

Analysis of Creative Thinking Skills Test Instrument Items on Water Cycle Material in Elementary Schools

¹⁾Adelia Dewanti, ²⁾Sri Yamtinah, ³⁾Sukarno

^{1,2,3)} Master Program in Elementary School Teacher Education, Universitas Sebelas Maret

*Correspondence Author: adelia942@student.uns.ac.id

Article Info

Keywords:

Test instrument;
creative thinking skills;
water cycle;
item analysis;
construct validity.

Informasi Artikel

Kata Kunci:

Instrumen tes;
keterampilan berpikir kreatif;
siklus air;
analisis butir;
validitas konstruk.

Article History

ABSTRACT

This study analyzed the quality of a test instrument designed to measure fifth-grade elementary school students' creative-thinking skills on the water cycle. A quantitative descriptive design was applied based on classical test theory. The instrument contained 15 multiple-choice items and 4 essay items representing fluency, flexibility, originality, and elaboration. Six validators, consisting of three lecturers and three elementary school teachers, assessed content validity using Aiken's V. The instrument was administered to 103 fifth-grade students. Item characteristics were examined in terms of content validity, reliability, discrimination, difficulty, and distractor functioning using ITEMAN 4.4 and ANATES V4. The results showed that most items met the content-validity criterion, with Aiken's V values generally at or above 0.78. Cronbach's Alpha was 0.838 for the pretest and 0.817 for the posttest, while ANATES produced reliability coefficients of 0.83 and 0.84. Most items showed moderate-to-good discrimination, while item difficulty ranged from easy to difficult. Distractor functioning was generally better in the posttest. The findings provide preliminary evidence of instrument quality at the initial development stage. However, construct validity was not examined using exploratory or confirmatory factor analysis. Therefore, further validation using larger and more diverse samples and construct-validity analysis is recommended.

ABSTRAK

Penelitian ini menganalisis kualitas instrumen tes untuk mengukur keterampilan berpikir kreatif murid kelas V sekolah dasar pada materi siklus air. Penelitian menggunakan desain deskriptif kuantitatif berdasarkan teori tes klasik. Instrumen terdiri atas 15 soal pilihan ganda dan 4 soal esai yang merepresentasikan indikator kelancaran, keluwesan, keaslian, dan elaborasi. Validitas isi dinilai oleh enam validator yang terdiri atas tiga dosen dan tiga guru sekolah dasar menggunakan Aiken's V. Instrumen diujicobakan kepada 103 murid kelas V. Analisis karakteristik butir meliputi validitas isi, reliabilitas, daya pembeda, tingkat kesukaran, dan fungsi pengecoh menggunakan ITEMAN 4.4 dan ANATES V4. Hasil penelitian menunjukkan bahwa sebagian besar butir memenuhi kriteria validitas isi dengan nilai Aiken's V secara umum $\geq 0,78$. Koefisien Cronbach's Alpha sebesar 0,838 pada pretest dan 0,817 pada posttest, sedangkan ANATES menghasilkan koefisien reliabilitas sebesar 0,83 dan 0,84. Sebagian besar butir memiliki daya pembeda sedang hingga baik, dengan tingkat kesukaran berkisar dari mudah hingga sukar. Fungsi pengecoh secara umum lebih baik pada posttest. Temuan ini memberikan bukti awal mengenai kualitas instrumen pada tahap pengembangan awal. Namun, validitas konstruk belum diuji melalui EFA maupun CFA. Oleh karena itu, diperlukan penelitian lanjutan dengan sampel lebih besar dan beragam serta analisis validitas konstruk untuk memperkuat bukti kualitas instrumen.

Received : 15/08/2026

Revised : 10/09/2026

Accepted : 21/10/2026

✉ **Corresponding Author:** (1) Adelia Dewanti, (2) Master Program in Elementary School Teacher Education, (3) Universitas Sebelas Maret, (4) Jl. Ir. Sutami 36A, Kentingan, Jebres, Surakarta, 57126, Indonesia, (5) Email: adelia942@student.uns.ac.id

1. Introduction

Creative thinking skills are an important competency in 21st-century education because they enable students to generate diverse ideas, consider alternative solutions, and develop original responses to problems. At the elementary school level, these skills need to be developed through learning experiences and assessments that capture the process of generating ideas rather than merely recalling information. Studies of elementary learning indicate that creative thinking can be observed through the indicators of fluency, flexibility, originality, and elaboration (Numaningsih et al., 2023). These four indicators have also been used in various developments of creativity instruments in elementary education (Alfarisa et al., 2025; Mirza et al., 2026; Noorkholisoh et al., 2024).

Science learning provides a relevant context for developing and measuring these skills because students encounter natural phenomena that can be analyzed from multiple perspectives. The water cycle, for example, involves evaporation, condensation, precipitation, infiltration, and environmental issues that are close to students' daily lives. Research on science learning and assessment shows that contextual materials can encourage students to propose alternative ideas and solutions (Simatupang & Anas, 2024). Therefore, an instrument on the water cycle needs to be designed so that students' responses can demonstrate the variety and quality of their ideas more clearly.

The need for instruments specifically designed to measure creativity is also related to assessment practices in schools. A needs analysis in elementary schools showed that teachers still face difficulties in selecting indicators and designing creative-thinking instruments (Tumangger et al., 2022). Research on assessment-task design also indicates that teachers' understanding of the characteristics of creative tasks is an important factor in producing appropriate assessments (Wijayanti et al., 2022). Other findings show that creativity instruments for elementary schools need evidence of validity and reliability so that the measurement results can be justified (Noorkholisoh et al., 2024).

Several studies have examined creative-thinking assessment instruments in elementary education, but differences remain in the constructs, content contexts, and analytical approaches used. Alfarisa et al. (2025) developed a creative-thinking essay test for elementary students using Aiken's V, construct analysis, and reliability evidence, while Noorkholisoh et al. (2024) examined a creativity instrument using the Rasch model. Rohmatika et al. (2025) also analyzed creative-thinking test items using the Rasch model, but focused on renewable energy material. These studies provide important evidence regarding instrument quality; however, they do not specifically combine expert-based content validity with classical test theory analysis of multiple choice and essay items for measuring creative thinking on water cycle material among fifth-grade elementary students. Therefore, this study addresses this gap by integrating Aiken's V with classical item analysis, including reliability, discrimination, difficulty, and distractor functioning, in the specific context of elementary science learning on the water cycle.

In test-quality analysis, classical test theory remains relevant because it provides practical information about test consistency, item difficulty, discrimination, and response patterns. Classical item analysis can help identify items that are too easy or too difficult and items that are less able to distinguish among students (DeCarlo, 2023; Ikhsanudin et al.,

2023; Priyani & Sugiharto, 2024; Setiawan et al., 2026). In multiple-choice tests, distractor analysis is also important because options that are not selected or are selected by a very small proportion of students may indicate the need for revision (Sari et al., 2022; Syahrial et al., 2023). Comparative research on classical and modern test theories indicates that examining distractors can provide additional information for making decisions about item quality (Syahrial et al., 2023).

Content validity is an important stage before an instrument is empirically tested. The use of Aiken's V enables researchers to summarize expert agreement regarding the relevance and representativeness of items to the construct being measured (Amini, 2023; Nurjanah et al., 2023). After expert evaluation, empirical reliability, evidence discrimination, through difficulty, and distractor effectiveness is needed to provide a more comprehensive picture of item performance (Haryanto et al., 2026; Ismail & Zubairi, 2022; Setiawan et al., 2026). This integrated approach is consistent with test-analysis research that examines several item characteristics simultaneously rather than validity or reliability alone (Ikhsanudin et al., 2023; Sari et al., 2022; Setiawan et al., 2026). Empirical evidence of instrument quality can also be obtained through item analysis across different subject contexts (Ismail & Zubairi, 2022).

Based on this rationale, a study is still needed that specifically analyzes a creative thinking skills instrument on the water cycle in elementary schools by combining content validity and empirical item characteristics. This study aimed to analyze the content validity, reliability, discrimination, difficulty, and distractor functioning of a creative-thinking skills test instrument on the water cycle. The results are expected to provide empirical evidence for the use and improvement of creative-thinking assessment instruments in elementary schools (Alfarisa et al., 2025; Noorkholisoh et al., 2024; Setiani et al., 2024; Van Hooijdonk et al., 2024). Assessment development at the secondary-school level also shows that validation and empirical trials remain necessary when an instrument is intended to measure creativity (Putriani et al., 2025).

2. Research Method

This study used a descriptive quantitative design to analyze the quality of a creative-thinking skills test instrument on the water cycle based on classical test theory. The data were obtained from test blueprints, pretest and posttest items, expert rating sheets, and responses from 103 fifth-grade elementary school students in the 2025/2026 academic year. The analyzed instrument consisted of 15 multiple-choice items and 4 essay items.

The trial sample was selected using purposive sampling. The participants consisted of 103 fifth-grade students from SDII Al Abidin Surakarta and MI Alkharim Nganjuk. The selection considered the following criteria: (1) the school implemented the Merdeka Curriculum, (2) the fifth-grade students had received the water cycle material, and (3) access to the participants was permitted by the school. The instrument administration was conducted under standardized conditions, including consistent test duration, supervision, and instructions for all participants. Data were collected through documentation and test administration.

Content validation involved six validators, consisting of three lecturers who teach elementary IPAS courses and three fifth-grade elementary school teachers with at least 15 years of

teaching experience. The lecturers had expertise in elementary science content and assessment, while the teachers had direct experience teaching fifth-grade students. All validators were familiar with the water cycle material and the assessment of elementary students' creative-thinking skills. Thus, the validator panel represented perspectives from elementary science content, classroom teaching experience, and assessment at the elementary school level. The validators assessed each item using a four-point rating scale ranging from 1 (lowest) to 4 (highest).

The analysis was conducted in two stages. The first stage calculated content validity using Aiken's V based on ratings from the six validators. The second stage analyzed students' responses to obtain reliability coefficients, discrimination, difficulty, and distractor functioning. For multiple-choice items, ITEMAN 4.4 was used to obtain the Alpha coefficient, item correlation, difficulty, and response-option patterns, while ANATES V4 was used to complement the item analysis, particularly for essay items. This approach enabled a comprehensive examination of the instrument's item characteristics, as recommended in item-analysis studies (DeCarlo, 2023; Ikhsanudin et al., 2023; Setiawan et al., 2026).

The content validity of the creative-thinking skills test instrument was calculated using Aiken's V. The Aiken's V coefficient was calculated using Equation (1):

$$V = \frac{\sum S}{n(c-1)} \quad (1)$$

V = Content validity coefficient

$\sum S$ = Sum of all S values; the S value is calculated for each item

N = Number of raters

C = Number of rating categories

Where V is the Aiken's V coefficient, s is the score assigned by each validator minus the lowest score on the rating scale, n is the number of validators, and c is the number of rating categories. The Aiken's V coefficient was used to determine the content validity of each test item based on the ratings provided by the six validators.

3. Results and Discussion

3.1 Results

The analysis results for the pretest and posttest administered in Grade V science during the second semester of the 2025/2026 academic year are presented as follows.

3.1.1 Content Validity

Expert assessment was conducted to establish the content validity of the creative-thinking skills test instrument on the water cycle. The validation process involved six experts, consisting of three lecturers who teach elementary IPAS courses and three fifth-grade elementary school teachers. The lecturers were selected based on their expertise in elementary science content and their familiarity with assessment instruments, while the teachers were selected based on their experience teaching fifth-grade students. All validators were familiar with the water cycle material and the assessment of elementary students' creative thinking skills. Thus, the validation panel combined expertise in elementary science content, classroom teaching experience, and assessment of creative-thinking skills. The validators reviewed each item based on content suitability and instrument construction, including: (a) alignment of the item with learning indicators and objectives; (b) clarity of item formulation; (c)

appropriateness of language for students' developmental level; and (d) functionality of answer options in multiple-choice items.

The assessment used a four-point scale, from 4 to 1. A score of 4 was assigned when an item was highly consistent with the indicators and learning objectives, clearly formulated, unambiguous, used language appropriate to students' developmental level, and provided consistent and effective answer options. A score of 3 was assigned when the item was consistent with the indicators and learning objectives, sufficiently clear, understandable, and had well-functioning answer options, although minor shortcomings remained that did not affect the meaning of the item. A score of 2 was assigned when the item was less consistent with the indicators or learning objectives, insufficiently clear, used language that was less appropriate for students' developmental level, and contained weaknesses in the answer options requiring revision. A score of 1 was assigned when the item was inconsistent with the specified indicators or learning objectives, ambiguous or open to multiple interpretations, used language inappropriate for students' developmental level, and contained ineffective response options.

The evaluative results from the validators were then analyzed using Aiken's V to determine the qualification of each item before administration to the student participants. Aiken's V indicates the degree of agreement among validators regarding the relevance of each item to the assessed aspect. Higher Aiken's V values indicate stronger content validity of the developed instrument. The pretest content-validity results are presented in Table 1.

Table 1. Pretest Instrument Validity Results.

No.	Validator	Soal																		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1	2	3	4
1.	Dosen 1	4	4	4	4	3	3	4	1	4	4	4	4	4	4	4	4	4	4	4
2.	Dosen 2	4	4	4	4	2	4	4	3	4	4	4	4	4	4	4	4	4	4	4
3.	Dosen 3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
4.	Guru 1	4	3	4	4	4	4	4	3	4	4	4	4	4	4	4	4	4	4	4
5.	Guru 2	4	4	4	4	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4
6.	Guru 3	4	4	4	4	4	4	3	3	4	4	4	4	4	4	4	4	4	4	4
Validitas		1	0,94	1	1	0,77	0,94	0,94	0,66	1	1	1	1	1	1	1	1	1	1	1
		V	V	TV	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V

Source: Processed Research Data (2026)

The posttest content-validity results are presented in Table 2.

Table 2. Posttest Instrument Validity Results.

No.	Validator	Soal																		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1	2	3	4
1.	Dosen 1	4	4	1	4	3	4	4	4	4	4	2	2	4	4	4	4	4	4	4
2.	Dosen 2	4	4	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
3.	Dosen 3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
4.	Guru 1	4	4	3	4	4	4	4	3	4	4	4	4	4	4	4	4	4	4	4
5.	Guru 2	4	4	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
6.	Guru 3	4	4	3	4	4	4	3	4	4	4	4	4	4	4	4	4	4	4	4
Validitas		1	1	0,61	1	0,94	1	0,94	0,94	1	1	0,88	0,88	1	1	1	1	1	1	1
		V	V	TV	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V

Source: Processed Research Data (2026)

The six validators' ratings showed that most items met the content-validity criteria. Based on Aiken's V values, several multiple-choice items were below the 0.78 threshold and therefore required revision, whereas the other items met the predetermined criteria. Expert assessment before empirical testing provides evidence that the items are aligned with the learning objectives, creative-thinking indicators, content, and characteristics of fifth-grade elementary school students. However, this evidence represents content validity and should not be interpreted as empirical evidence of construct validity. Further construct-validity analysis is required to examine whether the empirical structure of the instrument adequately represents the intended dimensions of creative thinking.

Substantively, content validity in this study was established through the alignment among the items, creative

thinking indicators, learning objectives, and characteristics of elementary school students. This process provides a basis for concluding that the instrument not only has an appropriate item format but is also directed toward the intended construct. Similar findings have been reported in elementary creativity instrument development using Aiken's V as evidence of expert agreement (Alfarisa et al., 2025; Nurjanah et al., 2023).

3.1.2 Reliability

Reliability was examined to determine the internal consistency of the instrument in producing scores. The analysis showed that the pretest instrument had an Alpha coefficient of 0.838 and the posttest 0.817 using ITEMAN. ANATES V4 produced coefficients of 0.83 for the pretest and 0.84 for the posttest. These values indicate high measurement consistency. Reliability evidence is an important component of instrument quality evaluation because good content validity is insufficient when the resulting scores are inconsistent (Amini, 2023; Haryanto et al., 2026). Studies developing creativity instruments for elementary schools likewise show that reliability is an important component in ensuring that an instrument can be used repeatedly in an appropriate measurement context (Alfarisa et al., 2025; Noorkholisoh et al., 2024). The pretest reliability analysis using ITEMAN is presented in table 3, while the posttest results are presented in table 4.

Table 3. Pretest Instrument Reliability Analysis Using ITEMAN.

Score	Alpha	Sem	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	S-B Random	S-B First-Last	S-B Odd-Even
Scored items	0,838	1,512	0,819	0,637	0,816	0,837	0,778	0,898

Source: Processed Research Data (2026)

Table 4. Posttest Instrument Reliability Analysis Using ITEMAN.

Score	Alpha	Sem	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	S-B Random	S-B First-Last	S-B Odd-Even
Scored items	0,817	1,621	0,680	0,539	0,736	0,810	0,700	0,848

Source: Processed Research Data (2026)

The reliability results indicate relatively stable consistency across both forms of the test. The Alpha coefficients ranged from 0.817 to 0.838, while the reliability coefficients for the essay items obtained through ANATES ranged from 0.83 to 0.84. These findings provide preliminary evidence of measurement consistency for the respective test forms. However, the reliability coefficients should be interpreted together with content-validity evidence and item-level characteristics. Because construct validity was not examined, these reliability results alone cannot establish that the instrument adequately represents the intended multidimensional creative-thinking construct. Other studies using classical analysis also place reliability among the main indicators for determining whether a test instrument can be used or requires revision (Ikhsanudin et al., 2023; Setiawan et al., 2026; Tanjung & Nasution, 2022). The findings are also consistent with the use of ITEMAN and other measurement models in instrument evaluation, where reliability serves as an indicator of measurement consistency (Nuryanti et al., 2018).

3.1.3 Discrimination Index

Discrimination indicates an item's ability to distinguish students based on variation in the measured ability. The analysis in this study showed that most pretest multiple-choice items were in the fair and good categories, while the posttest was dominated by fair and good items, with one item

categorized as poor. For essay items, the pretest items were in the fair and good categories, whereas all posttest items were in the fair category. Examining discrimination is important because items that cannot distinguish between students with different ability levels may reduce the information provided by test scores. The pretest item discrimination results are presented in Table 5, while the posttest results are presented in Table 6.

Table 5. Pretest Item Discrimination Index Analysis.

Category	Items	Number
Poor/Discarded		-
Fair	3, 5, 6, 11, dan 15	5
Good	1, 2, 4, 7, 8, 9, 10, 12, 13, dan 14	10
Good Sekali		-

Table 6. Posttest Item Discrimination Index Analysis.

Category	Items	Number
Poor/Discarded	1	1
Fair	2, 4, 6, 7, 8, dan 9	6
Good	3, 5, 10, 11, 12, 13, 14, dan 15	8
Good Sekali		-

The discrimination results for the pretest essay items are presented in Table 7, while the posttest results are presented in Table 8.

Table 7. Pretest Essay Item Discrimination Index Analysis.

No.	DI%	DI	Category
1	28,57	0,29	Fair
2	31,25	0,31	Fair
3	48,21	0,48	Good
4	52,68	0,53	Good

Table 8. Posttest Essay Item Discrimination Index Analysis.

No.	DI%	DI	Category
1	44,64	0,45	Fair
2	44,64	0,45	Fair
3	46,43	0,46	Fair
4	54,46	0,54	Fair

Overall, the discrimination distribution indicates that most items were able to provide differentiated information for students with relatively high and low creative-thinking abilities. This supports the suitability of most items; however, items with low discrimination should still be reviewed in terms of construction, indicator alignment, difficulty, and possible problems with the answer key. Similar analyses based on classical test theory and the Rasch model are used to determine which items should be retained, revised, or eliminated.

For the pretest multiple-choice items, 10 items were categorized as good and 5 as fair. For the posttest, 8 items were categorized as good, 6 as fair, and 1 as poor. The posttest item with poor discrimination requires particular attention because it may not adequately distinguish students according to their ability levels. The wording, stimulus suitability, difficulty, and answer options should be reviewed before the item is reused. Item-analysis studies indicate that discrimination should be interpreted together with difficulty and response-option

patterns so that revision decisions are not based on a single index.

For the essay items, the pretest discrimination consisted of two fair items and two good items, whereas all posttest items were categorized as fair. This result indicates that the essay items provided sufficient response variation to distinguish students' abilities. This characteristic is important in creativity measurement because open-ended responses allow students to demonstrate variation in ideas, flexibility, originality, and elaboration.

3.1.4 Difficulty Index

The difficulty index describes the proportion of students who were able to answer an item correctly. In this study, the pretest multiple-choice items consisted of 8 easy, 4 moderate, and 3 difficult items, while the posttest consisted of 2 easy, 8 moderate, and 5 difficult items. For the essay items, the pretest consisted of three moderate and one difficult item, whereas the posttest consisted of one easy and three moderate items. This distribution demonstrates variation in the level of challenge provided by the instrument. Item-analysis studies also indicate that a balanced difficulty distribution is necessary so that a test does not only measure basic mastery but can also provide information across a broader ability range. The difficulty index results for the pretest and posttest multiple-choice items are presented in Tables 9 and 10, respectively.

Table 9. Pretest Instrument Difficulty Index Analysis.

Category	Range	Items	Number
Easy	$P > 0,70$	1, 2, 3, 4, 7, 8, 9 and 11	8
Moderate	$0,31 \leq P \leq 0,70$	5, 6, 10 and 12	4
Difficult	$P < 0,30$	13, 14 and 15	3

Table 10. Posttest Instrument Difficulty Index Analysis.

Category	Range	Items	Number
Easy	$P > 0,70$	2 and 3	2
Moderate	$0,31 \leq P \leq 0,70$	1, 4, 5, 6, 7, 8, 9 and 10	8
Difficult	$P < 0,30$	11, 12, 13, 14 and 15	5

The difficulty results for the pretest and posttest essay items are presented in Tables 11 and 12, respectively.

Table 11. Pretest Essay Instrument Difficulty Index Analysis.

No Item	Difficulty %	DI	Category
1	47,32	0,47	Moderate
2	48,21	0,48	Moderate
3	39,73	0,39	Moderate
4	29,46	0,29	Difficult

Table 12. Posttest Essay Instrument Difficulty Index Analysis

No Item	Difficulty %	DI	Category
1	70,09	0,70	Easy
2	62,95	0,62	Moderate
3	58,48	0,58	Moderate
4	49,11	0,49	Moderate

The dominance of moderate difficulty among most posttest items indicates that the instrument provided a more varied level of challenge than the pretest. In creative-thinking assessment, moderately difficult items can provide students

with opportunities to generate more than one idea or consider alternative solutions. However, easy items remain necessary to ensure that students understand the basic context before facing more complex thinking demands.

For the multiple-choice items, the difference in distribution between the pretest and posttest shows that not all items had the same difficulty level. This variation can help researchers identify parts of the instrument that are too easy or too difficult. Classical item-analysis research indicates that the difficulty index should be interpreted together with discrimination because very easy or very difficult items may have different discriminatory power.

The essay items tended to have moderate difficulty, with one pretest item categorized as difficult and one posttest item categorized as easy. This pattern was relatively proportional and provided students with opportunities to develop open-ended responses. Creativity instruments developed in other studies also use open-ended tasks to obtain evidence of fluency, flexibility, originality, and elaboration.

3.1.5 Distractor Functioning

A distractor is considered functional when an incorrect answer option is selected by students rather than being ignored. In this study, an option selected by fewer than 5% of students was categorized as a non-functional distractor (NFD), following the criterion proposed by Sajjad et al. (2020). Distractor analysis is necessary because response patterns can reveal the quality of the alternatives provided and help identify items that require revision. Previous studies have shown that non-functional distractors may reduce item discrimination and affect the overall quality of multiple-choice items (Sajjad et al., 2020). The distractor functioning results for the pretest and posttest multiple-choice items are presented in Tables 13 and 14, respectively.

Table 13. Pretest Instrument Distractor Functioning Analysis.

No.	Distractor	Result	Category
1	0.204	20,4%	Functional distractor
	0.786*	78,6%	Correct answer
	0.010	1%	Nonfunctional distractor
	0.000	0%	Nonfunctional distractor
2	0.049	4,9%	Nonfunctional distractor
	0.078	7,8%	Functional distractor
	0.874*	87,4%	Correct answer
	0.000	0%	Nonfunctional distractor
3	0,078	7,8%	Functional distractor
	0.796*	79,6%	Correct answer
	0.087	8,7%	Functional distractor
	0.039	3,9%	Nonfunctional distractor
4	0.039	3,9 %	Nonfunctional distractor
	0.165	16,5%	Functional distractor
	0.738*	73,8%	Correct answer
	0.058	5,8%	Functional distractor
5	0.136	13,6%	Functional distractor
	0.214	21,4%	Functional distractor
	0.456*	45,6%	Correct answer
	0.194	19,4%	Functional distractor

6	0.524*	52,4%	Correct answer	4	0.097	9,7 %	Functional distractor
	0.272	27,2%	Functional distractor		0.379*	37,9%	Correct answer
	0.117	11,7%	Functional distractor		0.301	30,1%	Functional distractor
	0.087	8,7%	Functional distractor		0.223	22,3%	Functional distractor
7	0.058	5,8%	Functional distractor	5	0.087	8,7%	Functional distractor
	0.767*	76,7%	Correct answer		0.049	4,9%	Nonfunctional distractor
	0.087	8,7%	Functional distractor		0.689*	68,9%	Correct answer
	0.087	8,7%	Functional distractor		0.175	17,5%	Functional distractor
8	0.019	1,9%	Nonfunctional distractor	6	0.097	9,7%	Functional distractor
	0.126	12,6%	Functional distractor		0.165	16,5%	Functional distractor
	0.767*	76,7%	Correct answer		0.621*	62,1%	Correct answer
	0.087	8,7%	Functional distractor		0.117	11,7%	Functional distractor
9	0.049	4,9%	Nonfunctional distractor	7	0.155	15,5%	Functional distractor
	0.155	15,5%	Functional distractor		0.524*	52,4%	Correct answer
	0.068	6,8%	Functional distractor		0.097	9,7%	Functional distractor
	0.728*	72,8%	Correct answer		0.223	22,3%	Functional distractor
10	0.117	11,7%	Functional distractor	8	0.184	18,4%	Functional distractor
	0.233	23,3%	Functional distractor		0.456*	45,6%	Correct answer
	0.563*	56,3%	Correct answer		0.320	32%	Functional distractor
	0.087	8,7%	Functional distractor		0.039	3,9%	Nonfunctional distractor
11	0.058	5,8%	Functional distractor	9	0.184	18,4%	Functional distractor
	0.757*	75,7%	Correct answer		0.194	19,4%	Functional distractor
	0.049	4,9%	Nonfunctional distractor		0.485*	48,5%	Correct answer
	0.136	13,6%	Functional distractor		0.136	13,6%	Functional distractor
12	0.184	18,4%	Functional distractor	10	0.126	12,7%	Functional distractor
	0.223	22,3%	Functional distractor		0.262	26,2%	Functional distractor
	0.505*	50,5%	Correct answer		0.485*	48,5%	Correct answer
	0.087	8,7%	Functional distractor		0.126	12,6%	Functional distractor
13	0.272	27,2%	Functional distractor	11	0.146	14,6%	Functional distractor
	0.175	17,5%	Functional distractor		0.660*	66%	Correct answer
	0.476*	47,6%	Correct answer		0.126	12,6%	Functional distractor
	0.078	7,8%	Functional distractor		0.068	6,8%	Functional distractor
14	0.049	4,9%	Nonfunctional distractor	12	0.689*	68,9%	Correct answer
	0.709*	70,9%	Correct answer		0.136	13,6%	Functional distractor
	0.107	10,7%	Functional distractor		0.097	9,7%	Functional distractor
	0.136	13,6%	Functional distractor		0.078	7,8%	Functional distractor
15	0.184	18,4%	Functional distractor	13	0.126	12,6%	Functional distractor
	0.204	20,4%	Functional distractor		0.262	26,2%	Functional distractor
	0.544*	54,4%	Correct answer		0.049	4,9%	Nonfunctional distractor
	0.068	6,8%	Functional distractor		0.563*	56,3%	Correct answer
1	0.078	7,8%	Functional distractor	14	0.078	7,8%	Functional distractor
	0.146	14,6%	Functional distractor		0.718*	71,8%	Correct answer
	0.738*	73,8%	Correct answer		0.087	8,7%	Functional distractor
	0.039	3,9%	Nonfunctional distractor		0.117	11,7%	Functional distractor
2	0.126	12,6%	Functional distractor	15	0.068	6,8%	Functional distractor
	0.689*	68,9%	Correct answer		0.117	11,7%	Functional distractor
	0.117	11,7%	Functional distractor		0.748*	74,8%	Correct answer
	0.068	6,8%	Functional distractor		0.068	6,8%	Functional distractor

Table 14. Posttest Instrument Distractor Functioning Analysis.

The analysis showed that in the pretest, 7 items had effectively functioning distractors, whereas 8 items still had distractors that were not optimal. This condition indicates that not all answer alternatives were sufficiently attractive to students. Distractors that were rarely selected can be reviewed in terms of conceptual suitability, plausibility, language, and their relationship to the stimulus. Similar variation in distractor functioning has been reported in elementary-school test analysis (Syahrial et al., 2023).

Effective distractors should be plausible alternatives for students who have not adequately mastered the concept. Thus,

distractor quality is related not only to the number of students selecting an option but also to its relationship with possible misconceptions or errors in thinking. Examining response patterns in this way can enrich the information obtained from total test scores.

In the posttest, distractor functioning showed a better tendency than in the pretest. This indicates that some answer alternatives were able to attract student responses more proportionally. Nevertheless, each distractor should still be examined individually before the instrument is reused. Recent item-analysis research emphasizes that examining distractors, discrimination, and difficulty simultaneously provides a stronger basis for revision decisions.

3.2 Discussion

The findings provide preliminary evidence regarding the quality of the creative-thinking skills test instrument for water cycle material at the initial development stage. The combination of content-validity evidence and empirical item analysis provides information about several aspects of instrument performance, including reliability, discrimination, difficulty, and distractor functioning. The results showed that most items met the predetermined criteria for content validity and demonstrated acceptable empirical item characteristics. However, these findings should be interpreted within the scope of the analyses conducted in this study because empirical construct validity was not examined through exploratory factor analysis (EFA) or confirmatory factor analysis (CFA). Therefore, the findings provide evidence of content validity and item-level characteristics but do not establish the empirical dimensionality of the creative-thinking construct.

The content validity results show that most items achieved Aiken's V values that met the predetermined criterion, although several items required revision. This finding indicates that the items were generally aligned with the learning objectives, creative-thinking indicators, and characteristics of fifth-grade elementary school students. The use of six validators consisting of lecturers and elementary school teachers provided perspectives from both measurement and classroom practice. This is important because creative-thinking assessment requires not only conceptual alignment but also language and task characteristics that are appropriate for students' developmental levels. Similar findings have been reported in studies developing creative-thinking instruments for elementary students, where expert judgment and Aiken's V were used to establish evidence of content validity. Therefore, items with lower validity coefficients should be revised as part of the instrument refinement process. Nevertheless, the Aiken's V results should be interpreted specifically as evidence of content validity and should not be regarded as empirical confirmation of the multidimensional construct structure.

The reliability findings provide additional evidence concerning the consistency of the instrument. The Alpha coefficients obtained from ITEMAN were 0.838 for the pretest and 0.817 for the posttest, while the essay-item reliability coefficients obtained using ANATES were 0.83 and 0.84, respectively. These results indicate relatively high internal consistency for the respective forms of the instrument. The slight difference between the pretest and posttest coefficients suggests that the consistency of the instrument remained relatively stable across administrations. However, reliability should be interpreted together with validity and item-level characteristics because a high reliability coefficient alone does

not guarantee that every item is appropriate for measuring the intended construct (Amini, 2023). Thus, the reliability coefficients in this study should be considered as one component of the overall evidence regarding instrument quality rather than as sufficient evidence of construct validity. This interpretation is consistent with previous studies emphasizing reliability as one component of a broader process of instrument evaluation (Noorkholisoh et al., 2024).

The discrimination results show that most multiple-choice items were able to distinguish students with different levels of creative-thinking ability. In the pretest, 10 items were categorized as good and 5 as fair, whereas in the posttest, 8 items were good, 6 were fair, and 1 was poor. The essay items also demonstrated sufficient discrimination, with two fair and two good items in the pretest and all four items categorized as fair in the posttest. These findings suggest that most items provided useful variation in students' responses. Nevertheless, the poor-discrimination posttest item requires further examination. Possible aspects to review include item wording, stimulus clarity, difficulty level, answer-key accuracy, and the relationship between the item and the intended creative-thinking indicator. Previous item-analysis research has emphasized that discrimination should be interpreted together with difficulty and response patterns rather than considered independently. Accordingly, discrimination results in this study provide information for item refinement but do not by themselves establish that the instrument validly represents the underlying creative-thinking construct.

The distribution of item difficulty also provides important information about the characteristics of the instrument. The pretest multiple-choice items consisted of 8 easy, 4 moderate, and 3 difficult items, while the posttest contained 2 easy, 8 moderate, and 5 difficult items. The essay items were predominantly moderate in difficulty. The greater proportion of moderate items in the posttest indicates that the instrument provided a broader range of challenge for students than the pretest. For creative-thinking assessment, variation in difficulty is important because students need opportunities to demonstrate both basic understanding and more complex responses. However, items that are extremely easy or difficult should be reviewed because they may provide limited information about differences among students' abilities, particularly when their discrimination indices are also low. Therefore, difficulty indices can support decisions about item refinement while being interpreted together with other item characteristics.

The distractor analysis provides additional information concerning the quality of the multiple-choice items. In the pretest, 7 items had effectively functioning distractors, while 8 items still contained distractors that were selected by fewer than 5% of students. In contrast, the posttest showed a better tendency in distractor functioning. This improvement suggests that several alternatives in the posttest were more plausible and capable of attracting responses from students who had not completely mastered the concepts. Distractors are particularly important in multiple-choice assessment because ineffective alternatives may make an item easier for reasons unrelated to students' actual understanding. Previous research has similarly emphasized that distractor functioning can provide additional information about item quality and should be considered alongside difficulty and discrimination. Thus, distractor analysis in this study provides practical information for identifying alternatives that require revision.

Taken together, the findings demonstrate the value of

examining several complementary item characteristics rather than relying on a single indicator of instrument quality. An item with acceptable content validity may still require revision if its discrimination is poor or its distractors are ineffective. Conversely, an item with an appropriate difficulty level may provide more useful measurement information when it also has adequate discrimination and functional response options. Therefore, the present findings support an integrated approach to item analysis in which content validity, reliability, discrimination, difficulty, and distractor functioning are examined before an instrument is considered for subsequent use. For elementary school teachers, items requiring revision, particularly those that insufficiently elicit originality and elaboration, can inform the design of creative-thinking tasks that encourage students to generate diverse ideas and develop their responses in greater detail.

The findings also have practical implications for elementary science assessment. The water cycle provides a contextual topic that can be connected with evaporation, condensation, precipitation, infiltration, and environmental conditions. Such contextual content provides opportunities for students to demonstrate fluency, flexibility, originality, and elaboration rather than merely recall factual information. Appropriately designed contextual tasks can therefore provide opportunities for students to demonstrate different aspects of creative thinking. In this context, the item characteristics identified in this study can provide information for teachers and instrument developers when reviewing and refining assessment tasks. However, the instrument should be subjected to further validation before stronger claims are made regarding its capacity to represent the multidimensional creative-thinking construct.

3.3 Limitations

This study has several limitations. First, the study did not examine construct validity using exploratory factor analysis (EFA) or confirmatory factor analysis (CFA). Therefore, although the study provides evidence of content validity and empirical item characteristics, it does not establish the empirical dimensionality of the creative-thinking construct, particularly the dimensions of fluency, flexibility, originality, and elaboration. Second, the empirical trial involved 103 fifth-grade students from two schools, which may limit the generalizability of the findings to broader elementary school populations. Third, items identified as requiring revision were not subjected to a subsequent empirical trial after revision. Therefore, the present findings reflect the characteristics of the items in the initial administration and do not yet provide empirical evidence regarding the performance of the revised items. Future studies should involve larger and more diverse samples, conduct construct-validity analyses such as EFA and CFA, and re-administer the revised items to a new sample to evaluate their item characteristics and construct validity.

Overall, the findings should be understood as preliminary evidence of instrument quality at the initial development stage. Items with low content validity, poor discrimination, or ineffective distractors should be revised before subsequent use. Further construct-validity analysis is also needed to determine whether the proposed dimensions of creative thinking are empirically supported. Following these additional validation steps, the revised instrument may provide stronger evidence for assessing fifth-grade students' creative-thinking skills on water cycle material. h-grade students' creative-thinking skills on water cycle material.

4 CONCLUSION

This study provides preliminary evidence of the quality of a creative-thinking skills test instrument on the water cycle at the initial development stage. The instrument showed evidence of content validity, acceptable reliability, and varied item characteristics in terms of discrimination, difficulty, and distractor functioning. Most items met the predetermined criteria, although several items require revision based on their content validity, discrimination, or distractor functioning.

However, the present study did not examine construct validity through exploratory factor analysis (EFA) or confirmatory factor analysis (CFA). Therefore, the findings should not be interpreted as conclusive evidence of the multidimensional construct validity of the instrument. Further studies involving larger and more diverse samples and construct-validity analysis are needed to examine the empirical structure of fluency, flexibility, originality, and elaboration. Following these additional validation steps, the revised instrument may provide stronger evidence for assessing fifth-grade students' creative-thinking skills on water cycle material.

References

- Alfarisa, F., Arisetyawan, A., Maulana, F. R., Lathifah, N. Y., Afriani, S., & Mashudi, M. (2025). Creative thinking in elementary education: Reliable assessment instruments for science and social studies. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 29(1), 13–25. <https://doi.org/10.21831/pep.v29i1.79429>
- Amini, R. P. (2023). Analisis validasi dan reliabilitas instrumen kemampuan berfikir kritis siswa sekolah dasar. *Edukatika*, 1(1), 1–10. <https://doi.org/https://doi.org/10.26877/edukatika.v1i1.125>
- DeCarlo, L. T. (2023). Classical item analysis from a signal detection perspective. *Journal of Educational Measurement*, 60(3). <https://doi.org/https://doi.org/10.1111/jedm.12358>
- Haryanto, M. L., Ruffi'i, Sabariah, & Adi. (2026). Psychometric Quality Of An Elementary School Mathematics Test: A Classical Test Theory Analysis Of Item Difficulty, Discrimination, And Distractor Functioning. *Jurnal Penelitian*, 11(2), 94–101. <https://doi.org/https://doi.org/10.46491/jp.v11i2.2550>
- Ikhsanudin, I., Novaliah, N., Hidayatullah, H., & Almizi, M. (2023). A practical using of the QUEST program to analyze the characteristics of the test items in educational measurement. *JISAE: Journal of Indonesian Student Assessment and Evaluation*, 9(1), 37–43. <https://doi.org/https://doi.org/10.21009/jisae.v9i1.31163>
- Ismail, F. K. M., & Zubairi, A. M. B. (2022). Item analysis of a reading test in Sri Lankan context using classical test theory. *International Journal of Learning, Teaching and Educational Research*, 21(3).
- Mirza, A. N., Hastomo, A., & Firmansyah. (2026).

- Development of a Figural Creativity Assessment Instrument Based on Torrance Tests of Creative Thinking (TTCT) for Elementary School Students. *Journal of Educational Sciences*, 10(4), 424–442. <https://doi.org/https://doi.org/10.31258/jes.10.4.p.424-442>
- Noorkholisoh, L., Yustiana, Y. R., Budiman, N., & Suryana, D. (2024). Validity and Reliability Analysis Using the Rasch Model in Developing Creativity Tests Instruments for Elementary School Students. *Jurnal Ilmiah Bimbingan Konseling Undiksha*, 15(1), 128–135. <https://doi.org/10.23887/jibk.v15i1.81249>
- Nurjanah, S., Istiyono, E., Widiastuti, W., Iqbal, M., & Kamal, S. (2023). The Application of Aiken's V Method for Evaluating the Content Validity of Instruments that Measure the Implementation of Formative Assessments. *Journal of Research and Educational Research Evaluation*, 12(2), 2023–2125. <http://journal.unnes.ac.id/sju/index.php/jere>
- Nurnaningsih, Hanum, C. B., Sopandi, W., & Sujana, A. (2023). Keterampilan Berpikir Kritis dan Berpikir Kreatif Siswa Sekolah Dasar dalam Pembelajaran Berbasis RADEC. *Jurnal Basicedu*, 7(1), 872–879. <https://journal.uui.ac.id/ajie/article/view/971>
- Nuryanti, S., Masykuri, M., & Susilowati, E. (2018). Analisis iteman dan model rasch pada pengembangan instrumen kemampuan berpikir kritis peserta didik sekolah menengah kejuruan. *Jurnal Inovasi Pendidikan IPA*, 4(2), 224–233. <https://doi.org/https://doi.org/10.21831/jipi.v4i2.21442>
- Priyani, T., & Sugiharto, B. (2024). Analysis of biology midterm exam items using a comparison of the classical theory test and the Rasch model. *Jurnal Pendidikan Biologi Indonesia*, 10(3). <https://doi.org/https://doi.org/10.22219/jpbi.v10i3.34345>
- Putriani, D., Kusumah, R. G. T., & Sukarno, S. (2025). Development of Assessment to Measure Creative Thinking Skills of Junior High School Students on the Topic of Chemicals in Daily Life. *ISEJ: Indonesian Science Education Journal*, 4(1), 19–27. <https://doi.org/https://doi.org/10.62159/isej.v4i1.365>
- Rohmatika, R., Fatmawati, S., & Suhartono. (2025). Analisis butir soal instrumen tes keterampilan berpikir kreatif pada materi energi terbarukan menggunakan Rasch model. *Jurnal EDUCATIO: Jurnal Pendidikan Indonesia*, 11(1), 92–106. <https://doi.org/https://doi.org/10.29210/1202525556>
- Sajjad, M., Iltaf, S., & Khan, R. A. (2020). Nonfunctional distractor analysis: An indicator for quality of Multiple choice questions. *Pakistan Journal of Medical Science*, 36(5), 982–986. <https://doi.org/https://doi.org/10.12669/pjms.36.5.2439>
- Sari, V., Utomo, A., & Sumarwati, S. (2022). Kualitas soal Bahasa Indonesia di SMP Muhammadiyah 1 Pontianak: Analisis butir soal. *Jurnal Pendidikan Bahasa Dan Sastra Indonesia*, 11(2), 112–119. <https://doi.org/https://doi.org/10.15294/jpbsi.v11i2.58091>
- Setiani, R., Syafi'ah, R., Hartanto, S., Arifianti, E. R., & Dwikoranto. (2024). Science creativity assessment instruments in the Merdeka curriculum in elementary schools. *At-Thullab: Jurnal Pendidikan Guru Madrasah Ibtidaiyah*, 8(2), 135–148. <https://doi.org/https://doi.org/10.30736/atl.v8i2.2062>
- Setiawan, K., Ruffi'i, R., & Sabariah, S. (2026). Item analysis of buffer solution multiple-choice test. *Pendas: Jurnal Ilmiah Pendidikan Dasar*, 11(2), 492–499. <https://doi.org/https://doi.org/10.23969/jp.v11i02.47117>
- Simatupang, D. A., & Anas, N. (2024). Pengembangan media AQUACA pada materi siklus air untuk meningkatkan kemampuan berpikir kreatif ilmu pengetahuan alam di sekolah dasar. *Jurnal EDUCATIO: Jurnal Pendidikan Indonesia*, 10(1), 376–385. <https://doi.org/https://doi.org/10.29210/1202424246>
- Syahrial, S., M, N., Ratno, S., & Sitanggang, A. K. (2023). Perbandingan keberfungsian distraktor dalam tes: Pendekatan teori tes klasik dan modern dalam konteks pendidikan sekolah dasar. *School Education Journal PGSD FIP Unimed*, 13(2). <https://doi.org/https://doi.org/10.24114/sejpgsd.v13i2.47032>
- Tanjung, Y. I., & Nasution, I. R. (2022). The development of creative thinking test instruments with Torrance indicators on direct current electricity materials. *Jurnal Pendidikan Fisika Indonesia*, 18(2). <https://doi.org/https://doi.org/10.15294/jpfi.v18i2.32117>
- Tumangger, M. H., Kartono, K., & Ridlo, S. (2022). Analisis Kebutuhan Pengembangan Instrumen Penilaian Kemampuan Berpikir Kreatif pada Mata Pelajaran Matematika untuk Siswa Sekolah Dasar. *Jurnal Basicedu*, 6(5), 9170–9179. <https://doi.org/https://doi.org/10.31004/basicedu.v6i5.3623>
- Van Hooijdonk, M., Mainhard, TKroesbergen, E. H., & Van Tartwijk, J. (2024). *Can elementary school teachers assess students' creative problem solving abilities? Teaching and Teacher Education*. 104644. <https://doi.org/https://doi.org/10.1016/j.tate.2024.104644>
- Wijayanti, P., Siswono, T. Y. E., Juniati, D., Abadi, A., & Hartono, S. (2022). How do mathematics teachers design tasks to assess students' creative thinking ability? *Jurnal Riset Pendidikan Matematika*, 9(2), 137–146. <https://doi.org/https://doi.org/10.21831/jrpm.v9i2.55122>