Journal of General Education and Humanities
Vol. 5, No. 5, October 2026, pp. 6209 – 6229, https://doi.org/10.58421/gehu.v5i5.1606
ISSN 2963-7147

The Illusion of Analytical Depth: AI-Mediated Responses and Contextual Reasoning Challenges in Early Childhood Teacher Education

Ratna Pangastuti[1], Mukhoiyaroh[2], Rina Insani Setyowati[3], Dedy Setyawan[4]
[1,2]Universitas Islam Negeri Sunan Ampel Surabaya, Jawa Timur, Indonesia
[3]Universitas Nahdlatul Ulama Blitar, Jawa Timur, Indonesia
[4]STKIP Citra Bakti Ngada, Flores, Nusa Tenggara Timur, Indonesia

Article Info

Article history:

Received 2026-05-13
Revised 2026-07-28
Accepted 2026-09-13

Keywords:

AI-mediated learning
Analytical depth
Case-based learning
Contextual reasoning
Generative AI

ABSTRACT

Artificial intelligence (AI) is increasingly used in case-based learning, yet its implications for students' analytical reasoning remain insufficiently examined. Previous studies have emphasized learning efficiency, user perception, and general performance outcomes, while less attention has been given to whether AI-mediated responses reflect independent analytical development or textual alignment with assessment structures. This study examined differences in students' pedagogical analysis across unmediated and AI-mediated response conditions in early childhood teacher education. A one-group pretest-posttest design was employed involving 60 second-semester students in an Early Childhood Islamic Education Study Program. Data were collected through case-based essay tasks on child language development and assessed using an analytical rubric covering problem identification, theoretical use, language development analysis, impact analysis, and pedagogical recommendation. The posttest condition involved AI support and a mandatory prompt closely aligned with the rubric dimensions. Descriptive statistics and the Wilcoxon Signed-Rank Test were used to analyze score differences, while selected responses from cases of non-improvement or decline were examined to interpret possible changes in contextual reasoning. The mean score increased from 14.25 to 16.50 ($Z = -6.280$; $p < .001$; $r = .81$). Although 49 students improved, 7 showed no change, and 4 experienced score declines. These findings indicate that AI-mediated and prompt-structured conditions may support more organized written responses, but they do not necessarily demonstrate independently internalized analytical capacity. Selected response comparisons suggest that technical coherence may coexist with weakened contextual caution. The study highlights the limitation of product-based written assessment and the need for process-oriented assessment.

Corresponding Author:

Ratna Pangastuti
Program Studi Pendidikan Islam Anak Usia Dini, Universitas Islam Negeri Sunan Ampel Surabaya, Jawa Timur, Indonesia
Email: ratnapangastuti@uinsa.ac.id

1. INTRODUCTION

Pedagogical analysis is a central competence in early childhood teacher education because teachers are expected to interpret children's developmental situations before making instructional decisions. In early childhood education, instructional responses should not be based only on visible symptoms, but on the ability to connect those symptoms with possible causes, developmental theory, family context, and appropriate pedagogical action. Previous studies indicate that early childhood educators still face difficulties in translating professional knowledge into reflective responses to children's developmental needs [1], [2]. Research using Facione's critical thinking framework also suggests that early childhood teachers' professional judgment involves interpretation, analysis, evaluation, inference, and reflective decision-making, rather than merely procedural task completion [3]. Similar concerns appear in studies on parenting practices and early childhood educator professionalism, which show persistent gaps between curriculum expectations and educators' ability to read children's situations analytically [4], [5].

This issue is also evident among pre-service early childhood teachers. When students are asked to analyze developmental cases, their responses often stop at identifying the problem and do not consistently develop into connected pedagogical reasoning. Child language development cases are especially demanding because they cannot be understood only through general developmental norms. They require contextual reasoning, defined in this study as the ability to interpret a case by considering the child's developmental characteristics, family environment, communication patterns, cultural and linguistic background, and the professional boundaries of early childhood educators. In such cases, contextual reasoning also requires caution in using diagnostic terminology and sensitivity to the child's home communication environment. Without this form of reasoning, students may produce answers that appear complete but remain detached from the specific situation of the child.

Case-based learning has been widely used to strengthen students' analytical engagement with complex educational situations. Through case analysis, students are expected to move beyond memorized theory and construct responses that connect concepts, evidence, and pedagogical judgment [6], [7]. Evidence from professional education also suggests that case-based learning can support analytical and decision-making skills when students are required to integrate theoretical concepts with concrete problems [8]. From Vygotsky's perspective, this process can function as scaffolding, where assistance supports learners only when it gradually leads them toward more independent thinking [9]. Assistance becomes problematic when it remains present during evaluation and begins to replace, rather than support, the construction of reasoning. A cognitive-developmental perspective also raises a related concern: structurally advanced responses do not necessarily indicate internalized reasoning when the structure is supplied externally rather than constructed by the learner [10].

The increasing use of generative artificial intelligence has changed the nature of assistance in case-based learning. Large Language Model-based tools can generate organized explanations, connect theoretical concepts, and formulate recommendations within seconds. Studies in higher education show that students use AI to increase speed,

confidence, and productivity in academic tasks, although concerns remain regarding originality, dependence, and the weakening of independent thinking [11], [12]. Kasneci et al. [13] argue that generative AI creates both opportunities and risks in education because its outputs can appear fluent and authoritative even when the user has not fully understood the reasoning behind them. In this sense, AI does not merely provide technical support; it can reshape the form and structure of students' responses.

This distinction is important because AI may operate in several different ways during learning. First, AI may function as scaffolding when it helps students organize ideas while still requiring judgment, revision, and justification. Second, AI may become a form of cognitive offloading when students transfer the burden of analysis to the system instead of constructing arguments themselves [14]. Third, AI may produce pseudo-analytical performance, where responses appear systematic because they follow expected academic structures, but the appearance of structure does not necessarily demonstrate internalized reasoning. Fourth, AI may create challenges for contextual reasoning when technically correct terminology replaces careful attention to the child's concrete situation. These possibilities suggest that AI-mediated performance should not be treated automatically as evidence of analytical development.

Existing studies on AI in higher education have largely focused on student perceptions, learning efficiency, technology acceptance, and general academic performance [15], [16], [17]. Some studies suggest that AI can support critical thinking and problem-solving when embedded in structured learning environments [18], [19]. However, less attention has been given to the quality of reasoning contained in AI-mediated written responses, especially in case-based tasks for early childhood teacher education. What remains underexamined is the possibility that AI-mediated responses may satisfy rubric-based indicators of analytical quality while leaving the reasoning process behind those responses empirically uncertain. This gap is important because a written response can meet formal assessment criteria while still failing to demonstrate contextual sensitivity, professional caution, or independent analytical construction.

This study uses several key concepts to frame the problem. Analytical depth refers to the extent to which a response moves beyond surface-level problem identification and connects symptoms, causes, theory, developmental implications, and pedagogical recommendations. In the present study, analytical depth is operationalized through rubric dimensions assessing problem identification, theoretical use, developmental analysis, impact analysis, and pedagogical recommendation. AI-mediated performance refers to the quality of written responses produced under conditions in which students are allowed to use AI and structured prompts. This concept does not refer to independently reproducible competence, but to performance generated within a mediated analytical system. Cognitive disorientation is treated as an interpretive category derived from score decline and response-level comparison, rather than as a directly measured cognitive state. The notion of illusion of analytical depth is informed by the concept of illusion of explanatory depth, in which individuals may believe they understand a phenomenon more deeply than they actually do [20]. In this study, the term is used cautiously to describe the possibility that AI-generated

coherence may create an appearance of analytical depth without sufficient evidence of internalized understanding.

Based on this gap, the present study examines differences in students' pedagogical analysis across unmediated and AI-mediated response conditions in case-based learning. The study does not claim to measure internal cognitive capacity directly. Instead, it examines written analytical performance and interprets response-level changes within the limits of a one-group pretest-posttest design. The following questions guide the research:

1. How does students' analytical performance differ between unmediated and AI-mediated case analysis?
2. How are individual score changes distributed after AI-mediated analysis?
3. What response-level changes are observed in selected cases of non-improvement or score decline after AI-mediated analysis?

The novelty of this study lies in examining the mismatch between rubric-based score improvement and the possible weakening of contextual reasoning in AI-mediated case analysis for early childhood teacher education. Its contribution is not to determine whether AI is beneficial or harmful in general, but to question whether product-based written assessment remains adequate when textual coherence can be partly generated by external systems. By focusing on the tension between improved written performance and uncertainty about the reasoning process behind it, this study contributes to current debates on AI-mediated learning, assessment validity, and pedagogical reasoning in teacher education.

2. METHOD

This study employed a one-group pretest-posttest design without a control group [21]. The design was used to compare students' analytical performance before and after AI-mediated case-based learning within the same participant group. Because no control group was involved, this study does not claim to establish a causal effect of AI on students' reasoning. The comparison is interpreted as a difference between two analytical conditions: an unmediated condition, in which students completed case analysis without digital assistance, and an AI-mediated condition, in which students revised and expanded their responses using generative AI and a mandatory structured prompt.

The study was conducted in the Early Childhood Islamic Education Study Program, Universitas Islam Negeri Sunan Ampel Surabaya, during the even semester of the 2025/2026 academic year. The participants were second-semester students enrolled in the *Child Language and Early Literacy Development* course. The course included case-based learning activities related to children's language development, language acquisition, communication stimulation, and pedagogical responses to language-related developmental cases. The participants consisted of 60 students aged 18 to 20 years. Total sampling was used because all students who met the inclusion criteria were involved in the study [22]. The inclusion criteria were active enrollment in the course, minimum attendance of 80%, participation in both pretest and posttest activities, and written consent to participate in the study. Students who did not complete either the pretest or posttest were excluded from the analysis. The participants had not received prior formal training in analyzing child language development

cases, which allowed the study to observe their written analytical performance before and after structured case-based learning involving AI support.

The research activities were embedded in the regular course sequence from meeting 1 to meeting 9. Each meeting lasted 150 minutes. Meetings 1 to 6 introduced key concepts in child language development, early literacy, language acquisition, and case-based pedagogical analysis. The pretest was administered in meeting 7. In this phase, students analyzed a child language development case independently without AI tools, search engines, digital references, or peer assistance. Meeting 8 focused on AI-mediated revision and comparative analysis using a mandatory structured prompt. Students were trained to use AI output as material for revision and reflection, not as a final answer to be copied. The posttest was administered in meeting 9 using a different case from the pretest but with the same analytical structure. This parallel case structure was used to reduce the likelihood that score changes resulted merely from repeated exposure to the same case.

During the AI-mediated learning activity, students were introduced to generative AI as a tool for revising and strengthening their responses. The AI tools used by the students were mainly ChatGPT, while some students used Gemini. The study did not experimentally standardize the AI model used by each student because the learning activity reflected a natural classroom setting. This condition is acknowledged as a methodological limitation because different AI models may generate different response patterns.

AI use was controlled through a mandatory prompt that students were required to copy exactly. The prompt was used to improve students' initial answers, not to replace the task of analysis entirely. The mandatory prompt was administered in Indonesian, which was the language used by students during the learning activity. To preserve the original intervention while making it accessible to readers, the prompt is presented below in its original wording followed by an English translation.

Original prompt in Indonesian

Perbaiki jawaban berikut untuk setiap nomor soal (1–5) tentang kasus perkembangan bahasa anak dengan cara:

1. Memperjelas masalah utama sesuai konteks kasus
2. Memperkuat penjelasan teori yang relevan (misalnya teori pemerolehan bahasa)
3. Memperdalam analisis berdasarkan kondisi anak
4. Memberikan rekomendasi yang lebih spesifik dan aplikatif
   Pertahankan struktur jawaban per nomor (1–5). Jangan mengubah makna utama dari jawaban asli. Jangan menambahkan informasi di luar konteks kasus.

English translation

Improve the following answers for each item (1–5) on child language development cases by:

1. Clarifying the main problem according to the case context;
2. Strengthening the explanation using relevant theory, such as language acquisition theory;
3. Deepening the analysis based on the child's condition;

4. Providing more specific and applicable recommendations. Maintain the answer structure for each item (1–5). Do not change the main meaning of the original answer. Do not add information beyond the case context.

Students were required to maintain the structure of their answers for questions 1 to 5 and were not allowed to modify the mandatory prompt. The AI-mediated task consisted of three required parts: the original response, the AI-assisted revised response, and the student's comparative analysis. Students were explicitly prohibited from submitting AI output without personal analysis. They were also informed that their answers had to show personal understanding and that they had to be ready to explain their work orally if requested. Therefore, the posttest represented analytical performance under AI-mediated conditions, not independently reproducible analytical ability.

The research instrument consisted of five open-ended essay questions based on child language development cases. Each question corresponded to one component of pedagogical analysis: problem identification, use of theory, language development analysis, impact analysis, and pedagogical recommendation. The instrument was designed to assess the structural quality of students' written case analysis rather than to directly measure internal cognitive processes. Student responses were assessed using an analytical rubric developed for the midterm assessment of the *Child Language and Early Literacy Development* course. The rubric used a four-point scale for each question. A score of 4 indicated a very accurate, complete, and analytical response; a score of 3 indicated an accurate but less in-depth response; a score of 2 indicated a partially accurate or incomplete response; and a score of 1 indicated an inaccurate or very weak response. The total score ranged from 5 to 20.

The rubric emphasized the organization, completeness, relevance, and analytical structure of students' written responses. It was not designed to distinguish whether the reasoning was produced independently or supported by AI-generated output. For this reason, scores were interpreted as indicators of written analytical performance, not as direct evidence of internal analytical capacity.

Content validity was examined through expert judgment using Aiken's V. The instrument was reviewed by two lecturers from the Early Childhood Islamic Education Study Program with expertise in early childhood education and language-literacy learning. The Aiken's V values ranged from 0.78 to 0.92, indicating acceptable content validity. Student responses were assessed by two independent raters using the same analytical rubric. Each rater scored the responses independently, and the final score for each student was calculated as the mean of the two raters' scores. Before scoring, the raters discussed the rubric indicators to establish a shared understanding of the scoring criteria. An inter-rater reliability coefficient was not reported because the retained dataset contained only averaged final scores rather than separate rater-level scores. Therefore, the scoring procedure relied on a shared analytical rubric, independent scoring by two raters, and averaged scores to mitigate individual scoring bias. The absence of a reported inter-rater reliability coefficient is acknowledged as a methodological limitation of this study.

Table 1. Analytical Rubric for Assessing Case-Based Responses

Aspect	Related Question	Score 4	Score 3	Score 2	Score 1
Problem identification	Q1	Identifies the main problem accurately and specifically, such as limited vocabulary or verbal expression difficulties	Identifies the problem accurately but still generally	Identifies part of the problem but incompletely	Fails to identify the problem correctly
Use of theory	Q2	Uses relevant theory accurately, such as Vygotsky or Chomsky, and connects it to the case	Uses theory correctly but with limited depth	Mentions theory but inaccurately or without clear connection to the case	Does not use theory or shows conceptual error
Language development analysis	Q3	Determines the stage of language development accurately and provides logical justification	Determines the stage fairly accurately but with weak justification	Determines the stage inaccurately or provides unclear justification	Fails to determine the developmental stage
Impact analysis	Q4	Explains the impact comprehensively across cognitive, social, and communication aspects	Explains the impact but in a limited way	Mentions impact but inaccurately or superficially	Fails to explain the impact
Pedagogical recommendation	Q5	Provides concrete, relevant, and clearly separated recommendations for teachers and parents	Provides relevant recommendations but they are not specific enough	Provides recommendations that are too general or less relevant	Does not provide appropriate recommendations

Note. *Each aspect was scored on a scale of 1–4. The total score ranged from 5 to 20.*

The study followed ethical principles for research involving human participants. Written informed consent was obtained from all participants before the data were used for research analysis. Students were informed about the purpose of the study, the voluntary nature of their participation, and the use of their written responses for academic research. Consent concerned the use of anonymized assessment responses for research purposes, while academic assessment followed the regular course procedure. Participant identities were anonymized, and all responses were coded to prevent personal identification. The data were used only for academic research purposes. Students were also informed about responsible AI use in the learning activity. They were instructed not to submit AI output without personal analysis and were required to show their understanding through comparison and explanation of the revised response.

Data analysis began with descriptive statistics, including mean, standard deviation, standard error, minimum score, maximum score, median, frequency, and percentage. Gain scores were calculated by subtracting pretest scores from posttest scores. Total scores were grouped into three descriptive performance categories: low (5–11), moderate (12–14), and

high (15–20). This categorization was used only to summarize score distribution and was not treated as an independent measure of cognitive competence.

The Shapiro-Wilk test was used to examine the normality of pretest, posttest, and gain score distributions [23]. Because the pretest and posttest scores did not meet the assumption of normality, the Wilcoxon Signed-Rank Test was used to examine differences between the unmediated and AI-mediated conditions. The effect size for the Wilcoxon Signed-Rank Test was calculated using the formula $r = |Z| / \sqrt{N}$. Based on $Z = -6.280$ and $N = 60$, the effect size was $r = .81$, indicating a large difference between the two analytical conditions.

In addition to statistical analysis, selected student responses were compared to examine response-level changes after AI-mediated analysis. The excerpts were selected purposively from students who showed no improvement or experienced score decline because these cases provided relevant evidence for examining possible changes in contextual reasoning. This response-level comparison was not treated as a direct measurement of cognitive processes, but as interpretive evidence to support the discussion of AI-mediated analytical performance.

3. RESULTS AND DISCUSSION

This section presents the findings obtained from descriptive statistics, score distribution, normality testing, the Wilcoxon Signed-Rank Test, effect size calculation, and individual gain score distribution. The results are presented first to establish the empirical pattern of score change before the findings are interpreted in relation to AI-mediated analytical performance and contextual reasoning.

3.1 Changes in Analytical Performance Across Two System Conditions

Student analytical performance showed a clear difference between the unmediated condition and the AI-mediated condition. In the pretest, students completed the case analysis without AI assistance, digital references, search engines, or peer support. This condition was intended to capture their written analytical performance based on the cognitive and conceptual resources available to them at that point. The mean pretest score was 14.25 (SD = 2.01), with scores ranging from 9 to 18. The median score was 14.0, indicating that many students were positioned around the moderate-to-high boundary before AI mediation was introduced.

In the posttest, students completed a parallel case-based task with access to AI and the mandatory structured prompt. The mean posttest score increased to 16.50 (SD = 1.99), with scores ranging from 12 to 19. The median also increased from 14.0 to 17.0. The mean gain score was 2.25 (SD = 1.95), with individual gain scores ranging from -3 to +6. This pattern indicates a measurable increase in written analytical performance under the AI-mediated condition.

Table 2. Descriptive Statistics of Pretest and Posttest Scores

Variable	N	Mean	SD	SE	Min	Max	Median
Pretest	60	14.250	2.014	0.260	9	18	14.0
Posttest	60	16.500	1.987	0.257	12	19	17.0
Gain (Post-Pre)	60	2.250	1.945	0.251	-3	6	2.0

Note. SD = Standard Deviation; SE = Standard Error of the Mean. Gain = the difference between posttest and pretest scores for each respondent. Negative gain values indicate a score decline in 4 respondents.

The increase in mean and median scores suggests that students' written responses became more aligned with the structural demands of the analytical task after AI mediation. However, this increase must be interpreted carefully. The pretest and posttest did not occur under identical analytical conditions. The pretest required independent response construction, whereas the posttest allowed students to revise and strengthen their responses through AI support and a structured prompt. Therefore, the observed difference should be understood as a difference in written analytical performance across two system conditions, not as direct evidence of independently developed analytical capacity. This interpretation is consistent with the limitation of a one-group pretest-posttest design, which does not allow strong causal attribution without a control group [21].

The distribution of scores also changed across performance categories. In the pretest, 6 students (10.00%) were in the low category, 25 students (41.67%) were in the moderate category, and 29 students (48.33%) were in the high category. This distribution shows that almost half of the students had already produced responses that met the high category before AI mediation, while a substantial proportion remained in the moderate category.

Table 3. Frequency Distribution of Pretest Scores

Total Score	f	%	Cumulative %	Category
9	2	3.33	3.33	Low
10	1	1.67	5.00	Low
11	3	5.00	10.00	Low
12	4	6.67	16.67	Moderate
13	8	13.33	30.00	Moderate
14	13	21.67	51.67	Moderate
15	13	21.67	73.33	High
16	9	15.00	88.33	High
17	5	8.33	96.67	High
18	2	3.33	100.00	High
Total	60	100.00		

Note. Categories: Low = 5-11, Moderate = 12-14, High = 15-20.

In the posttest, no students remained in the low category. Eight students (13.33%) were in the moderate category, while 52 students (86.67%) were in the high category. The shift from 48.33% to 86.67% in the high category indicates that more students produced

responses that fulfilled the higher structural criteria of the analytical rubric under AI-mediated conditions.

Table 4. Frequency Distribution of Posttest Scores

Total Score	f	%	Cumulative %	Category
12	4	6.67	6.67	Moderate
14	4	6.67	13.33	Moderate
15	11	18.33	31.67	High
16	10	16.67	48.33	High
17	9	15.00	63.33	High
18	10	16.67	80.00	High
19	12	20.00	100.00	High
Total	60	100.00		

Note. Categories: Low = 5–11; Moderate = 12–14; High = 15–20.

The category shift shows that AI-mediated conditions were associated with stronger written performance according to the rubric. Nevertheless, because the rubric assessed the organization, completeness, relevance, and analytical structure of written responses, high scores should not automatically be equated with internalized reasoning. The descriptive findings therefore provide the first indication of the central tension examined in this study: students' written products improved structurally, but the extent to which this improvement reflected independent reasoning requires further interpretation.

3.2 Statistical Difference Between Unmediated and AI-Mediated Conditions

Before testing score differences, the normality of the pretest, posttest, and gain score distributions was examined using the Shapiro-Wilk test. The pretest scores were not normally distributed ($W = 0.9546$, $p = .026$), and the posttest scores were also not normally distributed ($W = 0.9161$, $p = .001$). The gain scores were normally distributed ($W = 0.9700$, $p = .145$). Although the gain scores met the normality assumption, the Wilcoxon Signed-Rank Test was selected as the main inferential test because the pretest and posttest distributions were non-normal and the scores were derived from an ordinal rubric-based assessment. This decision was also consistent with the need to use a conservative non-parametric approach for paired score comparison [23].

Table 5. Normality Test Results (Shapiro-Wilk)

Variable	*N*	Statistic (W)	*Sig.* (*p*)	Conclusion
Pretest	60	0.9546	.026*	Non-normal
Posttest	60	0.9161	.001**	Non-normal
Gain (Post-Pre)	60	0.9700	.145	Normal

*Note. * $p < .05$; ** $p < .01$. Pretest ($p = .026$) and posttest ($p = .001$) distributions are non-normal. Gain distribution is normal ($p = .145$). Due to the non-normal distributions of pretest and posttest scores, the Wilcoxon Signed-Rank Test was utilized.*

The Wilcoxon Signed-Rank Test showed a statistically significant difference between pretest and posttest scores ($Z = -6.280$, $p < .001$). The rank distribution showed 49 positive ranks, 4 negative ranks, and 7 ties. This means that most students obtained higher posttest scores, while a smaller number showed no score change or experienced a score decline.

Table 6. Wilcoxon Signed-Rank Test Results

Category	*N*	*Mean Rank*	*Sum of Ranks*
Negative Ranks	4	2.50	10.00
Positive Ranks	49	25.00	1225.00
Ties	7		
Total	60		

Note. Negative ranks = posttest < pretest; positive ranks = posttest > pretest; ties = no change.

Table 7. *Wilcoxon Signed-Rank Test: Test Statistics*

Statistic	Value
Z	-6.280
Asymp. Sig. (2-*tailed*)	< .001

Note. Based on negative ranks.

The effect size was calculated using the Wilcoxon-compatible formula $r = |Z| / \sqrt{N}$. With $Z = -6.280$ and $N = 60$, the effect size was $r = .81$. Using commonly applied benchmarks for non-parametric effect size interpretation, this value indicates a large difference between the two analytical conditions [24]. However, the effect size should be interpreted as the magnitude of performance difference between the unmediated and AI-mediated conditions, not as proof of causal effectiveness. The design does not isolate AI from other contextual factors, and the posttest condition included both AI access and structured prompting.

Table 8. Effect Size for the Wilcoxon Signed-Rank Test

Parameter	**Value**
Z	-6.280
N	60
Effect size formula	$r = \|Z\| / \sqrt{N}$
Effect size r	.81
Interpretation	Large

Note. The effect size represents the magnitude of difference between the unmediated and AI-mediated analytical conditions. It should not be interpreted as evidence of a causal effect because the study used a one-group pretest-posttest design without a control group.

These inferential results confirm that the difference between pretest and posttest scores was statistically detectable. The result supports the descriptive pattern showing improved written analytical performance under AI-mediated conditions. Yet the statistical finding remains limited to performance differences. It does not establish that students developed independently reproducible analytical capacity. This distinction is crucial because the posttest condition involved external mediation through AI and structured prompts, while the pretest condition did not.

3.3 Distribution of Score Changes Across Participants

The distribution of individual gain scores showed that score changes under the AI-mediated condition were not uniform across participants. Out of 60 students, 49 students (81.67%) showed score increases, 7 students (11.67%) showed no change, and 4 students (6.67%) experienced score declines. Gain scores ranged from -3 to +6, with the highest frequency occurring at a gain of +3. This distribution demonstrates that a majority pattern of improvement drove the overall increase in mean score, but it also reveals a smaller group of students for whom AI-mediated analysis did not produce higher written performance.

Table 9. Distribution of Gain Scores

Gain	f	%	Cumulative %	Description
-3	1	1.67	1.67	Decrease
-2	1	1.67	3.33	Decrease
-1	2	3.33	6.67	Decrease
0	7	11.67	18.33	No change
1	10	16.67	35.00	Increase
2	10	16.67	51.67	Increase
3	14	23.33	75.00	Increase
4	8	13.33	88.33	Increase
5	4	6.67	95.00	Increase
6	3	5.00	100.00	Increase
Total	60	100.00		

Note. Positive ranks (increase): n = 49 (81.67%). Negative ranks (decrease): n = 4 (6.67%). Ties (no change): n = 7 (11.67%).

The gain distribution is important because it prevents the findings from being reduced to a simple narrative of improvement. Although most students improved, 18.33% of the participants either showed no change or experienced a decline. This pattern suggests that AI-mediated conditions do not produce uniform outcomes across students. Some students may be able to use AI and structured prompts to improve the organization and completeness of their responses, while others may not benefit in the same way.

The seven unchanged scores indicate that access to AI did not automatically raise students' written analytical performance. These cases suggest that AI support may require a certain level of readiness, interpretive control, or ability to revise system-generated output. Without these capacities, AI may remain a surface-level aid that does not substantially change the quality of the submitted response.

The four declining scores require careful interpretation. A score decline under the AI-mediated condition does not, by itself, prove cognitive disorientation or deterioration of reasoning. Such a conclusion would require stronger process-based evidence, such as think-aloud protocols, interviews, prompt logs, or revision histories. However, these declining cases are analytically important because they identify responses that warrant closer examination. When read alongside selected response-level comparisons, score decline may

suggest a possible tension between AI-mediated textual restructuring and students' contextual reasoning.

Therefore, the gain distribution provides two important findings. First, AI-mediated analysis was associated with improved written performance for most students. Second, the non-uniform pattern of score change shows that AI mediation did not operate as a universally beneficial condition. The subsequent sections interpret this pattern more cautiously by distinguishing what the data directly show from what can only be inferred theoretically. The data directly show score improvement and uneven individual change; the interpretation that these changes may reflect an illusion of analytical depth remains a theoretical inference drawn from the mismatch between score improvement, AI-mediated response structure, and selected response-level comparisons.

3.4 Illusion of Analytical Depth in AI-Mediated Written Performance

The score increase observed in this study should not be interpreted only as a technical improvement in student performance. It also raises a deeper assessment issue: whether structurally improved responses necessarily indicate deeper analytical reasoning. The findings show that many students produced higher-scoring responses under AI-mediated conditions. However, because the posttest involved both AI support and a mandatory structured prompt, the improved written products cannot be separated from the external system that helped shape them. This distinction is important because generative AI can produce fluent, coherent, and authoritative responses even when the user's own understanding remains uncertain [13].

The analytical rubric assessed visible features of written responses, including problem identification, theoretical use, developmental analysis, impact analysis, and pedagogical recommendation. These features are necessary for evaluating case-based responses, but they do not fully reveal how the reasoning was constructed. A response may appear coherent, complete, and theoretically informed because it follows the expected structure of the rubric, while the student's independent understanding of the case remains empirically uncertain. This creates a possible mismatch between rubric-based performance and internalized analytical capacity.

This mismatch can be interpreted through the concept of the illusion of explanatory depth. Rozenblit and Keil [20] argue that individuals may believe they understand a phenomenon more deeply than they actually do. The issue in this study is not simply whether students overestimated their understanding. Rather, the AI-mediated condition may have produced written responses that looked analytically deeper than the reasoning process that could be verified from the available data. The "illusion" therefore lies in the assessment situation: textual coherence can make analytical depth appear more stable than it actually is.

The mandatory prompt used in the AI-mediated condition was not a neutral instruction. It explicitly asked students to clarify the main problem according to the case context, strengthen the explanation with relevant theory, deepen the analysis based on the child's condition, and provide more specific and applicable recommendations. These four instructions closely paralleled the rubric dimensions used to assess students' responses, particularly problem identification, theoretical use, developmental analysis, and pedagogical

recommendation. The prompt also instructed students to maintain the structure of answers for items 1–5, not to change the main meaning of the original answer, and not to add information beyond the case context. Therefore, the posttest responses were not only AI-mediated but also prompt-structured.

This AI-prompt-rubric alignment is central to interpreting the score increase. The prompt functioned as a response architecture that directed students toward the kinds of elements most likely to be rewarded by the scoring rubric. As a result, the improved posttest scores may reflect students' ability to work within a preformatted analytical pathway, rather than independently constructing the full structure of the analysis. This does not make the improvement meaningless. It means that the improvement should be interpreted as mediated written performance, not as direct evidence of internalized analytical capacity.

From a Vygotskian perspective, external assistance can support learning when it functions as scaffolding that gradually enables more independent performance [9]. The relevant issue in this study is that assistance was still present during the posttest condition. Therefore, the posttest scores represent mediated performance rather than unassisted competence. This distinction is crucial. AI may have helped students produce more organized and complete responses, but the study design does not demonstrate whether students could reproduce the same level of reasoning without AI support.

The findings can also be read in relation to cognitive offloading. Sweller's [14] cognitive load theory suggests that instructional support can reduce unnecessary cognitive burden, but excessive reliance on external support may also reduce the learner's need to organize and evaluate information independently. In AI-mediated writing, this risk becomes more pronounced because the system can supply not only information but also structure, terminology, argument flow, and recommendation patterns. Under such conditions, students may appear to engage in complex reasoning while part of the analytical work has been transferred to the system.

A Piagetian or neo-Piagetian interpretation may also be useful, but it must be used cautiously. Cognitive development theories emphasize that mature reasoning involves active construction, adaptation, and reorganization of thought rather than the passive acceptance of externally supplied structures [10], [25]. The data in this study do not prove that students bypassed cognitive restructuring. They only raise the possibility that AI-supported and prompt-structured responses may simulate advanced analytical organization without sufficient evidence that such organization has been internalized.

This interpretation is consistent with concerns in the AI education literature. Holmes et al. [26] warn that AI in education requires careful ethical and pedagogical control because system-generated outputs can reshape learning processes. Zawacki-Richter et al. [17] show that AI research in higher education has often emphasized technological application while giving less attention to educators' roles and pedagogical judgment. Brazão and Tinoca [27] similarly suggest that critical thinking through educational chatbots does not develop automatically, but depends on the quality of task design, mediation, and learner engagement. In the Indonesian higher education context, Rahiem [28] also reports that generative AI use may affect learning practices in ways that require attention to students' patterns of use and academic responsibility.

This interpretation does not imply that AI is inherently harmful to pedagogical reasoning. AI can support learning when students are required to question, revise, justify, and contextualize its output. The concern raised by this study is narrower and more specific: product-based written assessment may overestimate analytical development when AI contributes substantially to the structure and language of the submitted response. The issue is not AI use itself, but the validity of interpreting AI-mediated written products as direct evidence of independent reasoning.

Therefore, the score increase in this study is best understood as evidence of improved AI-mediated written analytical performance. It suggests that students were better able to produce responses aligned with the expected analytical structure under AI-supported and prompt-structured conditions. However, it does not provide sufficient evidence that their internal reasoning developed proportionally. This is the central meaning of the illusion of analytical depth in this study: the written product becomes more convincing, but the reasoning process behind it remains only partially visible.

3.5 Contextual Reasoning Challenges in Selected AI-Mediated Responses

The uneven distribution of score changes, especially the cases of non-improvement and decline, requires closer qualitative interpretation. The following excerpts were selected purposively from responses showing no improvement or score decline because these cases provided clearer evidence of possible tension between AI-mediated textual improvement and contextual reasoning. These examples are not treated as representative of all students. They are used as illustrative cases to examine how AI-mediated revision may alter the relationship between case evidence, terminology, and pedagogical judgment.

A translated excerpt from one student's response before AI mediation stated:

"The child experiences difficulty in recounting simple experiences, and the parents rarely engage the child in communication at home."

This answer was not terminologically sophisticated, but it maintained a direct relationship between the child's observable difficulty and the family communication context. The response identified the problem in ordinary language and connected it to the child's home environment. It remained within the professional boundaries of a pre-service early childhood teacher because it did not impose a diagnostic label beyond the information available in the case.

After AI mediation, the translated revised response stated:

"The primary issue is expressive language delay (speech delay) resulting from a lack of linguistic stimulation and social interaction."

The AI-mediated version appears more academic because it uses technical terminology. However, its contextual caution is weaker. The term speech delay is not inherently incorrect, but it becomes problematic when used as a diagnostic-like conclusion based on a limited case narrative. A pre-service early childhood teacher may observe and describe language difficulties, but assigning a diagnostic category requires more comprehensive assessment and professional authority. In this example, terminological sophistication appears to replace contextual restraint.

A similar issue appeared in a selected recommendation excerpt. A translated excerpt from the initial response stated:

"Parents should frequently engage the child in communication and ask about daily activities."

The translated AI-mediated revision stated:

"Parents need to establish a language-rich environment through self-talk and parallel talk to enhance the child's linguistic competence."

The revised recommendation is more technical and appears more structured. Yet the terms self-talk and parallel talk are presented as general solutions without sufficient explanation of how they fit the child's actual family situation. These strategies are not pedagogically wrong. They become weak when they are inserted as universal recommendations without considering parents' communication habits, time availability, cultural-linguistic background, or the practical conditions in which home-based stimulation would occur.

These examples show why contextual reasoning is central in early childhood teacher education, where teachers are expected to interpret children's developmental situations with professional caution and pedagogical sensitivity [2], [5]. In child language development cases, contextual reasoning requires students to connect symptoms with developmental variation, family communication patterns, cultural and linguistic background, and the limits of teachers' professional authority. A response that uses technical language but fails to justify its relevance to the child's concrete situation may appear more advanced while becoming less context-sensitive. This problem is consistent with broader concerns that AI-generated educational responses can appear authoritative even when their contextual grounding remains weak [13], [26].

The selected examples also indicate a possible form of AI-mediated authority. AI-generated language often carries a tone of confidence, coherence, and technical precision. Students who do not critically evaluate this output may accept the terminology as automatically superior to their original reasoning. Under such conditions, AI-mediated revision may shift students' responses from context-bound analysis toward generalized professional language when students do not critically evaluate the generated output. This shift does not necessarily lower the score, because rubrics often reward completeness, theoretical terminology, and formal coherence. However, it may weaken the contextual justification that should guide pedagogical decision-making.

The mandatory prompt may have intensified this possibility because it directed students to strengthen theory, deepen analysis, and make recommendations more specific. These are legitimate instructional goals. The problem is that AI may fulfill these instructions by adding technical terminology and generalized pedagogical strategies rather than by deepening the student's contextual interpretation of the case. In other words, the prompt encouraged analytical expansion, but the quality of that expansion depended on whether students could evaluate the relevance of AI-generated additions. This is why AI-supported critical thinking cannot be assumed to emerge from chatbot interaction alone; it requires structured evaluation, pedagogical mediation, and learner responsibility [27], [28].

This pattern should be interpreted cautiously. The study does not directly measure students' cognitive processes, and it cannot determine whether students internally misunderstood the case. What the response comparison shows is more limited: in selected cases, AI-mediated revision changed the form of the answer in ways that improved academic appearance but reduced contextual caution. This supports the broader concern that AI-mediated responses may produce pseudo-analytical performance, where the response appears conceptually mature but the underlying reasoning remains insufficiently grounded in the case.

The findings therefore complicate a simple positive interpretation of AI-mediated score improvement. Most students improved in score, but selected response-level evidence suggests that higher structural quality does not always mean stronger contextual reasoning. In early childhood education, this distinction matters because pedagogical analysis is not only about using correct theory or terminology. It requires careful judgment about what can be inferred from the case, what remains uncertain, and what kinds of recommendations are professionally appropriate.

3.6 Limitations and Implications for AI-Mediated Assessment

Several limitations must be acknowledged. First, this study used a one-group pretest-posttest design without a control group. As a result, the findings cannot establish a causal effect of AI on students' analytical reasoning. The observed score difference reflects performance across two different analytical conditions. However, the design cannot isolate AI from other possible influences, including repeated exposure to case-based tasks, course learning, familiarity with rubric expectations, and the use of a structured prompt. This limitation is consistent with the general caution required in interpreting pre-experimental designs [21].

Second, the study did not include a delayed posttest. Therefore, it cannot determine whether the improved posttest performance would be maintained after AI support was removed. This is a substantive limitation because the central issue in AI-mediated learning is not only whether students can produce better responses with AI, but whether they can internalize the reasoning process and reproduce it independently. From a scaffolding perspective, this distinction matters because support should ideally lead toward more independent performance rather than remain embedded in the assessment condition [9].

Third, the study assessed written products rather than the full reasoning process. No think-aloud protocol, interview, screen recording, prompt log, or revision history was systematically analyzed. Because of this, the study cannot directly verify how students interacted with AI, how they evaluated AI output, or how much of the final response reflected their own reasoning. The discussion of possible cognitive disorientation and the illusion of analytical depth must therefore be understood as theoretical interpretation, not direct cognitive measurement.

Fourth, the AI tools used by students were not fully standardized. Most students used ChatGPT, while some used Gemini. Different AI systems may generate different types of output, especially in terminology, explanation structure, and recommendation style. This variation limits the precision of claims about AI-mediated response construction. It also

supports the broader concern that educational AI use must be interpreted in relation to specific tools, contexts, and pedagogical controls rather than treated as a single uniform intervention [17], [26].

Fifth, another limitation concerns the role of the mandatory prompt. Because the prompt was intentionally structured around the same components assessed in the rubric, the study cannot determine whether the score increase was produced by AI capability, prompt structure, rubric alignment, or the interaction among these elements. The posttest condition should therefore be understood as an AI-prompt-rubric configuration rather than as AI use alone.

Sixth, the study does not report an inter-rater reliability coefficient because the retained dataset contained only averaged final scores rather than separate rater-level scores. Although two raters assessed the responses independently using the same rubric, the absence of a reported reliability coefficient limits the strength of the scoring evidence. Future studies should retain rater-level data and report inter-rater reliability to strengthen the credibility of rubric-based assessment.

Seventh, the response-level comparisons were selected purposively from cases of non-improvement and score decline. These excerpts are useful for illustrating possible contextual reasoning challenges, but they do not constitute a comprehensive qualitative analysis of all student responses. Future research should examine response patterns across all participants using systematic qualitative coding.

Despite these limitations, the findings have important implications for assessment design in AI-mediated learning. Written products alone may be insufficient as evidence of students' independent analytical capacity when AI is allowed during task completion. In case-based teacher education, assessment should include process-oriented components that require students to explain how they used AI, why they accepted or rejected AI suggestions, and how they connected the final response to the specific case context. This implication is consistent with recent arguments that responsible AI use in academic tasks requires transparency, critical reflection, and a clear distinction between human and system contributions [29].

Several assessment strategies are therefore recommended. Students should be required to submit their original response, AI-generated revision, and a reflective comparison explaining what changed and why. Prompt critique should also be included so that students evaluate whether the AI output is contextually appropriate, theoretically accurate, and professionally cautious. Oral defense can be used to test whether students understand the reasoning behind their submitted answers. Reflection logs and prompt records can help lecturers distinguish between AI-assisted organization and independent analytical construction.

For early childhood teacher education, these implications are especially important. Future teachers must learn not only to produce well-structured answers, but also to make careful judgments about children's developmental situations. AI can support this process when used as a tool for reflection and revision. It becomes problematic when its fluent output is treated as a substitute for contextual reasoning. Therefore, AI-mediated assessment should

not ask only whether a response looks analytical. It must also ask whether students can justify, contextualize, and defend the reasoning behind it.

4. CONCLUSION

This study examined differences in students' pedagogical analysis between unmediated and AI-mediated case-based response conditions in early childhood teacher education. The findings showed a statistically significant increase in written analytical performance, with the mean score rising from 14.25 in the pretest to 16.50 in the posttest. The Wilcoxon Signed-Rank Test indicated a significant difference between the two conditions, with a large effect size. At the individual level, 49 students improved, 7 showed no change, and 4 experienced score declines.

These findings indicate that AI-mediated and prompt-structured conditions can support the production of more organized and complete written responses. However, the score increase should not be interpreted as direct evidence of independently internalized analytical capacity. The posttest responses were produced under conditions involving AI support and a mandatory prompt that closely aligned with the scoring rubric. Therefore, the improvement is best understood as enhanced AI-mediated written performance rather than as proof of autonomous reasoning development.

The study also shows that higher structural quality does not always guarantee stronger contextual reasoning. Selected response-level comparisons suggest that AI-mediated revisions may introduce more technical terminology and formal coherence while weakening contextual caution in some cases. In early childhood language development analysis, this distinction is important because pedagogical judgment requires sensitivity to the child's developmental condition, family communication environment, cultural-linguistic context, and the professional boundaries of early childhood educators.

The study is limited by its one-group pretest-posttest design, the absence of a control group, the lack of delayed posttest data, the absence of direct process-based evidence such as prompt logs or interviews, the use of non-standardized AI tools, and the lack of a reported inter-rater reliability coefficient. The response-level excerpts were also selected purposively and should not be treated as representative of all students' responses.

Future research should use controlled or comparative designs, include delayed posttests, retain rater-level data, and examine AI interaction processes through prompt logs, revision histories, interviews, or oral defenses. For teacher education assessment, written products should be complemented with process-based evidence, including original responses, AI-revised responses, reflective comparison, prompt critique, and justification of pedagogical decisions. AI can support case-based learning, but its use in assessment requires mechanisms that make students' reasoning process visible, not only the final written product.

REFERENCES

[1] S. A. Wati, "Teachers' academic qualifications and their levels of proficiency on the essential teaching skills: A survey study among teachers of early childhood education programs," *mgs*, vol. 51, no. 1, pp. 75–85, May 2023, doi: 10.33508/mgs.v51i1.4234.

[2] E. Pollarolo, I. Størksen, T. H. Skarstein, and N. Kucirkova, "Children's critical thinking skills: perceptions of Norwegian early childhood educators," *European Early Childhood Education Research Journal*, vol. 31, no. 2, pp. 259–271, Mar. 2023, doi: 10.1080/1350293X.2022.2081349.

[3] R. Rohita, E. Yetti, and T. Sumadi, "Exploring Teachers' Views on Early Childhood Critical Thinking: Insights from Facione's Framework," *The International Journal of Learning in Higher Education*, Feb. 2026, doi: 10.18848/2327-7955/CGP/A323.

[4] A. S. Fatimaningrum, "Analisis Tematik Permasalahan dalam Praktik Pengasuhan Anak yang Dilakukan oleh Guru Pendidikan Anak Usia Dini," *Diklus: Jurnal Pendidikan Luar Sekolah*, vol. 5, no. 2, pp. 128–144, Sep. 2021, doi: 10.21831/diklus.v5i2.43030.

[5] I. Maria, S. Hartati, and N. Dhieni, "Analisis Pengembangan Profesional Pendidik Anak Usia Dini; Tren, Penelitian dan Praktik," *Jurnal Obsesi: Jurnal Pendidikan Anak Usia Dini*, vol. 7, no. 4, pp. 4276–4286, Aug. 2023, doi: 10.31004/obsesi.v7i4.4567.

[6] B. Bungatang, F. Filawati, and N. Fitrayani, "Pengaruh Metode Case Method terhadap Hasil Belajar Mahasiswa Bahasa dan Sastra Indonesia pada Materi Pemerolehan Bahasa Anak Universitas Negeri Makassar," *Jurnal Onoma: Pendidikan, Bahasa, dan Sastra*, vol. 10, no. 3, pp. 2409–2414, Jun. 2024, doi: 10.30605/onoma.v10i3.3839.

[7] E. Silitubun, M. Mustaji, and U. Dewi, "The effectiveness of case-based learning in enhancing students' critical thinking skills," *Journal of Neonatal Surgery*, vol. 14, no. 2, pp. 136–141, Feb. 2025, doi: 10.52783/jns.v14.1844.

[8] D. P. Wijaya, T. Nurmasitoh, and M. D. Pramaningtyas, "The effectiveness of case-based learning for physiology practicum session," *JPKI*, vol. 12, no. 3, p. 261, Oct. 2023, doi: 10.22146/jpki.46154.

[9] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press, 1978.

[10] M. M. Y. Aseeri, "Abstract Thinking of Practicum Students at Najran University in Light of Piaget's Theory and Its Relation to Their Academic Level," *Journal of Curriculum and Teaching*, vol. 9, no. 1, pp. 63–72, Feb. 2020, doi: 10.5430/jct.v9n1p63.

[11] N. K. Pratiwi, B. Yulianto, M. Mintowati, H. Supratno, S. Sodiq, and M. Mulyono, "Persepsi Mahasiswa terhadap Penggunaan ChatGPT: Peluang dan Tantangan bagi Pembelajaran Bahasa Indonesia sebagai Mata Kuliah Wajib pada Kurikulum Perguruan Tinggi," *Jurnal Onoma: Pendidikan, Bahasa, dan Sastra*, vol. 10, no. 3, pp. 2727–2742, Jun. 2024, doi: 10.30605/onoma.v10i3.3931.

[12] S. Rifky, "Dampak Penggunaan Artificial Intelligence Bagi Pendidikan Tinggi," *Indonesian Journal of Multidisciplinary Science and Technology*, vol. 2, no. 1, pp. 37–42, Feb. 2024, doi: 10.31004/ijmst.v2i1.287.

[13] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, Apr. 2023, doi: 10.1016/j.lindif.2023.102274.

[14] J. Sweller, "Cognitive Load During Problem Solving: Effects on Learning," *Cognitive Science*, vol. 12, no. 2, pp. 257–285, Apr. 1988, doi: 10.1207/s15516709cog1202_4.

[15] G. Pitts, V. Marcus, and S. Motamedi, "Student perspectives on the benefits and risks of AI in education," in *2025 ASEE Annual Conference & Exposition Proceedings*, Montreal, Quebec, Canada: ASEE Conferences, Jun. 2025, p. 57693, doi: 10.18260/1-2--57693.

[16] S. Sathish Kumar, D. Kesavan, N. Marianand, and X. Naveen Raj, "Artificial intelligence and students learning: A study on outcome of technological usage (AI) in higher education platform," in *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India: IEEE, Mar. 2023, pp. 2077–2083, doi: 10.1109/ICACCS57279.2023.10113027.

[17] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education: Where are the educators?," *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, p. 39, Dec. 2019, doi: 10.1186/s41239-019-0171-0.

[18] I. Aprianto, S. Sofyan, S. Rahmawati, and S. Sufyadi, "Artificial intelligence and its effects on critical thinking and problem-solving abilities in higher education," *Indonesian Journal of Educational Development*, vol. 6, no. 3, pp. 704–719, Nov. 2025, doi: 10.59672/ijed.v6i3.5030.

[19] A. Rahmawati, S. N. Amirah, and N. Wijaya, "Integrasi Kecerdasan Buatan dalam Pendidikan Tinggi Indonesia: Peluang, Tantangan, dan Kerangka Implementasi," *Jurnal Teknologi Sistem Informasi*, vol. 6, no. 1, pp. 114–126, Apr. 2025, doi: 10.35957/jtsi.v6i1.11329.

[20] L. Rozenblit and F. Keil, "The misunderstood limits of folk science: an illusion of explanatory depth," *Cognitive Science*, vol. 26, no. 5, pp. 521–562, Sep. 2002, doi: 10.1207/s15516709cog2605_1.

[21] J. W. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Thousand Oaks, CA: SAGE Publications, 2018.

[22] J. R. Fraenkel, N. E. Wallen, and H. H. Hyun, *How to design and evaluate research in education*, 8th ed. New York: McGraw-Hill Humanities/Social Sciences/Languages, 2012.

[23] J. Pallant, *SPSS survival manual: A step by step guide to data analysis using IBM SPSS*, 7th ed. London: McGraw-Hill Education, 2020.
[24] A. Field, *Discovering Statistics Using IBM SPSS Statistics*, 5th ed. London: SAGE Publications, 2018.
[25] A. Demetriou, M. Shayer, and A. Efklides, Eds., *Neo-Piagetian Theories of Cognitive Development: Implications and Applications for Education*. London: Routledge, 2016.
[26] W. Holmes et al., "Ethics of AI in Education: Towards a Community-Wide Framework," *International Journal of Artificial Intelligence in Education*, vol. 32, no. 3, pp. 504–526, Sep. 2022, doi: 10.1007/s40593-021-00239-1.
[27] P. Brazão and L. Tinoca, "Artificial intelligence and critical thinking: A case study with educational chatbots," *Frontiers in Education*, vol. 10, p. 1630493, Sep. 2025, doi: 10.3389/feduc.2025.1630493.
[28] M. D. H. Rahiem, "Generative AI in higher education in Indonesia: Patterns of use and learning impact," *Social Sciences & Humanities Open*, vol. 13, p. 102672, Jun. 2026, doi: 10.1016/j.ssaho.2026.102672.
[29] M. Dgebuadze and K. Kenkebashvili, "Integrating artificial intelligence in academic writing and research methods courses for master students: A teaching model for responsible use," *International E-Journal of Advances in Education*, 2025, doi: 10.5281/zenodo.14719550.