JPII 14 (3) (2025) 412-429

Jurnal Pendidikan IPA Indonesia

https://journal.unnes.ac.id/journals/jpii

EVALUATING ANALYTIC RUBRIC QUALITY FOR ASSESSING PRE-SERVICE BIOLOGY TEACHERS' CREATIVE THINKING SKILLS

A.U. T. Pada*[1], H. Yulisman[1,2], Melvina[3,4]

[1]Department of Biology Education, Faculty of Teacher Training and Education,
Universitas Syiah Kuala, Indonesia
[2]Research Center of Elephant Conservation and Forest Biodiversity, Universitas Syiah Kuala, Indonesia
[3]Department of Physics Education, Faculty of Teacher Training and Education,
Universitas Syiah Kuala, Indonesia
[4]Department of Teaching and Leadership, Doctoral program of Teaching & Curriculum,
Syracuse University, New York, USA

DOI: 10.15294/jpii.v14i3.21851

Accepted: February 27th, 2025. Approved: September 10th, 2025. Published: September 11th 2025

ABSTRACT

This study validates an analytic rubric designed to assess divergent thinking as a core dimension of pre-service biology teachers' creative thinking. The research addressed the absence of psychometrically sound assessment tools in Indonesian biology education, where creativity remains underdeveloped despite policy mandates, such as *Kurikulum Merdeka*, and international benchmarks, like PISA 2022. The study pursued two objectives: (1) theoretical validation of the rubric through expert judgment and inter-rater reliability, and (2) empirical validation through exploratory and confirmatory factor analyses. Eight subject-matter experts evaluated the rubric's descriptors, and 218 pre-service biology teachers completed divergent-thinking tasks based on human physiology scenarios. Content validity was calculated using Aiken's V, inter-rater agreement with Kendall's W, and construct validity and reliability through EFA and CFA. Findings indicated strong content validity (Aiken's V = 0.78–1.00) and fair to good inter-rater reliability (Kendall's W = 0.50–0.79). Factor analyses confirmed a unidimensional structure, with CFA demonstrating good model fit (RMSEA = 0.04; CFI = 0.97). All factor loadings exceeded 0.30, and composite reliability was ≥0.70 for three dimensions, though solution variety was marginally reliable. Notably, feasibility emerged as more stable than variety, suggesting that appropriateness is a stronger indicator of creativity in applied STEM contexts. The validated rubric provides theoretical insights into the structure of divergent thinking and practical tools for formative assessment in biology education. Future studies should extend validation across institutions, examine measurement invariance, and explore predictive validity to ensure broader applicability.

Keywords: analytic rubric; creative thinking skills; divergent thinking; pre-service biology teachers

INTRODUCTION

The advancement of the 21st-century education agenda, as echoed in Sustainable Development Goal 4 (SDG 4), emphasizes the importance of inclusive, equitable, and quality education that promotes lifelong learning and nurtures critical and creative thinking. However, formal education in Indonesia generally provides limited opportunities for students to develop their creative potential. Curricula and assessments in schools are predominantly geared toward cognitive recall and reasoning, with insufficient attention to higher-order thinking, such as creativity (Karunarathne & Calma, 2024). This results in classroom practices that constrain students' imaginative expression and engagement, particularly in science education. Empirical data indicate that early childhood education promotes spontaneous

***Correspondence Address**
E-mail: andi_ulfa@usk.ac.id

creativity, yet such abilities often diminish as students progress through rigid, test-oriented school systems. Yildiz and Yildiz (2021) argue that current preschool and primary curricula may lack structured opportunities for stimulating creative development. In secondary education, especially in biology classrooms, teachers continue to prioritize multiple-choice assessments that align with convergent thinking patterns. It limits the nurturing of divergent thinking, which is the foundation of creative problem-solving (Pada et al., 2016).

In Indonesian biology education, assessment still heavily relies on rote memorization and convergent thinking, leaving little room for creative problem-solving. The latest PISA data frame this challenge in stark terms: only 31 percent of Indonesian 15-year-olds achieved baseline proficiency (Level 3) in creative thinking, compared to an OECD average of 78 percent (OECD, 2024). These numbers starkly highlight the urgent need for valid tools that not only elevate creative thinking but also measure it with clarity. Aligned with this imperative, the *Kurikulum Merdeka* enshrines creativity and critical thinking at the heart of 21st-century education, promoting project-based instruction and performance-oriented formative assessments (Azmi et al., 2023; Wang et al., 2023). Together, these policy directives and empirical benchmarks make a compelling case for introducing a rigorously validated analytic rubric, which is a tool that honors the curriculums philosophy while enabling educators to assess creativity in transparent and reliable ways

A preliminary study by Pada et al. (2018) confirmed that the divergent thinking process is a component of creative process ability. Divergent thinking constructs or produces a variety of possible responses, ideas, options, or alternatives to a problem (Thakral et al., 2021). Thus, divergent thinking can be interpreted as the ability to produce various solutions to problems using proper procedures and sound reasoning.

Assessing creative thinking skills for pre-service biology teachers requires a rubric that is both valid and reliable. A rubric serves as a scoring tool, encompassing assessment domains, rating scales, and evaluative descriptors (Stevens & Levi, 2005). By using a rubric, scorers can discern and differentiate critical thinking constructs with clarity and accuracy (Williams, 1999), ensuring that all key elements are evaluated. An effective rubric is distinguished by the presence of assessment criteria that are aligned with the purpose of the evaluation and articulated across a continuum of performance levels. Such rubrics not only define appropriate criteria but also provide well-structured assessment domains, transparent rating scales, and precise descriptors for each score level, thereby ensuring clarity, consistency, and fairness in evaluation (Brookhart, 2018). Additionally, it should yield consistent scoring outcomes.

However, recent literature reveals a critical gap in the assessment of divergent thinking within teacher education programs, particularly in the field of biology. While numerous studies affirm the importance of creativity in science learning, few have developed validated, rubric-based instruments that reliably and objectively measure divergent thinking. Existing scoring methods often rely on subjective interpretation, which leads to inconsistent evaluations across different raters (Han et al., 2024; Sureeyatanapas et al., 2024). Furthermore, little attention has been paid to developing analytic rubrics specifically designed to assess the creative thinking skills of pre-service biology teachers.

While various scoring methods exist, rubric-based scoring relies on a predefined scale that includes detailed descriptions for different performance levels. Rubrics support a systematic grading process by covering all sub-components of the construct and providing descriptive statements for each level of performance. By using scoring rubrics, graders can comprehensively evaluate all elements of a test (Olson & Krysiak, 2021). The use of descriptors also helps ensure consistent scoring among different raters (Shafiei, 2024).

The rubric serves as an assessment tool to outline the expected performance for each criterion in order to attain specific outcomes (Cooper, 2023). Experts highlight several advantages of using rubrics: it contributes to time savings and streamlined feedback processes (Hettithanthri et al., 2023), it aligns the performance improvement to the established standards (Sadler, 2009), it is effective supervision and progress monitoring enhances student focus, leading to higher-quality work and improved grades (Reddy & Andrade, 2010), and it provides the accurate, fair, and transparent assessments that mitigate personal bias (Isbell & Goomas, 2014). There are two most common types of rubrics: analytic and holistic rubrics (Imbler et al., 2023).

The analytic rubric may substantially impact the scoring processes, even if they are time-consuming, as lecturers typically need to examine the product individually several times to assess various skills (Sumekto & Setyawati, 2018). The use of a rubric determines the feedback provided to students based on each scoring criterion (Panadero et al., 2025). The overall scores are calculated by adding the scores assigned to each feature. Scores can be conditionally weighted in variances so that features like thesis, organization, and development award are more significant in the overall scoring (Sumekto & Setyawati, 2018).

Analytic rubrics are beneficial for formative assessment but require more time for grading than holistic rubrics. In contrast, holistic rubrics allow for faster grading and are well-suited for summative assessments. However, the holistic rubrics only offer a single overall score, which does not provide clear guidance on areas for improvement (Chowdhury, 2018). To choose between both rubric types, follow the preferences of evaluators, the nature of the assignments, or the specific educational goals and competencies (Sudaryanto & Akbariski, 2021).

The use of an analytic rubric contributes to scoring a text in its substantial aspects. A reliable rubric should have a specific scale of scores, along with a description, to determine the evaluation of each skill. For each score, a concise description of all performance levels should be provided (Koswara et al., 2021). The characteristics of creativity are uniqueness and originality, which begin with a search for a variety of possible solutions. After that, students would know whether the solution was different from other solutions and whether such a solution had never been there. To find a variety of alternative solutions, a student requires divergent thinking skills (Subali & Suyata, 2013). Creative thinking is a cognitive activity that can lead to the production of useful creative thoughts and new groups or individuals through the use of analytic rubrics (Abdellatif & El-Wakeel, 2025; Gunawan et al., 2025).

The absence of validated assessment tools and the scarcity of empirical research suggest that raters often rely on subjective judgments, resulting in inconsistencies in scoring and interpretation (Han et al., 2024; Sureeyatanapas et al., 2024). Trainers may evaluate the same responses differently, producing unreliable outcomes. To address this issue, several scholars have recommended the adoption of structured scoring frameworks as a way to enhance fairness and consistency (Elosua, 2022; Asli et al., 2024).

From a theoretical perspective, this study integrates psychometric rigor with the realities of biology pedagogy by embedding rubric descriptors within authentic human-physiology scenarios. These domain-specific anchors ensure that students' divergent thinking is evaluated in ways that reflect disciplinary practices rather than relying solely on generic creativity tests. This approach is consistent with Baer (2012), who states that divergent thinking is domain-dependent, underscoring the need for tailored assessment instruments in education, particularly within STEM disciplines.

Building on this gap, the present study pursued two aims: (1) the theoretical validation of an analytic rubric for divergent thinking through expert judgment and inter-rater reliability, and (2) the empirical validation of the rubric using exploratory and confirmatory factor analyses to establish construct validity and internal consistency. Unlike existing approaches that rely heavily on subjective interpretation, this research introduces a construct-valid, rubric-based instrument that enables consistent and reliable evaluation. The novelty of this work lies in integrating a psychometric construct validation framework with rubric development—an approach rarely applied in biology education. By providing a validated analytic rubric, the study contributes to the advancement of educational measurement and supports curriculum reform and teacher professional development initiatives aimed at fostering creative thinking.

METHODS

The rubric development procedure follows the steps proposed by López-Pastor and Pérez-Pueyo in Garcimartín et al. (2024), which consists of seven stages. While Mertler's seven-stage framework guided the development of the rubric, this study implemented a notable modification to enhance contextual relevance and empirical grounding. Specifically, Step 6, which was initially intended to collect examples of student work after rubric construction, was modified. In this study, representative student responses were collected during the descriptor generation stage, which is Step 3. This modification allowed the researchers to ground performance descriptors in authentic student outputs, thereby increasing clarity and reducing ambiguity in interpreting performance levels. The adjustment aligns with the contextual challenges in evaluating creative thinking among pre-service biology teachers and enhances the validity and practical application of the rubric.

The development of the analytic rubric for assessing divergent thinking was adapted and modified from Garcimartín et al. (2024), who elaborated on a seven-stage rubric construction framework tailored for the context of biology education. The steps were implemented as follows:

1. Review of Learning Objectives: The rubric development began by aligning rubric indicators with specific course learning objectives (CLOs) associated with human physiology. This ensured congruence between learning goals and the rubric criteria; 2. Identification of Observable Traits: Key attributes of divergent thinking (fluency, flexibility, originality, and feasibility) were defined, along with their behavioral indicators. Common student errors were also considered; 3. Descriptor Generation: Descriptors were constructed to represent high, moderate, and low levels of performance for each criterion. These descriptors served as qualitative anchors in the rubric scale; 4. Writing Anchor Levels: Descriptions of excellent and poor performance were elaborated to delineate the extremes for each attribute. This step allowed precise differentiation of student achievement; 5. Completion of Performance Levels: Intermediate levels of performance were described, resulting in a complete four-point scale for each trait; 6.Collection of Work Samples: Student responses from divergent thinking tasks were collected to exemplify each score point. These examples served as scoring references; 7. Revision and Refinement: Based on expert feedback and pilot testing, the rubric underwent revision to enhance clarity, reliability, and alignment with assessment purposes.

Each of these stages was not only implemented sequentially but also critically analyzed in terms of its effectiveness and contextual relevance. For instance, in Stage 1, reviewing learning objectives allowed for a direct alignment with curriculum goals, ensuring the rubric measured intended learning outcomes. During Stage 2, the identification of traits was supported by theoretical constructs from creativity research, making the attributes pedagogically grounded. Stage 3 emphasized descriptor clarity, which was validated through iterative discussions with experts. In Stages 4 and 5, precise wording was tested with example responses to ensure interpretability by different raters. Stage 6, which was modified from Mertler's original model, further grounded the rubric in actual student output. Finally, Stage 7's revision process involved both qualitative feedback and initial scoring trials, reinforcing the instrument's reliability and usability in classroom settings.

Building upon the adapted seven-stage rubric development framework, the formulation of clear, performance-based criteria is vital to reduce subjectivity and promote fairness and instructional effectiveness in assessment practices. Although the process of developing criteria may appear straightforward, it often yields varying levels of achievement, each associated with a specific weighted score. This approach has prompted debate due to its potential reliance on subjective interpretation. To ensure transparency and consistency, rubrics should be shared with students before they begin an assignment, clearly informing them of how their work will be evaluated. These rubrics must provide detailed explanations of both acceptable and unacceptable performance. In addition, they should align with the intended learning outcomes, be measurable, and employ descriptive and action-oriented language (Olson & Krysiak, 2021).

The rubric was designed to assess the divergent thinking skills of pre-service biology teachers using situational conditions as stimuli. Specifically, these conditions involve deviations from normal organ function (such as diseases) or scenarios opposite to typical bodily processes. The intention behind these situational stimuli is to prompt students to generate relevant solutions for the various cases presented. Students are expected to provide diverse responses and demonstrate divergent thinking patterns based on the essential concepts they have studied in human physiology. These responses ultimately reflect their understanding of the subject matter (Pada et al., 2018).

The concept holds significant influence, with Morris and Sharplin (2013), among others, considering analytical marking as the most dependable and consistent approach for evaluating creative thinking. They contend that an ideal analytical assessment should encompass the attributes of creative thinking and establish achievement categories to delineate the extent to which each quality has been attained. According to Guilford (1959), divergent thinking includes four primary attributes: fluency, flexibility, originality, and elaboration. All criteria and their descriptors are described, accompanied by a brief explanation of each attribute.

First, the fluency index is associated with several responses (Runco, 1985). Fluency refers to the ability to generate multiple ideas to respond to a given problem or question (Guilford, 1959; Kind & Kind, 2007). Then, fluency can be defined as the capacity to produce a variety of ideas specifically tailored to solve the problem.

Second, flexibility, as defined by Barak and Levenberg (2016), refers to the inclination to generate a diverse array of responses or to explore various categories and themes when generating ideas. According to Weiss and Wilhelm (2022), flexibility represents the capacity to produce different types of solutions or ideas, which becomes evident through the variety of solution categories. This aspect of divergent thinking is crucial. In the context of divergent thinking, flexibility has been shown to lack unique variance beyond fluency and originality, given that it is primarily determined by these dimensions and moderated by cognitive mechanisms such as working memory and processing speed (Weiss & Wilhelm, 2022). Moreover, the ability to differentiate between gifted and non-gifted children is better achieved through assessing flexibility rather than relying solely on fluency or originality scores (Arabacı & Baki, 2023).

Third, originality plays a crucial role in creative thinking (Wang & Hou, 2018); it refers to the capacity to generate uncommon and unconventional solutions or ideas (Kind & Kind, 2007). While everything common, conventional, or not unique is deemed unoriginal and uncreative, originality is often connected with novelty (Alajami, 2020; Anderson & Graham, 2021). Since originality is defined as the ability to produce uncommon or unusual ideas, Forthmann et al. (2020) calculated originality scores based on the frequency of the test taker's responses. This method assigns a frequency score to each response, where the average of these scores for all questions is used to determine the test taker's overall originality score.

Originality serves as a fundamental aspect of creativity, yet it alone cannot suffice as the sole criterion. Ideas and products that are merely original- even if they are exceptional- can become ineffective. Therefore, originality alone is insufficient to evaluate creativity; it must also be effective. In defining creativity within academic discourse, originality has long been recognized as a central feature. Nevertheless, Marzano (2022) emphasized that originality alone is insufficient; it must be accompanied by appropriateness and contextual fit to ensure that creative contributions are both innovative and meaningful. While originality is essential, it must be complemented by feasibility. Furthermore, Runco and Alabbasi (2024) indicate that only fluency, flexibility, and originality possess predictive validity. As a result, in this study, the ability to elaborate on an aspect of divergent thinking is replaced with feasibility.

In this study, divergent thinking ability is assessed based on the following criteria: (1) Alternative solution or fluency measures the ability to generate multiple solutions to a problem, which is assessed by the number of relevant solutions provided; (2) Original solution reflects the ability to produce unique or uncommon relevant solutions, which are judged by the frequency of responses. A score of 4 is assigned to responses found in less than 10% of total tests, 3 for less than 25%, 2 for less than 50%, and 1 for more than 50%; (3) Solution feasibility evaluates the ability to generate practical and applicable solutions to a specific case which is measured by the number of practical responses; and (4) Variety of solution or flexibility assesses the ability to generate different categories of solutions, based on the diversity of relevant response categories across various test types.

Content representation of the creative thinking skills rubric can be evaluated through expert judgment, which involves a rational analysis of the test content's feasibility or relevance by a panel of experts (Azwar, 2012). Expert judgment is a formal process in which subject-matter experts (SMEs) provide evaluations of the appropriateness and relevance of test content (Hammitt & Zhang, 2013). In this study, the SMEs comprised eight experts: lecturers specializing in human physiology, biology education, educational measurement, and biology teachers. Additionally, 218 pre-service biology teachers enrolled in their fifth– to seventh–semester courses, who had completed the Human Physiology course, were selected as participants. Eligibility required willingness to participate and completion of the divergent-thinking tasks. Exclusion criteria included incomplete responses or noncompliance with task instructions. Recruitment was conducted using cluster random sampling across course sections. A target minimum of 200 participants was set, following the rule of 10–15 participants per item for factor analysis. The final sample exceeded this threshold, providing a stable estimation for both EFA and CFA.

Two types of data were collected in this study. First, expert ratings were obtained from eight subject-matter experts (SMEs). These experts evaluated each rubric descriptor in terms of clarity, construction, and relevance using a structured 5-point scale. Second, student responses were gathered from divergent-thinking tasks that required participants to generate multiple, original, feasible, and varied solutions to physiology-based scenarios. All responses were then independently scored by trained assessors using the analytic rubric, with double-rating procedures applied to ensure consistency and minimize subjectivity in the assessment process.

The initial stage of analysis involved establishing the content validity of the rubric. Data were collected through validation sheets distributed to subject-matter experts (SMEs), who assessed each aspect and descriptor in terms of construction, relevance, and clarity using a five-point scale (poor, fair, average, good, very good) (Pada et al., 2015). To determine consistency among validators, the content validity index is calculated using Aiken's V index, with a threshold of 0.75 or higher indicating acceptable agreement. The Aiken index is based on the assessments of "n" SMEs regarding how well each descriptor represents the construct being measured. Aiken's V is calculated using the formula **$V = \Sigma s / [n(c-1)]$**, where "s" is derived from the SME's rating (r) minus the lowest possible score (lo), and "c" represents the highest validity score (Aiken, 1985). Inter-rater reliability was then examined using Kendall's W, where values of 0.40 or higher reflected fair to good agreement among raters.

Following the establishment of content validity and inter-rater reliability, construct validity was tested through Exploratory Factor Analysis (EFA), preceded by the Kaiser-Meyer-Olkin (KMO) measure and Bartlett's test to confirm sampling adequacy. Model fit was subsequently examined using Confirmatory Factor Analysis (CFA), with fit indices including χ^2/df, RMSEA, CFI, and NFI serving as criteria. Finally, reliability was established through Composite Reliability (CR), with values ≥ 0.70 considered indicative of satisfactory internal consistency.

RESULTS AND DISCUSSION

Rubric quality must be supported by validity. Validity refers to the extent to which empirical data and theoretical analysis support the interpretation and use of assessment results (Asli et al., 2024). According to Hadi (2001), if a test is based on a strong theoretical construct, its results can be considered valid.

In this study, the quality of the rubric was examined through both theoretical and empirical approaches. Theoretical analysis involved expert judgment for content validity and inter-rater agreement. Empirical analysis was conducted using factor analysis to test the construct structure and reliability of the rubric.

The primary objective of this research was to validate a rubric designed to assess creative thinking skills among pre-service biology teachers, ensuring that it aligns with recognized psychometric standards. To accomplish this, the initial stage of analysis focused on establishing content validity through the application of Aiken's V index, which provides a systematic approach to quantifying expert agreement. This step was fundamental to reinforcing the rubric's reliability and enhancing its overall construct validity.

The most common approach to gathering evidence of validity based on content involves subject matter experts (SMEs) assessing three critical aspects: (a) the relevance of the aspect to the tested domain, (b) the construction quality of the aspect, and (c) the clarity of the aspect being tested (Lawshe, 1975). This study uses the Likert scale to measure the alignment between each descriptor and the domain aspects that SMEs are asked to assess. Even though this method is useful for evaluating overall domain representations, it does not provide information on how effective each aspect and descriptor is in measuring all intended targets (Sireci & Faulkner-Bond, 2014).

The content validity of the creative thinking skills rubrics was evaluated using a rating scale. A panel of eight subject-matter experts (SMEs) and a five-point rating system were employed, and the content validity coefficient was calculated using Aiken's V formula with a threshold of ≥ 0.75 (Aiken, 1985). An Aiken index of 0.75 or higher indicates consensus among SMEs that an aspect or descriptor is relevant to the content area. In contrast, an index of < 0.75 suggests disagreement, which means the aspect or descriptor's lack of relevance (Hayati et al., 2023). A low Aiken index for aspects or descriptors signifies insufficient agreement among SMEs on its relevance.

The rating scale approach provides valuable insights into how well each aspect and descriptor within a group meets a specific purpose. Then, whether the aspect and descriptors can measure the intended objectives, this data can be summarized based on the criteria used to assess content validity, as illustrated in Table 1. The Aiken Index, which spans from zero to one, essentially reflects the proportion of subject matter experts (SMEs) who support these aspects, allowing for statistical evaluation (Sireci, 1995; Sireci & Faulkner-Bond, 2014). To achieve statistical significance, the analysis of content validity using the Aiken index with six SMEs must yield a V coefficient equal to or greater than 0.75 (Aiken, 1985). This critical threshold is derived from the right-tailed binomial probability table established by Aiken

Table 1. The Aiken's V values and the 95% confidence interval scores for Rubric Construction, Relevance, and Clarity

No	Aspects	Descriptors	Aiken V Index		
			Co	R	Cl
1	**Alternative Solutions**	Provide more than 3 correct answers. Explanation directly addresses actions to solve the problem **(Score 4)**	0.75	0.94	0.97
		Provides 3 correct answers Explanation directly addresses actions to solve the problem **(Score 3)**	1.00	0.97	0.94
		Provides 2 correct answers Explanation directly addresses actions to solve the problem **(Score 2)**	0.97	0.78	0.78
		Provide only 1 correct answer Explanation directly addresses actions to solve the problem **(Score 1)**	0.97	0.97	0.78
2	**Solution Feasibility**	Provide more than 3 relevant and feasible answers Explanation directly addresses actions to overcome the problem. **(Score 4)**	0.97	0.94	0.78
		Provide 3 relevant and feasible answers Explanation directly addresses actions to overcome the problem. **(Score 3)**	0.78	1.00	0.97
		Provide 2 relevant and feasible answers Explanation directly addresses actions to overcome the problem. **(Score 2)**	0.81	0.78	0.97
		Provide only 1 relevant and feasible answer Explanation directly addresses actions to overcome the problem. **(Score 1)**	0.97	1.00	0.94
3	**Solution Variety**	Answer questions with more than 3 categories of relevant answers. The explanation of the answer refers explicitly to the action to help solve the problem **(Score 4).**	0.97	0.94	0.78
		Answer questions with 3 categories of relevant answers. The explanation of the answer explicitly refers to the action to help solve the problem **(Score 3).**	0.97	0.94	0.97
		Answer questions with only 2 categories of relevant answers. The explanation of the answer explicitly refers to the action to help solve the problem **(Score 2).**	1.00	0.81	0.88
		Answer questions with only 1 category of relevant answers. The explanation of the answer explicitly refers to the action to help solve the problem **(Score 1).**	0.78	0.94	0.97
4	**Original Solutions**	Answer questions with very unique and complex ideas (the given answers only have something in common with ≤ 10% of the answers of all tests). The explanation of the answer explicitly refers to the action helping to solve the problem **(Score 4).**	1.00	0.97	0.94
		Answer questions with very unique and complex ideas (the given answers only have something in common with ≤ 25% of the answers of all the tests) The explanation of the answer explicitly refers to the action helping to solve the problem **(Score 3).**	0.78	1.00	0.75
		Answer questions with very unique and complex ideas (the given answers only have something in common with ≤ 50% of the answers of all tests) The explanation of the answer explicitly refers to the action helping to solve the problem **(Score 2).**	0.97	0.94	0.97
		Answer questions with very unique and complex ideas (the given answers only have something in common with > 50% of the answers of all the tests) The explanation of the answer explicitly refers to the action helping to solve the problem **(Score 1).**	1.00	0.75	1.00

Note: Based on Aiken's (1985) table of critical values, the critical value of V is 0.75, with a Type I error rate of 0.05. CI = Confidence Interval, Co = Construction, R = Relevance, Cl = Clarity.

The analysis indicates that the developed rubric to assess the creative thinking skills of pre-service biology teachers, evaluated through the divergent approach, is effective in representing the extent to which the approach was developed and its relevance to the tested domain. In summary, we can conclude that all aspects and descriptors show that Aiken's V index values pass the established threshold of 0.75 (refer to Table 1). This signifies robust content validity for all aspects and descriptors. There is agreement among the SMEs regarding the basic construction of the questionnaire's validity descriptor. The strong content validity, as suggested by SMEs' consensus, supports the correspondence of the descriptors in interpreting rubric standards.

Table 1 displays the Aiken index of 8 SMEs, along with a 95% confidence interval for the descriptors' construction criteria. The 95% confidence interval provides information regarding the width characteristics of the interval, which is approximately ±0.30, and this value varies depending on the V value of each descriptor. By using the wide interval characteristic as the measurement of the accuracy of V as the estimator, the researcher can make a statement about the adequate accuracy of V. For example, in this study, the researcher set the criteria of the interval's width characteristic on the 95% confidence interval of 0.30 to ensure the accuracy of V. If the characteristic is wide then the confidence of score interval exceeds that value. The researcher has two options: re-examining the content of the descriptor due to its lack of potential relevance (annulment) or increasing the number of SMEs to enhance the value of V. Increasing the number of SMEs will improve the accuracy of V and thus reduce the width of the confidence interval.

Inter-rater reliability measures the level of agreement among subjective ratings provided by multiple raters. It addresses the question of whether the rating system is consistent. A high level of inter-rater reliability indicates that different raters give similar ratings for the same descriptors, while low reliability suggests inconsistencies among the ratings. High inter-rater reliability indicates that multiple raters' ratings for the same descriptors are consistent. In contrast, low reliability means they are inconsistent (Lange, 2011).

Kendall's coefficient of concordance, also known as Kendall's W, measures the level of agreement among multiple raters, emphasizing consensus rather than simply absolute agreement. Kendall's coefficient is highly beneficial, particularly for analyzing ordinal ratings. This coefficient ranges from 0 to 1, with higher values indicating stronger inter-rater reliability. A coefficient value above 0.9 is considered excellent, while a value of 1 signifies perfect agreement (Legendre, 2005).

Table 2. Inter-rater reliability values for Rubric Construction, Relevance, and Clarity

Rubric Aspects	Kendall's W values of construction	Kendall's W values of relevance	Kendall's W values of Clarity
Alternative Solutions	0.79	0.65	0.53
Solution Feasibility	0.79	0.71	0.42
Solution Variety	0.67	0.50	0.56
Original Solutions	0.77	0.72	0.46
Average Scores	**0.75**	**0.65**	**0.49**

The reliability values of inter-raters, as measured by Kendall's W, are presented in Table 2. The critical value of Kendall's W for the 5% significance level with 8 SMEs is ≥ 0.28 (Neupane & Bhattarai, 2024). In short, all aspects of the rubric demonstrate Kendall's W values exceeding the established threshold of 0.28 (see Table 2). Since all aspects of the rubric had Kendall's W values above this threshold, we can conclude that there is a statistically significant level of agreement among the raters across all rubric criteria, confirming the rubric's reliability.

Referring to Fleiss et al. (2003) agreement rule of thumb, generally, values above approximately 0.75 indicate excellent agreement beyond what could occur by chance. Values below 0.40 indicate poor agreement, while those between 0.40 and 0.75 signify fair to good agreement beyond chance levels. Since Kendall's W values for all aspects and criteria are above 0.40, indicating fair to good agreement among SMEs, this level of agreement implies that different raters consistently interpret the rubric criteria. Therefore, the agreement enhances the credibility of the rubric as a reliable assessment tool.

Empirical analysis is conducted through factor analysis, which consists of two main types: Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). EFA is utilized to explore theoretical concepts, while CFA is employed to confirm theories derived from EFA. CFA is a statistical technique widely used to validate construction in the psychological testing literature (Natalya & Purwanto, 2018; Goudarzian, 2023). This article demonstrates the application of EFA and CFA to provide evidence of validity construction in the development of evaluation instruments. Specifically, this study aims to validate the construction of the creative thinking test instrument for pre-service biology teachers. A review of CFAs, which is based on theoretical evidence, is an essential part of the validation process (DiStefano & Hess, 2005). Reliability is assessed through the internal consistency of indicators of construction, which reflect the degree to which each indicator represents the common latent construct. Data analysis starts with an exploratory factor analysis, followed by the Kaiser-Meyer-Olkin (KMO) and Bartlett's tests, which are used to determine the adequacy of the sample. The test results indicated that the KMO MSA value was 0.811 > 0.05 with a p-value of 0.000 < 0.05. These results suggest that the matrix data have sufficient correlation to be used for factor analysis. Furthermore, the analysis of the communality values revealed that all rubric descriptors had a value above the acceptable limit of 0.30 (Mooi & Sarstedt, 2011). As a result, all of the 16 descriptors tested had a communality value above 0.30, indicating that these 16 descriptors can be included in the subsequent analysis.

Table 3. KMO Dan Bartlett's test

Creative Thinking Skills Rubric		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.811
Bartlett's Test of Sphericity	Approx. Chi-Square	1526,891
	df	120
	Sig.	.000

Next, the dimensions of the measuring instrument can be determined based on the results of factor analysis, which can be observed through the acquisition of eigenvalues for each factor. Hattie (1985) examined unidimensionality based on the ratio of the most significant first and second eigenvalues of the tetrachoric correlation matrix. Hutten stated that the ratio criterion is a widely used procedure, where high values indicate unidimensionality and low values indicate multidimensionality. Another method mentions that the percentage of total variance explained by the first component is often considered an index of unidimensionality. Reckase (1979) recommended that for good calibration, the data should explain a total percentage of variance by the first component of 20% or more to meet unidimensional assumptions.

Table 4. Total Variance Explained

Component	**Initial Eigenvalues**			**Extraction Sums of Squared Loadings**		
	Total	**% of variance**	**Cumulative %**	**Total**	**% of variance**	**Cumulative %**
1	**5.315**	**33.218**	33.218	5.315	33.218	33.218
2	2.062	12.888	46.106	2.062	12.888	46.106
3	1.634	10.214	56.319	1.634	10.214	56.319
4	1.325	8.281	64.600	1.325	8.281	64.600
5	0.960	6.001	70.601			
6	0.771	4.820	75.421			
7	0.651	4.068	79.489			
8	0.580	3.626	83.115			
9	0.501	3.129	86.244			
10	0.445	2.779	89.023			
11	0.394	2.460	91.483			
12	0.368	2.303	93.786			

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
13	0.289	1.804	95.591			
14	0.259	1.620	97.211			
15	0.242	1.514	98.725			
16	0.204	1.275	100.000			

Extraction Method: Principal Component Analysis.

The test results indicate that the largest eigenvalue was 5.315, which explains that 33.218% of the variance exceeds the 20% threshold (Table 4). This suggests that the developed rubric is unidimensional. With the confirmation of unidimensionality, the assumption of local independence is automatically proven (Embretson & Reise, 2013). An assessment of the communality values revealed that all descriptors exceed the minimum acceptable threshold of 0.30 (Mooi & Sarstedt, 2011). The EFA identified four factors with eigenvalues greater than 1. This indicates, based on factor analysis, that four distinct factors can be identified, aligning with the aspects outlined in the rubric.

The findings from the EFA indicate that the rubric consisted of 16 descriptors categorized into four factors. These factors are: (1) alternative solutions, which represent the ability to propose multiple solutions to given problems; (2) original solutions, which assess the capacity to develop relevant and unique solutions; (3) solution feasibility which evaluates the effectiveness of solutions in addressing a specific problem; and (4) solution variety which measures the ability to generate diverse categories of solutions. Each of these factors comprises four descriptors.

The initial phase of the fit test aims to assess the overall compatibility or goodness of fit (GOF) between the data and the proposed model. There is no single metric available to evaluate a model's feasibility (Bollen & Long, 1993). Due to this limitation, several comparative fit indices have been developed to assess the fit of a given model relative to a baseline model. Since the sampling distribution for this index remains unknown, researchers rely on a general rule of thumb to determine an acceptable fit level (Shook et al., 2004). Until more definitive measures are established, multiple steps must be taken to demonstrate model compatibility (Breckler, 1990). Using various fit indices ensures that the researcher does not selectively report only those that support their model, thereby avoiding bias. Gerbing and Anderson (1992) suggest that among the most robust and reliable indices are the Normed Fit Index (NFI) and the Comparative Fit Index (CFI) (Hu & Bentler, 1998).

Table 5. Results of the Overall Model Fit

GOF Measure	Fit Degree Target	Estimated Results	Fit Degree
Normed χ^2	Normed $\chi^2 < 2$	1.22	good fit
p-value	$p \geq 0.05$	0.07	good fit
RMSEA	RMSEA ≤ 0.05	0.04	close fit
ECVI	Small & close to saturated ECVI values	M* = 1.76 S* = 2.39 I* = 12.38	good fit
AIC	Small & close to saturated AIC values	M* = 200.33 S* = 272.00 I* = 1470.89	good fit
NFI	NFI ≥ 0.90	0.91	good fit
GFI	GFI ≥ 0.90	0.89	marginal fit
CFI	CFI ≥ 0.92	0.97	Excellent fit
RMR	SRMR ≤ 0.09	0.07	good fit

**M = Model; S* = Saturated; I* = Independence*

According to Hair et al. (2013), several indicators are commonly used to evaluate the suitability of a measurement model. The Chi-square (χ^2) test is the primary and only statistical test in GOF analysis. Ideally, a lower χ^2 value is preferred with a significance level of 0.05 or higher ($p \geq 0.05$), which indicates a good model fit. The following section presents the GOF statistical measurements for the assessment model of creative thinking skills, which support the conative aspect of pre-service biology teachers.

As shown in Table 5, the p-value is 0.07 ($p \geq 0.05$), indicating the model demonstrates a good fit (Rosnawati et al., 2015). Another important fit index is the Normed Chi-square (Normed χ^2), which represents the ratio of the Chi-square to the degrees of freedom. A recommended range for this value is between 1.00 and 2.00; with a value of 1.22, the model is considered a good fit (Mezo & Short, 2012).

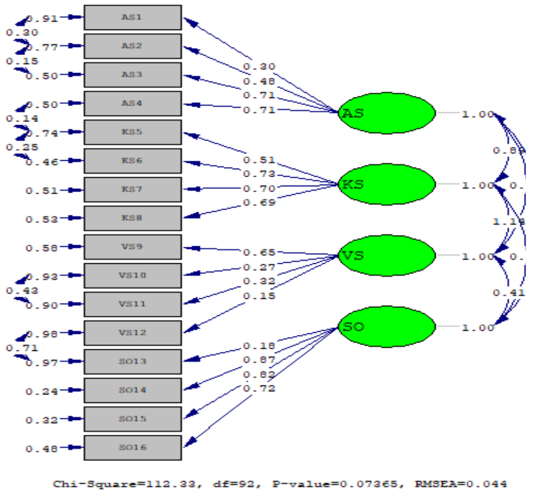

Figure 1. Final Construct of Creative Thinking Skills Rubric

Additionally, the Root Mean Square Error of Approximation (RMSEA) was used to assess model fit (DiStefano & Hess, 2005). An RMSEA value of 0.04 (RMSEA ≤ 0.05) suggests that the model has a close fit (Browne & Cudeck, 1993). In model evaluation, the Expected Cross-Validation Index (ECVI) plays a crucial role, where a value closer to the saturated ECVI indicates a good fit. In contrast, proximity to the independent ECVI suggests a poor fit. This principle also applies to the Akaike Information Criterion (AIC). The assessment model for creative thinking skills supports the conative aspect of pre-service biology teachers, with ECVI and AIC values closer to the saturated model, indicating a relatively good model fit. As shown in Figure 1, the Chi-Square value is 112.33 with degrees of freedom (df) = 92 and a p-value of 0.07365. These results suggest that the model demonstrates a good fit and aligns with the criteria presented in Table 5.

Convergent validity is the method used to evaluate construct validity. The term *"construct"* refers to a developed theoretical framework used to explain certain phenomena (Wiersma, 2000). According to van Dalen (1973), constructs typically represent complex concepts that include several related factors. An indicator is considered to have strong validity in relation to its construct or latent variable if it meets the following criteria: (a) the *t-value* of its factor loading exceeds the critical threshold (t-value ≥ 1.96) (Doll et al., 1994; Hair et al., 2013); and (b) the standardized factor loading is ≥ 0.30 (Gorsuch, 2013; Mooi & Sarstedt, 2011). The *t-values* and standardized factor loadings for the creative thinking skills rubrics are presented in Table 6.

Based on the summary of the *t-value* analysis results, it is evident that all *t-values* for the factor loadings of the variables or descriptors exceed 2 (*t-value* > 2). This indicates that the factor loadings of the variables in the model are statistically significant and not equal to zero. In addition, the standardized factor loading values for each variable are above the minimum threshold (*standardized factor loadings* > 0.3). In conclusion, all observed variables demonstrate strong validity in relation to their corresponding latent constructs.

Table 6. t-value, Standardized Factor Loadings, and Construct Reliability

Descriptors	SLF*	t-value	CR**
Alternative Solution Factor			
AS_01	0.446	2.89	0.713
AS_02	0.570	4.71	
AS_03	0.754	7.48	
AS_04	0.690	7.56	
Solution Feasibility Factor			
KS_05	0.617	5.50	0.789
KS_06	0.796	8.69	
KS_07	0.677	8.18	
KS_08	0.684	7.96	
Solution Variety Factor			
VS_09	0.771	5.54	0.708
VS_10	0.593	2.71	
VS_11	0.536	3.22	
VS_12	0.547	2.05	
Original Solution Factor			
SO_13	0.453	2.38	0.818
SO_14	0.889	10.83	
SO_15	0.812	9.97	
SO_16	0.716	8.43	

*SLF = Standardized loading factor
**CR = Construct Reliability

The evaluation of the reliability of the measurement model can be carried out using *a composite reliability measure,* also known as *construct reliability* (CR) (Ghadi et al., 2012; Hair et al., 2013). Reliability scores between 0.6 and 0.7 may still be considered acceptable, provided that other indicators of construct validity within the model are satisfactory. A high level of construct reliability suggests strong internal consistency, meaning that all measurements consistently reflect the same underlying construct. CR can be calculated by the sum of the squares of the factor charge (Li) for each construct divided by the sum of the number of factor load squares plus the number of error variances of a construct (ei).

Based on the results of the calculation of the reliability of the construct in Table 6, it can be concluded that the construct reliability of the creative thinking skill rubric is good (CR ≥ 0.70), with most factors, such as Solution Feasibility (CR = 0.789) and Original Solutions (CR = 0.818), demonstrating strong internal consistency. However, the Solution Variety factor exhibits marginal reliability (CR = 0.708). The overall reliability of the rubric was deemed acceptable, considering the strong validity and consistency of the other factors. CFA is carried out to estimate the load of variable factors. The factor charge represents the level of the regression path from the latent variable to the indicator. The CR level serves as an alternative guide for evaluating convergent validity (Ghadi et al., 2012).

The present study pursued two main objectives: to conduct a theoretical validation of an analytic rubric for divergent thinking through expert judgment and inter-rater reliability, and to provide empirical validation using exploratory and confirmatory factor analyses. Both aims were achieved. Expert evaluations confirmed that the rubric descriptors were clear, relevant, and appropriately constructed, while inter-rater reliability indicated fair to good agreement among scorers. Empirical testing through EFA and CFA further supported a coherent structure, strong model fit, and acceptable internal consistency. Taken together, these findings demonstrate that the rubric fulfills both conceptual and psychometric standards, establishing it as a dependable tool for assessing creative thinking in pre-service biology education.

Our findings align with recent research on rubric-based assessment instruments that have demonstrated robust psychometric properties, including high CFI and favorable model fits. For instance, Amelia et al. (2024) developed and validated a rubric for measuring science experiment design skills among prospective science teachers; it achieved strong construct validity with GFI = 0.94, NFI = 0.99, and CFI = 1.00, as well as RMSEA = 0.071, indicating high measurement coherence. Similarly, a study in simulation-based peer assessment for medical education reported high inter-rater reliability and clear performance criteria, supporting the effectiveness of rigorously designed rubrics in yielding consistent and valid assessments (Lertsakulbunlue & Kantiwong, 2024). However, unlike other investigations where divergent-thinking facets often loaded inconsistently across multiple factors (Beghetto & Karwowski, 2017; Said-Metwaly et al., 2017), our study revealed a clear unidimensional structure. This supports the view that divergent-thinking tasks in science education frequently converge into a general creative factor, confirming the argument that domain-specific tasks may activate creativity in a more integrated manner.

A particularly compelling finding from our study is that the feasibility dimension consistently showed stronger reliability than the variety dimension. This aligns with theoretical insights indicating that while variety often overlaps with fluency and originality, reducing its unique measurement variance, feasibility captures the appropriateness and practicality of solutions—an aspect that tends to be rated more consistently across scorers. Recent creativity research reinforces this distinction. Chan and Schunn (2023) demonstrate that feasibility is empirically separable from impact and can be reliably assessed, even when the constructs are conceptually close, providing clearer insights into creative performance. Similarly, Kern et al. (2024) highlight that creativity assessments that include feasibility alongside novelty and value tend to yield more reliable and meaningful results, as feasibility offers tangible anchors for evaluators, especially in applied and STEM contexts.

Furthermore, distinct from earlier studies that often emphasized content validation through expert judgment alone (Schilling et al., 2007; Connell et al., 2018; Elangovan & Sundaravel, 2021), this study implemented a two-tier validation approach that combined expert input with empirical analysis through exploratory and confirmatory factor analysis (EFA and CFA). This methodological integration reinforces the construct validity and structural coherence of the instrument.

Furthermore, this rubric was purposefully designed for use in human physiology tasks that simulate authentic real-world scenarios. This aligns with Guilford's (1959) claim that situational stimuli can drive behavioural changes. Guilford also stated that creative actions reflect learning outcomes, showcasing behavioural shifts triggered by stimuli and responses. This represents a departure from previous rubrics that were adapted without empirical revalidation (Lee et al., 2020). By embedding situational stimuli, the current instrument more effectively captures the critical components of divergent thinking, particularly fluency, flexibility, and originality. These findings align with recent literature emphasizing the importance of authentic assessment contexts in enhancing validity in creativity research (Karunarathne & Calma, 2024).

A notable innovation in this study lies in the use of student-generated responses during rubric development. By integrating actual student outputs into the descriptor formulation stage, the rubric offers clearer performance anchors and reduces ambiguity in interpretation. This strategy, which was seldom applied in earlier rubric designs, is supported by recent assessment literature (Scanlon et al., 2023; Elkington & Chesterton, 2025), recommending that assessment tools be contextualized within the learner experience to enhance validity.

The study also addresses inter-rater reliability, a long-standing challenge in creativity assessment. The application of a four-point scale with precise descriptors yielded fair to good levels of agreement across all measured dimensions. While previous research has highlighted variability in scoring as a limitation (Barth & Stadtmann, 2021), this study demonstrates that systematic descriptor development can lead to improved consistency in scoring.

Applying rubrics in educational settings presents several practical challenges that educators must navigate to ensure the effectiveness of assessments and learning outcomes. One of the significant challenges is the perception and understanding of rubrics among educators. Research indicates that if teachers do not recognize rubrics as tools to improve teaching practices, the use of rubrics might not result in the intended educational impact (Mui So & Hoi Lee, 2011). This highlights the necessity for professional development, which emphasizes the pedagogical benefits of the rubrics. Moreover, ensure that educators view the rubrics as integral to the learning process, rather than merely as compliance tools.

Additionally, the design and clarity of rubrics are crucial for their successful implementation. Studies have shown that rubrics must possess good validity and reliability to be effective (Connell et al., 2018; Ramazanzadeh et al., 2023). However, many existing rubrics are adaptations of previous models rather than being empirically tested, which can lead to issues of construct validity and misinterpretation of assessment criteria (Lee et al., 2020). This highlights the importance of developing rubrics with careful consideration of their psychometric properties, ensuring both aspects are clear and appropriate for the specific educational context (Mrangu, 2022).

Furthermore, involving students in the rubric development process can enhance their effectiveness. Engaging students in creating rubrics not only clarifies expectations but also fosters a sense of ownership over their learning (Nsabayezu et al., 2022). Therefore, this participatory approach can mitigate resistance to rubric use and promote a culture of feedback and improvement within educational settings.

These findings have important implications for curriculum designers, educators, and policymakers. Analytic rubrics developed through a psychometrically sound process can facilitate transparent evaluation and informed pedagogical decisions. As Jönsson and Panadero (2017) and Pancorbo et al. (2020) noted, rubrics that are theoretically sound and contextually grounded contribute meaningfully to both formative and summative assessment practices.

Looking ahead, future research could refine rubrics through cross-cultural validation, which is essential for ensuring their applicability across diverse educational contexts. As educational systems vary significantly worldwide, it is essential to evaluate how rubrics function in diverse cultural contexts. This could involve comparative studies that examine the effectiveness of rubrics in various educational frameworks by considering factors such as local pedagogical practices, student demographics, and cultural attitudes toward assessment (Nkhoma et al., 2020). Such research would not only contribute to the robustness of rubric design but also facilitate the development of universally applicable assessment tools that respect cultural nuances.

Moreover, future studies should focus on the iterative process of rubric development by incorporating feedback from both educators and students to enhance clarity and usability (Reddy & Andrade, 2010). By employing systematic methodologies to test the validity and reliability of rubrics in diverse contexts, researchers can contribute to a more nuanced understanding of how these tools can be optimized for various educational environments. Leveraging digital platforms for rubric-based scoring could enhance efficiency and scalability. Overall, this study provides a well-founded, empirically validated model for assessing creative thinking that bridges the gap between theory and classroom practice in science education.

CONCLUSION

This study has demonstrated that the analytic rubric developed for assessing divergent thinking skills in pre-service biology teachers is both valid and reliable. Expert reviewers confirmed that the descriptors were clear, relevant, and well-constructed, while inter-rater evaluations showed fair to good consistency. Empirical validation through EFA and CFA further supported a coherent structure, strong model fit, and acceptable reliability. A key finding is that the feasibility dimension proved to be more stable than variety, highlighting the importance of appropriateness as a meaningful indicator of creativity in applied biology contexts. Theoretically, these results strengthen the view that divergent thinking operates as a general factor, while also offering fresh insight into how feasibility contributes to the broader construct of creativity. Practically, the rubric equips educators and teacher trainers with a dependable framework for formative assessment, enabling them to provide constructive feedback on fluency, originality, feasibility, and variety. Its application is closely aligned with the *Kurikulum Merdeka* and responds to the urgent need to enhance students' creative thinking abilities, as reflected in Indonesia's performance on PISA 2022. For curriculum and policy development, the rubric provides a transparent way to integrate creativity into biology education, ensuring it is not only promoted as an ideal but also assessed fairly and consistently. Moving forward, future research should extend this validation to different institutions and cultural contexts, explore group-based differences, and test predictive validity in relation to authentic teaching and learning outcomes. The use of digital scoring platforms also holds promise for making this tool more efficient, scalable, and accessible in diverse educational settings.

REFERENCES

Abdellatif, R., & El-Wakeel, H. (2025). Assessing creative outcomes in studio-based learning: a comparative assessment of analytical rubrics.

International Journal of Design Creativity and Innovation, *13*(1), 41–66.

Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, *45*(1), 131–142.

Alajami, A. (2020). Beyond originality in scientific research: Considering relations among originality, novelty, and ecological thinking. *Thinking Skills and Creativity*, *38*, 100723.

Amelia, R. N., Listiaji, P., Dewi, N. R., Heriyanti, A. P., Atmaja, B. D., Shoba, T. M., & Sajidi, I. (2024). Developing and Validating a Rubric for Measuring Skills in Designing Science Experiments for Prospective Science Teachers. *Jurnal Inovasi Pendidikan IPA*, *10*(1), 32–46.

Anderson, R. C., & Graham, M. (2021). Creative potential in flux: The leading role of originality during early adolescent development. *Thinking Skills and Creativity*, *40*, 100816.

Arabacı, D., & Baki, A. (2023). An analysis of the gifted and non-gifted students creativity within the context of problem-posing activity. *Journal of Pedagogical Research*.

Asli, N. F., Matore, M. E. E. M., & Yunus, M. M. (2024). Construct validity of primary trait writing rubrics based on assessment use argument (AUA) validation framework. *Heliyon*, *10*(22), e40053.

Azmi, C., Hadiyanto, H., & Rusdinal, R. (2023). National Curriculum Education Policy "Curriculum Merdeka And Its Implementation." *International Journal of Educational Dynamics*, *6*(1), 303–309.

Azwar, S. (2012). *Reliabilitas dan Validitas* (4th ed.). Pustaka Pelajar. https://pustakapelajar.co.id/product/reliabilitas-dan-validitas/

Baer, J. (2012). Domain Specificity and the Limits of Creativity Theory. *The Journal of Creative Behavior*, *46*(1), 16–29.

Barak, M., & Levenberg, A. (2016). Flexible thinking in learning: An individual differences measure for learning in technology-enhanced environments. *Computers & Education*, *99*, 39–52.

Barth, P., & Stadtmann, G. (2021). Creativity assessment over time: Examining the reliability of cat ratings. *Journal of Creative Behavior*, *55*(2), 396–409.

Beghetto, R. A., & Karwowski, M. (2017). Toward Untangling Creative Self-Beliefs. In *The Creative Self* (pp. 3–22). Elsevier.

Bollen, K. A., & Long, J. S. (1993). *Testing Structural Equation Models*. Sage. https://books.google.co.id/books/about/Testing_Structural_Equation_Models.html?id=FvIxxeYDLx4C&redir_esc=y

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin*, *107*(2), 260–273.

Brookhart, S. M. (2018). Appropriate Criteria: Key to Effective Rubrics. *Frontiers in Education*, *3*, 22.

Browne, M. W., & Cudeck, R. (1993). Alternative Ways of Assessing Model Fit. In *Testing Structural Equation Models* (Vol. 154). SAGE Publications, Inc. https://us.sagepub.com/en-us/nam/testing-structural-equation-models/book3893#contents

Chan, J., & Schunn, C. D. (2023). The Importance of Separating Appropriateness into Impact and Feasibility for the Psychology of Creativity. *Creativity Research Journal*, *35*(4), 629–644.

Chowdhury, F. (2018). Application of rubrics in the classroom: A vital tool for improvement in assessment, feedback and learning. *IES*, *12*(1), 61.

Connell, J., Carlton, J., Grundy, A., Taylor Buck, E., Keetharuth, A. D., Ricketts, T., Barkham, M., Robotham, D., Rose, D., & Brazier, J. (2018). The importance of content and face validity in instrument development: lessons learnt from service users when developing the Recovering Quality of Life measure (ReQoL). *Quality of Life Research*, *27*(7), 1893–1902.

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *J Sci Educ Technol*, *32*(3), 444–452.

DiStefano, C., & Hess, B. (2005). Using Confirmatory Factor Analysis for Construct Validation: An Empirical Review. *Journal of Psychoeducational Assessment*, *23*(3), 225–241.

Doll, W. J., Xia, W., & Torkzadeh, G. (1994). A confirmatory factor analysis of the end-user computing satisfaction instrument. *MIS Quarterly*, *18*(4), 453.

Elangovan, N., & Sundaravel, E. (2021). Method of preparing a document for survey instrument validation by experts. *MethodsX*, *8*, 101326.

Elkington, S., & Chesterton, P. (2025). Embedding assessment flexibilities for future authentic learning. *Teaching in Higher Education*, *30*(3), 700–716.

Elosua, P. (2022). Validity evidences for scoring procedures of a writing assessment task. A case study on consistency, reliability, unidimensionality and prediction accuracy. *Assessing Writing*, *54*, 100669.

Embretson, S. E., & Reise, S. P. (2013). *Item Response Theory* (0 ed.). Psychology Press.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical Methods for Rates and Proportions* (1st ed.). Wiley.

Forthmann, B., Paek, S. H., Dumas, D., Barbot, B., & Holling, H. (2020). Scrutinizing the basis of originality in divergent thinking tests: On the measurement precision of response propensity estimates. *British Journal of Educational Psychology*, *90*(3), 683–699.

Garcimartín, C. F., Pastor, V. M. L., Nieto, T. F., & Alcalá, D. H. (2024). Creating Assessment Rubrics for Final Teacher Education Degree Projects: A Qualitative Case Study. *The Qualitative Report*.

Gerbing, D. W., & Anderson, J. C. (1992). Monte Carlo Evaluations of Goodness of Fit Indices for Structural Equation Models. *Sociological Meth-*

ods & Research, *21*(2), 132–160.

Ghadi, I., Alwi, N. H., Abu Bakar, K., & Talib, O. (2012). Construct validity examination of critical thinking dispositions for undergraduate students in University Putra Malaysia. *HES*, *2*(2), p138.

Gorsuch, R. L. (2013). *Factor Analysis* (0 ed.). Psychology Press.

Goudarzian, A. H. (2023). Challenges and recommendations of exploratory and confirmatory factor analysis: A narrative review from a nursing perspective. *Journal of Nursing Reports in Clinical Practice*, *1*(3), 133–137.

Guilford, J. P. (1959). Three faces of intellect. *American Psychologist*, *14*(8), 469–479.

Gunawan, Ferdianto, F., Mulyatna, F., & Untarti, R. (2025). The profile of creative thinking process: Prospective mathematics teachers. *Jurnal Eduscience (JES)*, *2*(12). https://jurnal.ulb.ac.id/index.php/eduscience/article/view/6915

Hadi, S. (2001). *Metodologi Research Jilid III*. Andi. https://onesearch.id/Record/IOS2726.slims-67126?widget=1

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2013). Partial Least Squares Structural Equation Modeling: Rigorous Applications, Better Results and Higher Acceptance. *Long Range Planning*, *46*(1–2), 1–12.

Hammitt, J. K., & Zhang, Y. (2013). Combining experts' judgments: Comparison of algorithmic methods using synthetic data. *Risk Analysis*, *33*(1), 109–120.

Han, C., Zheng, B., Xie, M., & Chen, S. (2024). Raters' scoring process in assessment of interpreting: an empirical study based on eye tracking and retrospective verbalisation. *The Interpreter and Translator Trainer*, *18*(3), 400–422.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and ltenls. *Applied Psychological Measurement*, *9*(2), 139–164.

Hayati, K., Ulfa Tenri Pada, A., & Mawarpury, M. (2023). Content validity of collective efficacy questionnaire for natural disasters based on Aceh Local wisdom. *E3S Web Conf.*, *447*, 4004.

Hettithanthri, U., Hansen, P., & Munasinghe, H. (2023). Exploring the architectural design process assisted in conventional design studio: a systematic literature review. *International Journal of Technology and Design Education*, *33*(5), 1835–1859.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453.

Imbler, A. C., Clark, S. K., Young, T. A., & Feinauer, E. (2023). Teaching second-grade students to write science expository text: Does a holistic or analytic rubric provide more meaningful results? *Assessing Writing*, *55*, 100676.

Isbell, T., & Goomas, D. T. (2014). Computer-assisted rubric evaluation: Enhancing outcomes and assessment quality. *Community College Journal of Research and Practice*, *38*(12), 1193–1197.

Jönsson, A., & Panadero, E. (2017). The Use and Design of Rubrics to Support Assessment for Learning. In D. Carless, S. M. Bridges, C. K. Y. Chan, & R. Glofcheski (Eds.), *Scaling up Assessment for Learning in Higher Education* (Vol. 5, pp. 99–111). Springer Singapore.

Karunarathne, W., & Calma, A. (2024). Assessing creative thinking skills in higher education: deficits and improvements. *Studies in Higher Education*, *49*(1), 157–177.

Kern, F. B., Wu, C., & Chao, Z. C. (2024). Assessing novelty, feasibility and value of creative ideas with an unsupervised approach using GPT-4. *British Journal of Psychology*.

Kind, P. M., & Kind, V. (2007). Creativity in science education: Perspectives and challenges for developing school science. *Studies in Science Education*, *43*(1), 1–37.

Koswara, D., Dallyono, R., Suherman, A., & Hyangsewu, P. (2021). The analytical scoring assessment usage to examine Sundanese students' performance in writing descriptive texts. *CP*, *40*(3), 573–583.

Lange, R. T. (2011). Inter-rater Reliability. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of Clinical Neuropsychology* (p. 1348). Springer New York.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563–575.

Lee, J. E., Recker, M., & Yuan, M. (2020). The Validity and Instructional Value of a Rubric for Evaluating Online Course Quality: An Empirical Study. *Online Learning*, *24*(1).

Legendre, P. (2005). Species associations: the Kendall coefficient of concordance revisited. *JABES*, *10*(2), 226–245.

Lertsakulbunlue, S., & Kantiwong, A. (2024). Development of peer assessment rubrics in simulation-based learning for advanced cardiac life support skills among medical students. *Advances in Simulation*, *9*(1), 25.

Marzano, G. (2022). *Sustaining Creativity and the Arts in the Digital Age*. IGI Global.

Mezo, P. G., & Short, M. M. (2012). Construct validity and confirmatory factor analysis of the Self-Control and Self-Management Scale. *Canadian Journal of Behavioural Science / Revue Canadienne Des Sciences Du Comportement*, *44*(1), 1–8.

Mooi, E., & Sarstedt, M. (2011). *A Concise Guide to Market Research*. Springer Berlin Heidelberg.

Morris, G., & Sharplin, E. (2013). The assessment of creative writing in Senior Secondary English: A colloquy concerning criteria. *English in Education*, *47*(1), 49–65.

Mrangu, L. (2022). Rubric as assessment tool for lecturers and students in higher education institution. *Acta Pedagogia Asia*, *1*(1), 26–33.

Mui So, W. W., & Hoi Lee, T. T. (2011). Influence of teachers' perceptions of teaching and learning on the implementation of Assessment for

Learning in inquiry study. *Assessment in Education: Principles, Policy & Practice*, *18*(4), 417–432.

Natalya, L., & Purwanto, C. V. (2018). Exploratory and Confirmatory Factor Analysis of the Academic Motivation Scale (AMS)–Bahasa Indonesia. *Makara Human Behavior Studies in Asia*, *22*(1), 29.

Neupane, S. M., & Bhattarai, P. C. (2024). Constructing the scale to measure entrepreneurial traits by using the modified delphi method. *Heliyon*, *10*(7), e28410.

Nkhoma, C., Nkhoma, M., Thomas, S., & Quoc Le, N. (2020). *The Role of Rubrics in Learning and Implementation of Authentic Assessment: A Literature Review*. 237–276.

Nsabayezu, E., Iyamuremye, A., Mukiza, J., Habimana, J. C., Mbonyiryivuze, A., Gakub, E., Nsengimana, T., & Niyonzima, F. N. (2022). Teachers' and students' perceptions towards the utilization of formative assessment rubric for supporting students' learning of organic chemistry. *Journal of Educational Sciences*, *45*(1), 124–134.

OECD. (2024). *Pisa 2022 Results (Volume III): Factsheets – Indonesia*. https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/06/pisa-2022-results-volume-iii-country-notes_72b418f8/indonesia_cf276198/a7090b49-en.pdf?utm_source=chatgpt.com

Olson, J. M., & Krysiak, R. (2021). *Rubrics as Tools for Effective Assessment of Student Learning and Program Quality* (pp. 173–200).

Pada, A. U. T., Kartowagiran, B., & Subali, B. (2015). Content validity of creative thinking skills assessment. *Proceeding of International Conference On Research, Implementation And Education Of Mathematics And Sciences*. https://core.ac.uk/download/pdf/33519344.pdf

Pada, A. U. T., Kartowagiran, B., & Subali, B. (2016). Separation index and fit items of creative thinking skills assessment. *REiD*, *2*(1), 1–12.

Pada, A. U. T., Mustakim, S. S., & Subali, B. (2018). Construct validity of creative thinking skills instrument for biology student teachers in the subject of human physiology. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(2), 119–129.

Panadero, E., Delgado, P., Zamorano, D., Pinedo, L., Fernández-Ortube, A., & Barrenetxea-Mínguez, L. (2025). Putting excellence first: How rubric performance level order and feedback type influence students' reading patterns and task performance. *Learning and Instruction*, *99*, 102168.

Pancorbo, G., Primi, R., John, O. P., Santos, D., Abrahams, L., & De Fruyt, F. (2020). Development and psychometric properties of rubrics for assessing social-emotional skills in youth. *Studies in Educational Evaluation*, *67*, 100938.

Ramazanzadeh, N., Ghahramanian, A., Zamanzadeh, V., Valizadeh, L., & Ghaffarifar, S. (2023). Development and psychometric testing of a clinical reasoning rubric based on the nursing process. *BMC Med Educ*, *23*(1), 98.

Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. *Journal of Educational Statistics*, *4*(3), 207–230.

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*(4), 435–448.

Rosnawati, R., Kartowagiran, B., & Jailani, J. (2015). A formative assessment model of critical thinking in mathematics learning in junior high school. *REiD*, *1*(2), 186–198.

Runco, M. A. (1985). Reliability and convergent validity of ideational flexibility as a function of academic achievement. *Percept Mot Skills*, *61*(3_suppl), 1075–1081.

Runco, M. A., & Alabbasi, A. M. A. (2024). Interactions among dimensions of divergent thinking as predictors of creative activity and accomplishment. *Thinking Skills and Creativity*, *53*, 101583.

Said-Metwaly, S., Noortgate, W. Van den, & Kyndt, E. (2017). Approaches to Measuring Creativity: A Systematic Literature Review. *Creativity. Theories – Research - Applications*, *4*(2), 238–275.

Scanlon, D., MacPhail, A., Walsh, C., & Tannehill, D. (2023). Embedding assessment in learning experiences: enacting the principles of instructional alignment in physical education teacher education. *Curriculum Studies in Health and Physical Education*, *14*(1), 3–20.

Schilling, L. S., Dixon, J. K., Knafl, K. A., Grey, M., Ives, B., & Lynn, M. R. (2007). Determining content validity of a self-report instrument for adolescents using a heterogeneous expert panel. *Nursing Research*, *56*(5), 361–366.

Shafiei, S. (2024). A proposed analytic rubric for consecutive interpreting assessment: implications for similar contexts. *Language Testing in Asia*, *14*(1), 13.

Shook, C. L., Ketchen, D. J., Hult, G. T. M., & Kacmar, K. M. (2004). An assessment of the use of structural equation modeling in strategic management research. *Strategic Management Journal*, *25*(4), 397–404.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *1*(26), 100–107.

Sireci, S. G. (1995). The central role of content representation in test validity. *The Construct of Content Validity: Theories and Applications*. https://files.eric.ed.gov/fulltext/ED387508.pdf

Stevens, D. D., & Levi, A. J. (2005). *Introduction to Rubrics: An Assessment Tool to Save Grading Time, Convey Effective Feedback and Promote Student Learning*. Stylus Publishing, LLC. https://eric.ed.gov/?id=ED515062

Subali, B., & Suyata, P. (2013). Standardisasi penilaian berbasis sekolah. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *17*(1), 1–18.

Sudaryanto, M., & Akbariski, H. S. (2021). Students' competence in making language skill assessment rubric. *REiD*, *7*(2), 156–167.

Sumekto, D. R., & Setyawati, H. (2018). Students' Descriptive Writing Performance: The Analytic Scoring Assessment Usage. *CP*.

Sureeyatanapas, P., Sureeyatanapas, P., Panitanarak, U., Kraisriwattana, J., Sarootyanapat, P., & O'Connell, D. (2024). The analysis of marking reliability through the approach of gauge repeatability and reproducibility (GR&amp;R) study: a case of English-speaking test. *Language Testing in Asia*, *14*(1), 1.

Thakral, P. P., Yang, A. C., Addis, D. R., & Schacter, D. L. (2021). Divergent thinking and constructing future events: dissociating old from new ideas. *Memory*, *29*(6), 729–743.

van Dalen, D. B. (1973). *Understanding Educational Research: An Introduction*. McGraw-Hill. https://books.google.co.id/books/about/Understanding_Educational_Research.html?id=r1lbngEACAAJ&redir_esc=y

Wang, C., Zhang, M., Sesunan, A., & Yolanda, L. (2023). *Driving education reform in Indonesia through technology: Exploring the current status of the Merdeka Belajar program*. Oliver Wyman. https://www.oliverwyman.com/our-expertise/insights/2023/dec/technology-driven-education-reform-indonesia.html?utm_source=chatgpt.com

Wang, Y., & Hou, Q. (2018). Insight or Originality: A Spray in the River of Creative Thinking. *OALib*, *05*(09), 1–6.

Weiss, S., & Wilhelm, O. (2022). Is Flexibility More than Fluency and Originality? *Journal of Intelligence*, *10*(4), 96.

Wiersma, W. (2000). *Research Methods in Education: An Introduction* (7th ed.). Allyn and Bacon. https://books.google.co.id/books/about/Research_Methods_in_Education.html?id=MAUmAQAAIAAJ&redir_esc=y

Williams, R. L. (1999). Operational definitions and assessment of higher-order cognitive constructs. *Educational Psychology Review*, *11*(4), 411–427.

Yildiz, C., & Yildiz, T. G. (2021). Exploring the relationship between creative thinking and scientific process skills of preschool children. *Thinking Skills and Creativity*, *39*, 100795.