala, R. J. (2022). *The Theory and Practice of Item Response Theory*. Guilford Publications.

DeMars, C. E. (2018). Classical test theory and item response theory. In *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 49–73). Wiley Online Library. https://doi.org/10.1002/9781118489772.ch2

DeVellis, R. F. (2006). Classical test theory. *Medical Care*, *44*(11), S50–S59. https://doi.org/10.1097/01.mlr.0000245426.10853.30

Gravett, K. (2025). Authentic assessment as relational pedagogy. *Teaching in Higher Education*, *30*(3), 608–622. https://doi.org/10.1080/13562517.2024.2380997

Griffin, P. (2017). *Assessment for teaching* (Second). Cambridge University Press.

Guo, L., Zhou, W., & Li, X. (2024). Cognitive Diagnosis Testlet Model for Multiple-Choice Items. *Journal of Educational and Behavioral Statistics*, *49*(1), 32–60. https://doi.org/10.3102/10769986231165622

Hagger, M. S., Cheung, M. W.-L., Ajzen, I., & Hamilton, K. (2022). Perceived behavioral control moderating effects in the theory of planned behavior: A meta-analysis. *Health Psychology*, *41*(2), 155. https://doi.org/10.1037/hea0001153

Hagger, M. S., & Hamilton, K. (2019). Grit and self-discipline as predictors of effort and academic attainment. *British Journal of Educational Psychology*, *89*(2), 324–342. https://doi.org/10.1111/bjep.12241

Hajar, A. (2024). Transforming Islamic Education for Environmental and Social Sustainability. *Sinergi International Journal of Islamic Studies*, *2*(2), 82–95. https://doi.org/10.61194/ijis.v2i2.601

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Hernandez-Martinez, P., & Vos, P. (2018). "Why do I have to learn this?" A case study on students' experiences of the relevance of mathematical modelling activities. *ZDM*, *50*(1),

245–257. https://doi.org/10.1007/s11858-017-0904-2

Holland, J., & Stevens, N. (2021). *Guidelines for the development of multiple choice items & assessments*. Royal College of Surgeons in Ireland. https://doi.org/10.25419/rcsi.13947164.v1

Jiang, R. (2022). Understanding, Investigating, and promoting deep learning in language education: A survey on chinese college students' deep learning in the online EFL teaching context. *Frontiers in Psychology*, *13*, 955565. https://doi.org/10.3389/fpsyg.2022.955565

Kablan, Z., & Uğur, S. S. (2021). The relationship between routine and non-routine problem solving and learning styles. *Educational Studies*, *47*(3), 328–343. https://doi.org/10.1080/03055698.2019.1701993

Koh, K. H. (2017). Authentic assessment. In *Oxford research encyclopedia of education*.

Krajcik, J., & Shin, N. (2023). Student conceptions, conceptual change, and learning progressions. In *Handbook of research on science education* (pp. 121–157). Routledge.

Li, J.-B., Bi, S.-S., Willems, Y. E., & Finkenauer, C. (2021). The association between school discipline and self-control from preschoolers to high school students: a three-level meta-analysis. *Review of Educational Research*, *91*(1), 73–111. https://doi.org/10.3102/0034654320979160

Lickona, T. (1996). Eleven principles of effective character education. *Journal of Moral Education*, *25*(1), 93–100. https://doi.org/10.1080/0305724960250110

Lin, C.-L. (2018). The development of an instrument to measure the project competences of college students in online project-based learning. *Journal of Science Education and Technology*, *27*(1), 57–69. https://doi.org/10.1007/s10956-017-9708-y

Lukitasari, M., Hasan, R., Sukri, A., & Handhika, J. (2021). Developing Student's Metacognitive Ability in Science through Project-Based Learning with E-Portfolio. *International Journal of Evaluation and Research in Education*, *10*(3), 948–955. https://doi.org/10.11591/ijere.v10i3.21370

Ma, W., Wang, C., & Xiao, J. (2023). A testlet diagnostic classification model with attribute hierarchies. *Applied Psychological Measurement*, *47*(3), 183–199. https://doi.org/10.1177/01466216231165315

Maxwell, A. E., Warner, T. A., & Guillén, L. A. (2021). Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—Part 1: Literature review. *Remote Sensing*, *13*(13), 2450. https://doi.org/10.3390/rs13132450

Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education*, *9*(3), e3292. https://doi.org/10.1002/rev3.3292

Morse, J. M. (2016). *Mixed method design: Principles and procedures*. Routledge.

Moss, C. M., & Brookhart, S. M. (2019). *Advancing formative assessment in every classroom: A guide for instructional leaders*. ASCD.

Mundofi, A. A. (2025). Integration of Deep Learning Approach in Transforming Islamic Religious Education Learning in Schools: A Pedagogical and Technological Study. *Journal of Asian Primary Education (JoAPE)*, *2*(1), 79–90. https://doi.org/10.59966/joape.v2i1.1787

Olfos, R., & Zulantay, H. (2007). Reliability and validity of authentic assessment in a web based course. *Journal of Educational Technology & Society*, *10*(4), 156–173. https://www.jstor.org/stable/jeductechsoci.10.4.156

Oreondo, L. L., & Antonio, E. M. D. (1984). *Evaluating Educational Outcomes*. Rex Book Store. https://books.google.co.id/books?id=8xNMBn4bn8oC

Plomp, T. (2013). Educational design research: An introduction. In T. Plomp & N. Nieveen (Eds.), *Educational design research* (Vol. 1).

Prayitno, S. H., & Jaedun, M. P. D. (2018). Authentic assessment competence of building construction teachers in indonesian vocational schools. *Journal of Technical Education and Training*, *10*(1). https://doi.org/10.30880/jtet.2018.10.01.008

Rahman, N. A. (2025). Competency-Based and Ethical Assessment Models in Contemporary Islamic Pedagogy. *Sinergi International Journal of Islamic Studies*, *3*(1), 57–69. https://doi.org/doi.org/10.61194/ijis.v3i1.710

Saadah, L., Rusnaini, R., & Muchtarom, M. (2023). The internalization of school environmental care through Adiwiyata program. *Jurnal Civics: Media Kajian Kewarganegaraan*, *20*(2), 205–213. https://doi.org/10.21831/jc.v20i2.56549

Sabri, M., & Retnawati, H. (2019). The implementation of authentic assessment in mathematics learning. *Journal of Physics: Conference Series*, *1200*(1), 12006. https://doi.org/10.1088/1742-6596/1200/1/012006

Sachdeva, S., & Eggen, P.-O. (2021). Learners' critical thinking about learning mathematics. *International Electronic Journal of Mathematics Education*, *16*(3), em0644. https://doi.org/10.29333/iejme/11003

Safitri, D. I., Mudzanata, M., & Putri, A. D. S. (2020). The Implementation of Authentic Assessment in Thematic Learning in Elementary Schools. *International Journal of Elementary Education*, *4*(2), 255–260. https://doi.org/doi.org/10.23887/ijee.v4i2.25551

Sahin, A. (2018). Critical issues in Islamic education studies: Rethinking Islamic and Western liberal secular values of education. *Religions*, *9*(11), 335. https://doi.org/10.3390/rel9110335

Schmucker, R., & Moore, S. (2025). The Impact of Item-Writing Flaws on Difficulty and Discrimination in Item Response Theory. *ArXiv Preprint ArXiv:2503.10533*. https://doi.org/10.48550/arXiv.2503.10533

Seah, R., & Horne, M. (2020). The construction and validation of a geometric reasoning test item to support the development of learning progression. *Mathematics Education Research Journal*, *32*(4), 607–628. https://doi.org/10.1007/s13394-019-00273-2

Selvaraj, A. M., Azman, H., & Wahi, W. (2021). Teachers' Feedback Practice and Students' Academic Achievement: A Systematic Literature Review. *International Journal of Learning, Teaching and Educational Research*, *20*(1), 308–322. https://doi.org/10.26803/ijlter.20.1.17

Sewagegn, A. A. (2020). Learning Objective and Assessment Linkage: Its Contribution to Meaningful Student Learning. *Universal Journal of Educational Research*, *8*(11), 5044–5052. https://doi.org/10.13189/ujer.2020.081104

Tavakol, M., & Wetzel, A. (2020). Factor Analysis: a means for theory and instrument development in support of construct validity. *International Journal of Medical Education*, *11*, 245. https://doi.org/10.5116/ijme.5f96.0f4a

Torshizi, M. D., & Bahraman, M. (2019). I explain, therefore I learn: Improving students' assessment literacy and deep learning by teaching. *Studies in Educational Evaluation*, *61*, 66–73. https://doi.org/10.1016/j.stueduc.2019.03.002

Triyandana, A., Ibrohim, I., Yanuwiyadi, B., Amin, M., & Hajar, M. U. (2024). Strategies to Enhance Eco-Friendly Culture and Environmental Awareness by Green Curriculum Integration in Indonesian Elementary Science Classroom. *International Electronic Journal of Elementary Education*, *17*(1), 217–232. https://doi.org/10.26822/iejee.2024.374

Utaya, S., & Wafaretta, V. (2021). The vision, mission, and implementation of environmental education of adiwiyata elementary school in Malang City. *IOP Conference Series: Earth and Environmental Science*, *802*(1), 12048. https://doi.org/10.1088/1755-1315/802/1/012048

Vos, P. (2018). "How real people really need mathematics in the real world"—Authenticity in mathematics education. *Education Sciences*, *8*(4), 195. https://doi.org/10.3390/educsci8040195

Wijaya, A., Van den Heuvel-Panhuizen, M., Doorman, M., & Veldhuis, M. (2018). *Opportunity-to-learn to solve context-based mathematics tasks and students' performance in solving these tasks–lessons from Indonesia*. https://doi.org/10.29333/ejmste/93420

Zhang, D., Wang, C., Yuan, T., Li, X., Yang, L., Huang, A., Li, J., Liu, M., Lei, Y., & Sun, L. (2023). Psychometric properties of the Coronavirus Anxiety Scale based on Classical Test Theory (CTT) and Item Response Theory (IRT) models among Chinese front-line healthcare workers. *BMC Psychology*, *11*(1), 224. https://doi.org/10.1186/s40359-023-01251-x

Zhang, J.-L. (2020). The application of human comprehensive development theory and deep learning in innovation education in higher education. *Frontiers in Psychology*, *11*, 1605. https://doi.org/10.3389/fpsyg.2020.01605