Journal of Social and Society (JOSS)
Volume 1, Issue 2, (2026), 221-236
Doi: https://doi.org/10.66784/joss.v1i2.84

Development and Validation of The Content of an Instrument for Creative Thinking Abilities in Writing Narratives for Elementary School Students

Yanti Arasi Sidabutar [1], Sumarno[2]

[1] PGSD, Universitas HKBP Nommensen Pematangsiantar, Indonesia
[2]Pendidikan Teknik Mesin, Universitas Negeri Medan, Indonesia
Correspondensi: arasiyanti@gmail.com

Received: October 10, 2026 *Revised: November 23, 2026* *Accepted: November 27, 2026*

Abstract

This study focuses on the development and content validation of an instrument designed to measure the creative thinking skills of elementary school students in narrative writing. It follows a research and development (R&D) framework, limited to the development of an initial product and its content validation. A quantitative descriptive approach was employed to calculate expert consensus, supplemented by a qualitative analysis of the experts' comments and suggestions. The instrument—comprising a test blueprint, the test itself, and a scoring rubric—was deemed suitable for use, subject to revisions, by three experts (with an overall mean V-coefficient of 0.75; specifically 0.85 for the test, 0.71 for the blueprint, and 0.67 for the rubric). Key revisions include separating indicators 4 and 6, refining the operationalization of "distinct ideas," clarifying the criteria for originality, and adjusting the final scoring procedure. A limited pilot test revealed good inter-rater reliability for the total score (ICC(2,1) = 0.82), though reliability was lower for the "originality" dimension; the results also indicated a correlation between the scores and the length of the written text. The instrument cannot yet be used as a final measurement tool; it requires further revision, re-validation, and testing with a larger sample, utilizing raters who have undergone calibration.

Keywords : Development, Validation, Instrument, Creative Thinking Abilities, Writing

This is an open access article under the CC BY-SA 4.0 .

INTRODUCTION

Think creative viewed as a key competency that needs to be developed from the elementary school level. At the international level, attention to this competency increased when *the Programme for International Student Assessment* (PISA) first assessed the creative thinking skills of 15-year- old students in the 2022 cycle. PISA defines creative thinking as the ability to engage in generating, evaluating, and refining original and diverse ideas (Jundana et al., 2026; Rodriguez-Marin et al., 2022; Sa'ban et al., 2024; Zahra et al., 2026). The results showed Indonesia's position, which requires careful attention: the average score of Indonesian students was 19 out of 60 points, far below the OECD average of 33, and only 5% of students reached level 5 or 6 (Juita et al., 2018). These achievements cannot be directly linked to a specific level, but they emphasize the need for more systematic efforts to monitor and develop creative

thinking from elementary school. Such efforts require measurement tools that teachers and researchers can use in real-life learning contexts.

Narrative writing is one promising context for observing elementary school students' creative thinking. In writing a story, students must generate ideas, choose the direction of the story, respond to the problems faced by the characters, and detail the events in their own way. Therefore, creative writing is seen as a measure of creativity that has ecological validity, that is, it is closer to real creative activities than artificial laboratory tasks (Pangestu et al., 2021; Siahaan et al., 2026; Simanjuntak et al., 2026). At the elementary school level, creativity in writing has long been a measurement concern. Cheung et al. (2001), for example, developed *the Chinese Creative Writing Scale,* adapted from *the Carlson Originality Scale,* to assess creativity in Hong Kong elementary school students' writing, with a factor structure encompassing flexibility, originality, and fluency. Thus, the idea of assessing creativity through elementary school students' written products has strong conceptual and empirical foundations.

However, assessing creativity in writing presents significant methodological challenges. First, many rubrics used to assess writing in schools address aspects not specifically related to creativity, such as spelling, grammar, punctuation, and text organization. However, few rubrics specifically designed to assess creative aspects of writing exist (Astutik et al., 2021; Harefa et al., 2026; Reden et al., 2026; Sidabutar et al., 2026). Second, the reliability and validity of analytical rubrics for writing assessment have not been adequately studied. Third, originality *is* highly context-dependent because its assessment relies on the comparison group and the assessor's knowledge. Bahar and Maker (2025) identify three commonly used approaches to assessing children's originality: norm- based, sample-based, and expert-referenced, and note that teachers are often ill-equipped to assess originality when clear guidelines and tools are not available. Fourth, writing length can potentially interfere with assessment: in a study of children's written stories using *the Consensual Assessment Technique,* story length explained 63% of the variance in *creative expressiveness assessments* (Toivainen et al., 2018). These findings suggest that a good instrument should formulate indicators, tasks, and scoring descriptors clearly and not rely on the quantity of writing (Setiyadi et al., 2025).

In Indonesia, studies on creative thinking in elementary school students' narrative writing are generally still descriptive. Rhosalia (2016), for example, described the fluency, flexibility, and originality of thinking in narrative writing in 97 fifth-grade public elementary school students in Surabaya, and found that only six students met the originality indicator, far fewer than the fluency (81 students) and flexibility (56 students). This finding is in line with the difficulties in measuring originality described above, while also indicating that originality is the most difficult dimension to develop and measure. Meanwhile, research on the development of creative thinking instruments for elementary school students in Indonesia has largely focused on specific subjects, including science and social studies (Carey et al., 2017) and critical and creative thinking in fourth-grade learning materials (Abdurahman, 2017) . These studies generally used expert validation with the Aiken's V index,

empirical testing, and reliability testing, thus providing a good methodological reference, but the domain they measured was not narrative writing.

Based on description Therefore, the urgency of this research lies in three things. *First,* the low achievement of creative thinking skills in Indonesian students in the 2022 PISA requires the availability of measurement tools that can monitor this ability starting in elementary school (Lisanty et al., 2023). *Second,* narrative writing is a common activity in Indonesian language learning in elementary schools, making narrative-based instruments easy to integrate into classroom practice (Simran & Adam, 2023). *Third,* the available instruments do not specifically separate the creativity aspect from the writing mechanics, do not provide operational guidelines for assessing originality, and have not been systematically tested for content suitability in the context of fifth-grade narrative writing. Without evidence of content validity, the scores generated by the instruments are difficult to interpret, let alone use as a basis for learning decisions. Therefore, content validation by experts is an important initial step before the instruments are empirically tested on a larger scale. wide.

Novelty study This lies in five aspects. *First,* the instrument was developed as an integrated package consisting of a grid, a series of picture-based narrative writing tasks, and an analytical rubric, so that the interrelationships of each component can be examined. *Second,* the construct is formulated in four dimensions of creative thinking (fluency, flexibility, originality, and detail) which are described into eight indicators, and the assessment deliberately separates creativity from spelling, grammar, neatness, and length of writing. *Third,* authenticity is operationalized as novelty relative to the group response pattern, in line with the contextual nature of authenticity in the literature (Mariana et al., 2018). *Fourth,* content validation examines not only each component, but also the suitability between components (grids, tasks, and rubrics) so that inconsistencies, such as indicators that are not triggered by tasks or overlapping score descriptors, can be identified before the trial. *Fifth,* this study targets a context that is still rarely studied, namely the assessment of creative thinking in narrative writing of fifth grade elementary school students, after previous studies have focused more on descriptions of abilities (Sari Togero & Usman, 2024) or on instruments for specific subjects (Erniwati et al., 2022). Based on the literature search conducted by the author, the combination of these aspects has not been found in a single reported instrument. on journal accredited.

Study This aims to (1) develop an instrument for creative thinking skills in writing narratives for elementary school students consisting of a grid, narrative writing tasks, and an analytical assessment rubric; and (2) obtain evidence of the validity of the instrument content based on expert assessment, along with input for improvements.

METHOD

This research is research and development *limited* to initial product development and content validation. The approach used is descriptive quantitative to calculate expert agreement, which is supplemented by qualitative analysis of expert comments and suggestions. The development procedure adapts an instrument development model consisting of sequential stages, from preparing specifications to interpreting results, as applied in the development of creative thinking instruments for fifth grade

elementary school students (Maulana et al., 2025). Because this research is small-scale, the stages covered are the preparation of specifications, writing instruments, content validation, and revision; large-scale empirical testing (factor analysis and internal reliability) is not covered and is noted as a limitation and further plan.

The research product consists of three interrelated instrument components.

1. **Grid.** Contains the construct of creative thinking ability in writing narratives in four dimensions, namely fluency, flexibility, originality, and elaboration, which refers to the divergent thinking framework (Guilford, 1967; Torrance, 2008). Each dimension is described into two indicators, resulting in a total of eight indicators, each with an operational definition and observed performance evidence.
2. **Narrative writing assignment.** In the form of a performance task sheet with a stimulus of four serial images about a child character who finds an old box, instructions for working on it, and a writing space. Fifth grade students work on assignments independently in 60 minutes. The stimulus is designed to be open enough so that students can develop the content, actions, problems, and ending of the story according to their own imagination.
3. **Analytical rubric.** Contains eight indicators with four levels of descriptors (scored 1–4). The total score ranges from 8–32 and the final score is calculated as (total score ÷ 32) × 100. The assessment focuses on creativity and does not take into account spelling, punctuation, neatness, or length of writing.

Content validation was carried out by a panel of experts selected purposively based on suitability of expertise. The panel consists of three people representing (a) Indonesian language learning in elementary schools, (b) educational evaluation and measurement, and (c) basic education. Creativity experts and fifth-grade teachers were not involved; this was noted as a limitation. In content validation, experts assess the extent to which the instrument represents the construct to be measured, so the number and qualifications of experts must be reported transparently (Wulandari et al., 2024).

A limited trial was conducted on 24 fifth grade students from one elementary school class [write the name/school code]. Each student worked on the narrative writing task independently, and their writing was given the code S01–S24. The writings were assessed by two independent assessors. Assessor 1 (P1) conducted an initial assessment assisted by a language model (Claude, Anthropic) based on handwritten transcription, while Assessor 2 (P2) was an Indonesian Language and Literature Education student in the School Field Practice who assessed the students' worksheets.

Data were collected using an Integrated Expert Validation Sheet containing the validator's identity, assessment instructions, and three assessment sections: (A) the grid (8 aspects), (B) the narrative writing test (13 aspects), and (C) the assessment rubric (12 aspects). Each aspect was assessed on a 4-level scale (1 = not appropriate, 2 = less appropriate, 3 = appropriate, 4 = very appropriate) with a column for notes and suggestions (Verawati et al., 2023). Aspects assessed included suitability to the construct and dimensions, accuracy and measurability of indicators, clarity of stimuli and instructions, opportunities for tasks to elicit each dimension, appropriateness of

time and format, clarity of descriptors for each level, and consistency between components. The sheet also included five qualitative review questions and a final decision (suitable without revision, suitable with revision, or not yet suitable) for each component and the entire instrument.

Research Procedures

Table 1. Stages Development And Content Validation

Stage	Activity	Output
1. Theoretical study and needs analysis	Examining the theory of creative thinking, the characteristics of narrative writing in fifth grade students, the curriculum, as well as the instruments and results of previous research.	Formulation of constructs, dimensions, and initial indicators
2. Preparation of specifications	Developing the grid: operational definitions, indicators, and evidence of performance	Grid
3. Writing assignments and rubrics	Developing a series of image stimuli, task instructions, a 4-level analytical rubric, and scoring rules	Draft assignments and rubrics
4. Validate content	Experts review grids, assignments, and rubrics with integrated validation sheets	Expert scores and comments
5. Analysis and revision	Calculating agreement index, grouping suggestions, and revising instruments	Revised instrument

Before validation, author Prepare all components in the same document for each expert to ensure uniform assessment of the material. Experts provide scores and comments independently. Validation results are used to determine which items to retain, revise, or delete. Revisions were made by prioritizing input related to the accuracy of indicators, the ambiguity of descriptors, and the suitability between tasks and rubrics (Resien et al., 2020).

Quantitative content validity. Expert agreement was analyzed using Aiken's V index (Aiken, 1980, 1985) on each aspect of the assessment, with the formula:

$$V = \frac{\Sigma s}{[n(c-1)]}$$

with $s = r - l_0$, r is the score given by the expert, l_0 is the lowest score on the scale (1), n is the number of experts, and c is the number of scale categories (4). The V value ranges from 0–1. An aspect is said to have adequate content validity if V meets the critical value according to the Aiken table for the number of experts and categories used; as a practical benchmark, many studies on the development of creative thinking instruments in elementary schools use a limit of around 0.80 or more (Sathanarugsawait et al., 2021). The V value is then averaged per component (grid, assignment, rubric) and overall. As a comparison, the content validity ratio *according*

to Lawshe (1975) can be calculated if experts are asked to give an essential or non-essential assessment of each indicator.

Qualitative analysis. Expert comments and suggestions were grouped according to components and themes (e.g., indicator overlap, descriptor clarity, stimulus transparency, and intercomponent fit), and then used as the basis for revision decisions. Items with V below the critical limit are revised or removed, while items with adequate V but receiving substantive suggestions are still reviewed.

Eligibility decision. The suitability of each component is determined based on a combination of the average V and the final expert decision. A component is declared feasible without revision if all aspects meet the critical value and there are no substantive suggestions, feasible with revision if some aspects require improvement, and not yet feasible if many aspects do not meet the critical value and experts state that it is not yet feasible. If the average V of the component is below 0.80 but all experts state that it is feasible with revision, the component is determined to be feasible with revision and must be revalidated after revision (Bagassi & Macchi, 2020).

Assessor consistency (limited advanced stage). Because the rubric is used to assess written products, inter-rater consistency needs to be demonstrated. In a limited pilot study, two independent raters assessed fifth-grade students' writing using the rubric, and agreement was analyzed using the selected *intraclass correlation coefficient (ICC) and reported according to guidelines (Hallgren, 2012; Koo & Li, 2016).* Prior to scoring, raters were trained using writing samples at each score level, and authenticity was benchmarked against the response patterns of the student group that served as the pilot study sample. This approach is in line with the practice of assessor training and improvement of unclear descriptors in creative writing assessment research (Asmarany et al., 2023).

The ICC models reported are ICC(2,1) two-tailed, absolute agreement, single rater as the primary index, ICC(3,1) for consistency, and ICC(2,k) for the average of two raters, each with a 95% confidence interval. ICC values are interpreted according to Koo and Li (2016): below 0.50 is poor, 0.50–0.75 is moderate, 0.75–0.90 is good, and above 0.90 is excellent (Khoiriyah & Husamah, 2018). Agreement per indicator is also reported as percentage of exact agreement, agreement within one score, and quadratic weighted kappa. Systematic differences between raters were tested using paired t-tests and Wilcoxon tests. As an exploratory analysis, Spearman correlations between word count and scores were calculated to examine whether creativity scores were affected by writing length.

RESULTS AND DISCUSSION

Validation Results

The three experts stated that the grid, test, and rubric were worthy of revision; the overall average V was 0.75 and only 5 of the 33 aspects achieved V ≥ 0.80. The assessment was carried out by three experts, namely an Indonesian language learning expert (P1), an educational evaluation expert (P2), and an elementary education expert (P3), on 33 aspects (grid 8, test 13, rubric 12) so that 99 assessments were obtained. [Author's note: complete the name, agency, and qualifications of the validator according to the original sheet before the manuscript is sent.] Of the 99 assessments,

there was no score 1; score 2 was given 5 times (5.1%), score 3 65 times (65.7%), and score 4 29 times (29.3%).

Table 2. Summary of Aiken's V index per component (n = 3 experts, 4 categories)

Component	Number of aspects	Average expert score (1–4)	Average V	V Range	Aspect V ≥ 0.80	Rating ≥ 3
A. Grid	8	3.12	0.71	0.56–0.78	0	95.8%
B. Narrative writing test	13	3.54	0.85	0.67–1.00	5	100%
C. Assessment rubric	12	3.00	0.67	0.44–0.78	0	88.9%
Overall	**33**	**3.24**	**0.75**	**0.44–1.00**	**5**	**94.9%**

The test received the highest score, while the rubric received the lowest. Aspects with V = 1.00 were all on the test: suitability of student characteristics (B2), openness of stimulus (B4), integration of stimulus and task (B5), fairness of implementation (B12), and appropriateness of format (B13). The lowest aspects were clarity of score levels 1–4 (C7) and practicality of rubric use (C12), each with V = 0.44, followed by accuracy of indicators on the grid (A3, V = 0.56). These three aspects were the only ones given a score of 2 (less appropriate) by the expert (Mohd Dzin & Lay, 2021).

V values below 0.80 need to be read with caution. On a four-category scale, the three experts who scored 3 (suitable) produced a V = 0.67, so the low V here reflects a rating of "suitable, with improvement" rather than "not suitable". Additionally, with three experts and four categories, the probability of obtaining V = 1.00 by chance is 0.016 (1/64), V = 0.89 is 0.063, and V = 0.78 is 0.156; so only V = 1.00 is significant at the 0.05 level. P3 experts tended to give higher scores (mean 3.55) than P1 (3.06) and P2 (3.12). P1 and P2 gave identical scores on 94% of aspects, the three experts were identical on 16 of 33 aspects (48.5%), and the difference between experts was never greater. from One points.

Table 3. Score expert And Aiken's V index per aspect

Code	Rated aspect	P1	P2	P3	Σs	V
A1	Construct suitability	3	3	4	7	0.78
A2	Dimensional conformity	3	3	4	7	0.78
A3	Indicator accuracy	2	3	3	5	0.56
A4	Clarity of formulation	3	3	4	7	0.78
A5	Operational definition	3	3	3	6	0.67
A6	Measurability of indicators	3	3	4	7	0.78

A7	Conformity of indicators and tasks	3	3	3	6	0.67
A8	Grid consistency	3	3	3	6	0.67
B1	Suitability of purpose	3	3	4	7	0.78
B2	Suitability of student characteristics	4	4	4	9	1.00
B3	Clarity of stimulus	3	3	4	7	0.78
B4	Openness of stimulus	4	4	4	9	1.00
B5	Integration of stimulus and task	4	4	4	9	1.00
B6	Clarity of instructions	3	3	4	7	0.78
B7	Opportunity to measure fluency	3	3	3	6	0.67
B8	Opportunity to measure flexibility	3	4	3	7	0.78
B9	Opportunity to measure originality	3	3	3	6	0.67
B10	Opportunity to measure elaboration	3	3	4	7	0.78
B11	Time and load feasibility	3	3	4	7	0.78
B12	Justice of implementation	4	4	4	9	1.00
B13	Format eligibility	4	4	4	9	1.00
C1	Conformity to construct	3	3	4	7	0.78
C2	Compliance of indicators	3	3	4	7	0.78
C3	Fluency descriptor	3	3	4	7	0.78
C4	Flexibility descriptor	3	3	3	6	0.67
C5	Originality descriptors	3	3	3	6	0.67
C6	Elaboration descriptor	3	3	3	6	0.67
C7	Clarity score level 1–4	2	2	3	4	0.44

C8	Gradation of score levels	3	3	3	6	0.67
C9	Observability of evidence	3	3	4	7	0.78
C10	Consistency between components	3	3	3	6	0.67
C11	Clarity of procedures	3	3	3	6	0.67
C12	Practicality of use	2	2	3	4	0.44

Based on rule decision In the Method, the grid and rubric had an average V below 0.80 and all experts stated that they were suitable with revision, so both were determined to be suitable with revision and required revalidation after revision. The test had an average V of 0.85, but three experts still provided substantive input, especially regarding the opportunity to measure variations in conflict or resolution, so the test was also determined to be suitable with revision. The overall decision of all three experts was the same : suitable with revision (Zhang et al., 2019).

Expert Input and Revision Action

The qualitative input from the three experts was highly convergent: nine of the ten themes were shared by two or three experts simultaneously (Table 4). The strongest theme was the overlap of indicators 4 and 6, which all three experts mentioned in review questions 1, 3, and 4, and reference comparator authenticity.

Table 4. Summary Input Experts and Action Proposed Revisions

No.	Expert findings	Expert	Revision action
1	Indicators 4 (conflict variation/resolution) and 6 (creative resolution) overlap.	P1, P2, P3	Indicator 4 = variety of alternative conflict/resolutions; indicator 6 = novelty and logic of the final solution; editorial is the same in the grid and rubric
2	Indicator score threshold 4: a score of 4 (“more than one”) includes a score of 3 (“two”)	P3	Thresholds are made exclusive and tiered (e.g. three or more, two, one, none)
3	"Different ideas" has not been defined; evidence may be double-counted	P1, P2, P3	Coding rules: ideas must be distinct, relevant, visible in the story; one piece of evidence does not count towards two indicators.
4	Authenticity does not yet have a comparative reference and anchor example	P1, P2, P3	List of common responses from the test writing collection and examples of anchors for each score level.

5	The limits of creativity and writing ability need to be emphasized.	P1, P2, P3	Confirmation that spelling, grammar, neatness, and length of writing are not assessed
6	One final story does not necessarily show several alternative conflicts/resolutions.	P1, P2, P3	Instructions for alternative fishing tasks without giving examples; the phrase "issue or incident" is made more inclusive
7	The print quality of the stimulus and the understandability of the instructions (the terms "narrative", "conflict")	P1, P2, P3	Adequate resolution color print; readability test on a number of fifth grade students
8	There is no final scoring procedure for two raters, training, and ICC model.	P1, P2, P3	Final score = average per indicator or adjudication if the difference is large; rater calibration; the ICC model is stated
9	The terminology is not yet uniform; theoretical references and traceability matrix are needed.	P1, P2, P3	Indonesian equivalents are consistent, theoretical references are included, indicator–task–rubric matrix is created
10	Administrative protocols and sufficiency of 60 minutes need to be confirmed.	P1, P2, P3	Standard protocol (assistance, stationery, time) and actual time recording on trials

Results Test Try Limited and Inter-Appraiser Agreement

P1 gave a mean total score of 20.8 out of 32 (SD 4.8; range 12–28; score 65.1) and P2 gave 22.3 (SD 4.8; range 14–31; score 69.7); both identified the conflict or resolution variation (indicator 4) as the indicator that least frequently achieved high scores. On indicator 4, no students received a score of 4 from both raters, and each rater gave a score of 1 to six students. In contrast, indicator 2 showed a ceiling effect: 14 students (58%) received a score of 4 from P1 and 8 students (33%) from P2, with no single score of 1. The largest differences between raters occurred on the authenticity dimension: P1 gave a score of 1 on indicator 5 to 9 students and on indicator 6 to 11 students, while P2 gave almost no scores of 1 (none on indicator 5, one student on indicator 6).

Table 5. Score Average And Agreement Inter-rater

Indicators / dimensions	Average P1	P2 Average	ICC(2,1) [95% CI]	ICC(3,1)	Exact deal	weighted κ
1. Ideas for developing the story	3.04	3.12	0.67 [0.38; 0.84]	0.67	62%	0.66
2. Series of events	3.58	3.33	0.38 [0.01; 0.67]	0.42	67%	0.37
3. Variation of character actions/responses	3.04	2.71	0.64 [0.30; 0.83]	0.69	50%	0.63
4. Variations of conflict/resolution	2.21	2.04	0.67 [0.38; 0.84]	0.68	71%	0.66
5. Novelty of ideas/events	1.92	2.79	0.51 [−0.10; 0.83]	0.79	21%	0.50
6. Creative solutions	1.67	2.62	0.32 [−0.11; 0.68]	0.61	8%	0.31
7. Characters, setting, atmosphere	2.25	2.79	0.47 [−0.02; 0.75]	0.61	38%	0.45
8. Supporting details of the event	3.12	2.88	0.81 [0.58; 0.92]	0.83	67%	0.80
Fluency (eng. 1–2)	6.62	6.46	0.66 [0.36; 0.84]	0.66	–	–
Flexibility (eng. 3–4)	5.25	4.75	0.72 [0.42; 0.87]	0.75	–	–
Originality (eng. 5–6)	3.58	5.42	0.41 [−0.08; 0.78]	0.79	–	–
Elaboration (eng. 7–8)	5.38	5.67	0.83 [0.64; 0.92]	0.84	–	–
Total score (8–32)	**20.83**	**22.29**	**0.82 [0.57; 0.92]**	**0.85**	**–**	**–**

On For total scores, interrater agreement was good: ICC(2,1) = 0.82 (95% CI 0.57–0.92), ICC(3,1) = 0.85 (95% CI 0.69–0.93), and ICC(2,k) for the average of the two raters = 0.90; Pearson correlation 0.85 and Spearman 0.80. Exact agreement across all pairs of indicator scores was 47.9 %, and agreement within one score was 96.3%. The two raters also selected nearly identical students in the top (four of five: S04, S15, S20, S24) and bottom (four of five: S10, S14, S17, S18) groups (Sukmawati et al., 2022). At the indicator level, the highest agreement was on supporting details of events (ICC(2,1) = 0.81) and the lowest on creative resolution (0.32; exact agreement only 8%) and sequence of events (0.38).

P2 gave a higher total score than P1 by a mean difference of 1.46 points (SD difference 2.59; t(23) = 2.76; p = 0.011; Wilcoxon test p = 0.009; limits of agreement −3.6 to +6.5). This difference was concentrated in originality: the mean dimension was 3.58 (P1) versus 5.42 (P2), with ICC(2,1) = 0.41 but ICC(3,1) = 0.79. This means that both raters ranked students similarly, but placed the standard for novelty at different levels (Hill et al., 2022).

Exploratory analysis showed that scores were strongly related to writing length. The average word count was 83.4 (range 40–159, calculated from transcriptions). Spearman correlations between word count and total score were 0.79 for P1 and 0.81 for P2 (both $p < 0.001$); the highest correlations were for elaboration (0.89 and 0.87), followed by flexibility (0.75 and 0.56), fluency (0.58 and 0.81), and originality (0.43 and 0.56).

Discussion

This instrument already has a framework accepted by experts, but it is not yet operational. All three experts rated the construct and four dimensions as appropriate (A1 and A2, V = 0.78), and the test obtained the highest V, especially for stimulus openness and its integration with the task (V = 1.00). Weaknesses were concentrated in the grid and rubric, namely indicator accuracy (A3), clarity of score levels (C7), and practicality of use (C12). This pattern is reasonable because the assessment of creativity in writing is highly dependent on how the assessor interprets the criteria: writing assessment rubrics are generally not specifically designed for creative aspects and originality is contextual (Mulyasa, 2022), while without clear guidelines and tools, assessors tend to rely on their own interpretations (Bahar & Maker, 2025). The expert validation approach using Aiken's V is in line with the development of other creative thinking instruments for elementary school students (Yilmaz et al., n.d.). However, with three experts only V = 1.00 is statistically significant, so V needs to be read together with the distribution of scores and qualitative input.

The pilot test data corroborates the experts' input. P1, P2, and P3 stated that a single final story does not necessarily demonstrate multiple alternative conflicts or resolutions. Similarly, none of the students achieved a score of 4 on indicator 4 according to both assessors. This could stem from the assignment requiring one problem and one ending, from overlapping scoring thresholds, or both, requiring joint revision of the assignment and rubric. On indicator 2, a ceiling effect was evident, and it is likely that the low ICC (0.38) reflects narrow scoring variation rather than rater disagreement, as many students simply followed the four-picture sequence to reach the four events (Setiyadi et al., 2025).

The findings on the originality dimension point to a benchmarking issue. The low originality mean for P1 (3.58 out of 8) aligns with Rhosalia (2016), who found that only 6 of 97 fifth-grade students met the originality indicator. However, the P2 mean is much higher (5.42), and the pattern of low ICC(2,1) (0.41) with high ICC(3,1) (0.79) indicates systematic differences, not inconsistencies in student ordering. This pattern supports the three experts' call for benchmarks and anchor examples (Table 4, theme 4), and aligns with Bahar and Maker's (2025) discussion of norm-based, sample-based, or expert-referenced approaches to scoring originality. Differences may also stem from differences in the assessors' backgrounds (language models and students) that were not calibrated together, necessitating repeated estimation of the originality ICC after calibration.

The strong correlation between scores and length ($\rho \approx 0.8$) is a finding that warrants further investigation, as the rubric explicitly states that length is not assessed. This correlation does not necessarily indicate bias: longer pieces allow more space for ideas and details, and count-based descriptors (ideas, events, actions) structurally

favor longer pieces. Nevertheless, this finding aligns with research on children's written stories, which found that story length explained 63% of the variance in *creative expressiveness ratings* (Hakan & Seval, 2021), and aligns with P2's warning against using length as a proxy for creativity. Descriptors need to emphasize diversity and quality, not quantity, and further analysis needs to control word count.

Based on these findings, priority revisions include (1) separating indicators 4 and 6 and creating exclusive score thresholds; (2) establish rules for coding ideas and events; (3) develop a list of common responses and examples of anchors for authenticity; (4) adding task instructions that provoke conflict alternatives or solutions without providing examples; and (5) set the final score as the average of two raters, because ICC(2,k) = 0.90 is higher than the single rater's ICC(2,1) (0.82). After revision, the instrument needs to be revalidated by experts and tested on a larger sample.

CONCLUSION

The IKBKMN grid, test, and rubric were deemed appropriate after revision by three experts (overall average V 0.75 ; test 0.85, grid 0.71, rubric 0.67). Priority revisions concerned the separation of indicators 4 and 6, the operationalization of different ideas, the reference for originality, and the final scoring procedure. A limited pilot test showed good interrater agreement on the total score (ICC(2,1) = 0.82), but weak on originality and showed a correlation between the score and the length of the writing. The instrument cannot be used as a final measuring tool until it is revised, revalidated, and tested on a wider sample with calibrated assessor.

REFERENCES

Abdurahman, B. (2017). Authoritative Agency for Tourism Zone: An Innovative Instrument for Destination Development? *Jurnal Bina Praja*, *9*(1), 15–27. https://doi.org/10.21787/jbp.09.2017.15-27

Asmarany, A. I., Mudriadi, W., Solissa, E. M., Sudadi, S., & Sudyana, I. N. (2023). Effectiveness Analysis of Students' Creative Thinking Skills Program by Optimizing The Development of Adobe Animate-Based Vlab Static Electricity Media. *Mudir: Jurnal Manajemen Pendidikan*, *5*(2), 392–397.

Astutik, W. B., Yuwana, S., & Hendratno. (2021). Development of Non-Fiction Text Digital Learning Media in Narrative Writing Skills for Fourth Grade Elementary School Students. *IJORER : International Journal of Recent Educational Research*, *2*(3), 275–292. https://doi.org/10.46245/ijorer.v2i3.99

Bagassi, M., & Macchi, L. (2020). Creative Problem Solving as Overcoming a Misunderstanding. *Frontiers in Education*, *5*(December), 1–10. https://doi.org/10.3389/feduc.2020.538202

Carey, E., Hill, F., Devine, A., & Szűcs, D. (2017). The modified abbreviated math anxiety scale: A valid and reliable instrument for use with children. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00011

Erniwati, E., Mertosono, S. R., Rofiqoh, R., & Gente, R. A. (2022). Picture Series in Teaching Writing Skills: A Literature Review. *Premise: Journal of English Education*, *11*(3), 679. https://doi.org/10.24127/pj.v11i3.5848

Hakan, K., & Seval, F. (2021). CIPP evaluation model scale: development, reliability and validity. *Procedia - Social and Behavioral Sciences*, *15*(1), 592–599. https://doi.org/10.1016/j.sbspro.2011.03.146

Harefa, J. S., Anwar, K., Nailatulliza, N., Lubis, Z. Z., & Phonna, M. (2026). Analysis of the Influence of Price Changes on The Demand of Stop Materials in the Community. *Journal of Social and Society*, *1*(2), 140–150.

Hill, J., Ogle, K., Gottlieb, M., Santen, S. A., & Artino, A. R. (2022). Educator's blueprint: A how-to guide for collecting validity evidence in survey- based research. *AEM Education and Training*, *6*(6). https://doi.org/10.1002/aet2.10835

Juita, E., Zulva, R., & Edial, H. (2018). Profile of Development of Critical Thinking Student Skills Assessment Instrument in the Course Geography of Disaster. *Sumatra Journal of Disaster, Geography and Geography Education*, *2*(1), 147. https://doi.org/10.24036/sjdgge.v2i1.142

Jundana, Q., As' ari, A. ari, Irwan, I., Zusmelia, Z., & Siska, F. (2026). Digital-Based Innovations in Nagari Administrative Services to Improve the Quality of Public Services in Pasaman Regency. *Marsiurupan: Journal of Community Development and Empowerment*, *1*(2), 130–142.

Khoiriyah, A. J., & Husamah, H. (2018). Problem-based learning: Creative thinking skills, problem-solving skills, and learning outcome of seventh grade students. *JPBI (Jurnal Pendidikan Biologi Indonesia)*, *4*(2), 151–160. https://doi.org/10.22219/jpbi.v4i2.5804

Lisanty, F. I., Rahayu, S., & Aulia, N. T. (2023). Teaching Narrative Text By Using Brainwriting Learning Model. *DEIKTIS: Jurnal Pendidikan Bahasa Dan Sastra*, *3*(1), 1–11.

Mariana, I., Purnomo, S., & Firdaus, A. R. (2018). Improving Student Writing Skills in Indonesian Language Learning Narrative Writing Materials Using Writing Process Approach. *COLLASE (Creative of Learning Students Elementary Education)*, *1*(3), 99–107.

Mohd Dzin, N. H., & Lay, Y. F. (2021). Validity and reliability of adapted self-efficacy scales in malaysian context using pls-sem approach. *Education Sciences*, *11*(11), 676. https://doi.org/10.3390/educsci11110676

Mulyasa, B. (2022). Reliability and validity of the student courtesy scale. *Journal of School Psychology*, *61*, 12–27. https://doi.org/10.1016/j.jsp.2022.12.001

Pangestu, W. A., Adiwijaya, P. A., & Purnami, N. M. A. (2021). Improving Students' Writing Skill Through The Implementation of Animated Short Movie in Teaching Writing Narrative Text. *Language and Education Journal Undiksha*, *4*(2), 114–124. https://doi.org/https://doi.org/10.23887/leju.v4i2.37669

Reden, S., Simamora, Y. T., Banjarnahor, D. N., & Thomas, R. A. (2026). The Urgency and Strategic Role of Students as Agents of National Identity Transformation in the Society 5.0 Era: The Concept of Patriotic Digital Collaboration. *Journal of Social and Society*, *1*(2), 98–105.

Resien, C., Sitompul, H., & Situmorang, J. (2020). The effect of blended learning strategy and creative thinking of students on the results of learning information and communication technology by controlling prior knowledge. *Budapest International Research and Critics in Linguistics and Education (BirLE) Journal*, *3*(2), 879–893. https://doi.org/https://doi.org/10.33258/birci.v3i2.99

Rodriguez-Marin, M., Saiz-Alvarez, J. M., & Huezo-Ponce, L. (2022). A Bibliometric Analysis on Pay-per-Click as an Instrument for Digital Entrepreneurship Management Using VOSviewer and SCOPUS Data Analysis Tools. *Sustainability*, *14*(24), 16956. https://doi.org/10.3390/su142416956

Sa'ban, L. M. A., Sadat, A., Pora, R., & Lumintao, J. (2024). Digital Transformation of Government: A Study of E-Government in Asean Countries. *Jurnal Ilmu Sosial Mamangan*, *13*(2), 208–226.

Sari Togero, S. S., & Usman, S. (2024). Improving Writing Skill of the Ninth Grade Students Through Peer Editing Technique. *Journal of English Language Teaching Society (ELTS)*, *2*(1), 1–14.

Sathanarugsawait, B., Samat, C., & Wattanachai, S. (2021). Enhancing Learners' Creative Thinking in the Massive Open Online Course (Moocs) Learning Environment Model for Higher Education. In *Innovative Technologies and Learning: 4th International Conference, ICITL 2021, Virtual Event, November 29–December 1, 2021, Proceedings 4* (pp. 533–541). Springer. https://doi.org/10.1007/978-3-030-91540-7_54

Setiyadi, M. W., Syamratulangi, S., Rapi, N. K., Adnyana, P. B., & Swandi, S. (2025). Development of science process skills instrument on ecology material: A validity and reliability study. *EduMatSains: Jurnal Pendidikan, Matematika Dan Sains*, *10*(1), 98–114. https://doi.org/10.33541/edumatsains.v10i1.6642

Siahaan, K. W. A., Simanjuntak, H., Yanti, F., Custilas, A., & Zouhair, E. (2026). Enhancing Student Competence Through SETS-Based Chemistry Instruction on the Colligative Properties of Solutions. *Marlajar: Journal of Science, Technology, and Innovation*, *1*(2), 151–167.

Sidabutar, L. S., Simatupang, T. L., & Panggabean, S. (2026). Analysis of Syntactic Functions in the Text of the Indonesian Proclamation of Independence. *Journal of Social and Society*, *1*(2), 151–158.

Simanjuntak, I. M., Telaumbanua, E. H., & Marbun, R. C. (2026). Management of Non-Academic Formative Assessment in Enhancing Self-Confidence and Collaboration in the Sunday School English Club. *Journal of Social and Society*, *1*(2), 123–139.

Simran, S., & Adam, R. (2023). Legal Analysıs Of Cbdc's Role As A Dıgıtal Payment Instrument Regulatory System In Indonesıa. *Asian Journal of Management, Entrepreneurship and Social Science*, *3*(03), 270–286.

Sukmawati, F., Setiawati, F. A., & Asmawi, S. (2022). The Impact of Z-Score Transformation Scaling on the Validity, Reliability, and Measurement Error of Instrument SATS-36. *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia*, *11*(2), 166–180. https://doi.org/10.15408/jp3i.v11i2.26591

Verawati, N. N. S. P., Harjono, A., Wahyudi, Gummah, S., & Prayogi, S. (2023). The Learning Tools of Inquiry-Creative Integrated with Ethnoscience: Its Validity and Effectiveness in Training Prospective Physics Teachers' Critical Thinking. In *International Joint Conference on Arts and Humanities 2023 (IJCAH 2023)* (pp. 1915–1923). Atlantis Press. https://doi.org/10.2991/978-2-38476-152-4_193

Wulandari, N. O., Sutrio, Doyan, A., & Rahayau, S. (2024). The Influence of Project Based Learning Model on Creative Thinking Skills and Physics Learning Outcomes. *Jurnal Penelitian Pendidikan IPA*, *10*(12), 10660–10669. https://doi.org/10.29303/jppipa.v10i12.9738

Yilmaz, F. G. K., Yilmaz, R., & Ceylan, M. (n.d.). Generative Artificial Intelligence Acceptance Scale: A Validity and Reliability Study. *International Journal of Human–Computer Interaction*, 1–13. https://doi.org/10.1080/10447318.2023.2288730

Zahra, N. A., Alamsjah, A., & Syah, F. (2026). The Effect of Implementing an Auction Item System and Social Factors on Profit and Loss Reports at PT. Pegadaian UPC Sukaria. *Journal Social Society*, *6*(2), 817–827.

Zhang, M., Yao, L., Haberman, S. J., & Dorans, N. J. (2019). Validity considerations for using automated scoring in speaking assessment. In *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech* (pp. 32–57). Taylor and Francis. https://doi.org/10.4324/9781315165103-3