



Item Quality Analysis of the Engineering Design Process Skills and Creativity Instruments in Biology Learning

Mulikah, *Ary Susatyo Nugroho, Endah Rita Sulistya Dewi

Biology Education Department, Universitas PGRI Semarang, Semarang, Indonesia

*Corresponding Author e-mail: arysusatyo@upgris.ac.id

Received: August 2026; Revised: September 2026; Published: October 2026

Abstract

Accurate assessment of Engineering Design Process Skills (EDPS) and creativity in Biology learning requires instruments with adequate empirical item characteristics. This study aimed to examine the item quality and internal consistency of EDPS and creativity assessment instruments within the context of the Human Respiratory System. A quantitative descriptive design was employed involving 67 eleventh-grade students from SMA Negeri 1 Mijen, Indonesia. The instruments comprised 14 EDPS items and eight creativity items, each using a 0–4 scoring scale. Data were analyzed using descriptive statistics, item difficulty, corrected item-total correlation, discrimination power, and Cronbach's alpha. The findings showed that 20 of the 22 items (90.91%) met the adopted item-total correlation criterion, comprising 13 EDPS items and seven creativity items. Of these, 19 items (86.36% of all items) also demonstrated satisfactory discrimination and were considered suitable for retention. EDPS_9 met the item-total correlation criterion but required revision because of its relatively low discrimination index (0.292), whereas EDPS_13 and Creativity_8 failed to meet the criterion and demonstrated very low discrimination. Twenty-one items were classified as moderately difficult, while one EDPS item was categorized as easy. The EDPS and creativity instruments demonstrated high internal consistency, with Cronbach's alpha coefficients of 0.933 and 0.869, respectively. These findings provide preliminary empirical evidence supporting the retention of most assessment items while emphasizing the need for targeted revisions and further psychometric evaluation before broader implementation.

Keywords: Assessment instrument; Biology learning; Engineering design process skills; Human respiratory system; Item analysis

How to Cite: Mulikah, M., Nugroho, A. S., & Dewi, E. R. S. (2026). Item Quality Analysis of the Engineering Design Process Skills and Creativity Instruments in Biology Learning. *Prisma Sains: Jurnal Pengkajian Ilmu Dan Pembelajaran Matematika Dan IPA IKIP Mataram*, 14(4), 2301–2319. <https://doi.org/10.33394/j-ps.v14i4.22308>



<https://doi.org/10.33394/j-ps.v14i4.22308>

Copyright© 2026, Mulikah et al.

This is an open-access article under the [CC-BY License](#)



INTRODUCTION

The increasing demands of twenty-first-century education have shifted the emphasis of learning from the acquisition of factual knowledge toward the development of competencies that enable students to apply knowledge, address complex problems, and generate meaningful solutions. Critical thinking, creativity, problem-solving, communication, and collaboration are increasingly recognized as essential competencies for participating effectively in contemporary educational and professional environments (Mardhiya et al., 2025; Mardhiyah et al., 2021; Thornhill-Miller et al., 2023). These competencies require learning experiences that encourage students to investigate problems, construct explanations, evaluate alternative ideas, and apply their understanding to authentic situations. Consequently, educational assessment must extend beyond the measurement of conceptual knowledge to capture students' abilities to engage in complex cognitive and problem-solving processes. The availability of appropriate assessment instruments is therefore essential for obtaining meaningful evidence of students' competencies and supporting informed decisions about teaching and learning.

Biology learning provides a relevant context for developing and assessing higher-order

competencies because biological concepts are closely connected to natural phenomena, human activities, health issues, and environmental challenges. Students are expected not only to understand biological structures and processes but also to interpret scientific information, identify problems, formulate explanations, and propose solutions based on biological principles. Learning that connects disciplinary knowledge with authentic situations can provide opportunities for students to apply scientific understanding in meaningful contexts (Herrington et al., 2014; Wulandari et al., 2021). However, the assessment of these competencies requires instruments capable of representing the cognitive processes involved rather than merely measuring students' ability to recall biological facts. This consideration is particularly relevant when Biology learning incorporates engineering-oriented problem-solving activities that require students to transform scientific knowledge into proposed designs or practical solutions.

Engineering Design Process Skills (EDPS) represent competencies associated with the systematic identification, investigation, and resolution of problems through engineering design activities. The engineering design process generally involves identifying problems, recognizing needs and constraints, generating alternative ideas, designing potential solutions, evaluating outcomes, and making improvements based on the results of evaluation (Aini et al., 2024; Lin et al., 2021). Unlike conventional tasks that emphasize predetermined answers, engineering design activities encourage students to consider multiple possible solutions, justify design decisions, and refine proposed solutions in response to particular requirements and limitations. These characteristics position EDPS as a relevant competency within interdisciplinary science education, particularly in learning environments that integrate science, technology, engineering, arts, and mathematics (STEAM). Such integration provides opportunities to connect conceptual understanding with design-oriented problem-solving and the application of scientific knowledge to contextual challenges (Erol et al., 2023; Quigley et al., 2017).

Creativity is closely associated with engineering design because the development of meaningful solutions requires students to generate, evaluate, elaborate, and refine ideas. Creativity in educational problem-solving is not limited to producing unusual or novel responses but also involves developing ideas that are relevant, appropriate, and potentially useful within a particular context. The processes of exploring different perspectives, generating alternatives, and refining solutions are therefore important components of creative engagement in design activities (Chang et al., 2022; Murray et al., 2019). In educational assessment, creativity can be represented through several dimensions, including fluency, flexibility, originality, and elaboration. Fluency concerns the generation of multiple relevant ideas, flexibility reflects variation in approaches or categories of ideas, originality refers to the novelty of responses, and elaboration involves developing ideas through additional detail and explanation. Although creativity contributes to engineering design activities, EDPS and creativity represent conceptually distinguishable constructs. EDPS emphasizes the systematic processes involved in developing and evaluating solutions, whereas creativity emphasizes the characteristics and development of ideas generated during these processes. Accordingly, both competencies require assessment indicators that reflect their respective conceptual characteristics.

The Human Respiratory System provides a meaningful biological context for assessing EDPS and creativity among senior high school students. This topic encompasses the structure and function of respiratory organs, mechanisms of ventilation, gas exchange, and disorders affecting respiratory health (Haddad & Sharma, 2026). These concepts can be connected to contextual problems involving respiratory health, environmental conditions, and the development of preventive or practical solutions. For example, students may be required to identify problems associated with respiratory function, consider relevant biological principles, propose alternative solutions, and evaluate the feasibility of their designs. Such tasks can elicit evidence of students' engineering design reasoning while also providing opportunities to

demonstrate fluency, flexibility, originality, and elaboration. Nevertheless, the relevance of a biological context does not automatically guarantee that assessment items adequately measure the intended competencies. The quality of the resulting measurements depends on the alignment between assessment tasks, construct indicators, scoring procedures, and students' observed responses.

Developing appropriate instruments for assessing EDPS and creativity requires careful consideration of both conceptual representation and empirical item characteristics. Assessment instruments are intended to produce scores that support interpretations of students' competencies; therefore, weaknesses in item construction or scoring may compromise the accuracy and usefulness of those interpretations (Baumert et al., 2009; DeVon et al., 2007). Constructed-response instruments, including essay-type assessments, can provide opportunities for students to explain reasoning, generate alternatives, justify decisions, and elaborate proposed solutions. These response characteristics are relevant to the assessment of engineering design and creativity because the competencies involve processes that may not be adequately represented by selecting predetermined response options. However, constructed-response formats also introduce measurement challenges associated with task clarity, differences in students' interpretation of prompts, variation in response complexity, and the consistency of scoring decisions. Consequently, empirical item analysis is necessary to evaluate whether individual items function adequately within the intended assessment context.

Item analysis provides information about the empirical characteristics of assessment tasks and supports decisions regarding item retention, revision, or replacement. Important parameters include item difficulty, item-total correlation, and discrimination power. For polytomously scored constructed-response items, item difficulty can be represented by the proportion of the mean item score to the maximum possible score, providing an indication of students' achievement on each task. Corrected item-total correlation reflects the association between performance on an individual item and the total score of the remaining items, while discrimination power indicates the extent to which an item differentiates students according to their measured performance (Maulina & Noviriany, 2020; Mustaqim & Sulisti, 2024). These parameters provide complementary information because an item may exhibit an appropriate difficulty level while demonstrating a weak relationship with the overall instrument or limited discriminatory capacity. Therefore, decisions about item quality should consider multiple empirical indicators rather than relying on a single statistical criterion.

In addition to item-level characteristics, internal consistency is an important consideration in evaluating assessment instruments. Cronbach's alpha is commonly employed to estimate the consistency of scores across items intended to measure a particular construct (Ekolu & Quainoo, 2019; Kimberlin & Winterstein, 2008). Nevertheless, a high internal consistency coefficient does not necessarily demonstrate that every item functions adequately or that the instrument comprehensively represents its intended construct. Reliability evidence should therefore be interpreted alongside item-level findings and the conceptual alignment of assessment content (DeVon et al., 2007; Geldhof et al., 2014). This distinction is particularly important for EDPS and creativity instruments because both competencies encompass multiple indicators involving different cognitive processes. A comprehensive interpretation of instrument quality requires attention to the performance of individual items, the consistency of the resulting scores, and the limitations of the evidence provided by the selected analyses.

Previous research has demonstrated the educational relevance of engineering design and STEAM-oriented learning for supporting creativity, problem-solving, and design-related competencies (Erol et al., 2023; Ozkan & Topsakal, 2021; Thammaariyasakun et al., 2025). Other studies have examined the development of assessment instruments and highlighted the importance of evaluating item characteristics to improve measurement quality (Ghorbiy et al., 2026; Long et al., 2022; Mustaqim & Sulisti, 2024). However, these research directions do not necessarily provide sufficiently specific evidence regarding the empirical functioning of

constructed-response items designed to assess EDPS and creativity within the same Biology learning context. In particular, evidence concerning item difficulty, corrected item-total correlations, discrimination power, and internal consistency remains important for evaluating context-specific instruments intended to represent these competencies. Although general approaches to educational instrument evaluation are well established, further context-specific item analyses are needed to inform the refinement of assessments that represent engineering design processes and creative thinking in secondary Biology education.

The intended contribution of the present study is to examine the empirical quality of two complementary constructed-response instruments within a common biological assessment context. Rather than evaluating the effectiveness of a particular instructional intervention, this study focuses on the empirical functioning of individual assessment items designed to measure Engineering Design Process Skills and creativity among eleventh-grade students. The EDPS instrument represents six indicators: identifying problems, identifying needs and constraints, generating ideas, designing solutions, evaluating outcomes, and making improvements. Meanwhile, the creativity instrument represents four indicators: fluency, flexibility, originality, and elaboration. Through the analysis of item difficulty, corrected item-total correlation, discrimination power, and internal consistency, this study seeks to generate empirical evidence that can inform decisions regarding item retention, revision, and further evaluation. Its intended contribution is primarily methodological and context-specific, emphasizing preliminary evidence of item quality rather than comprehensive validation of the instruments.

Accordingly, this study aims to analyze the empirical quality of Engineering Design Process Skills and creativity assessment items in Biology learning on the Human Respiratory System. Specifically, the study examines the difficulty levels, corrected item-total correlations, discrimination power, and item-level validity classifications of 14 EDPS items and eight creativity items, together with the internal consistency of each instrument. The findings are expected to provide an empirical basis for refining assessment items, improving the quality of constructed-response measurement, and informing subsequent evaluation of EDPS and creativity instruments in secondary Biology education.

METHOD

This study employed a quantitative descriptive research design to examine the empirical quality of Engineering Design Process Skills (EDPS) and creativity assessment instruments in Biology learning. The investigation focused on evaluating the characteristics of individual assessment items and the internal consistency of each instrument using classical test theory-based analyses. The parameters examined included item difficulty, corrected item-total correlation, discrimination power, item-level validity classification, and instrument reliability. This approach was intended to provide preliminary empirical evidence regarding the functioning of assessment items and to identify items requiring retention, revision, or further evaluation. The study did not involve an experimental intervention or a comparison of instructional effectiveness.

The research was conducted at SMA Negeri 1 Mijen, Indonesia, involving eleventh-grade students aged 16–17 years. The target population comprised students distributed across ten Grade XI classes, of which five classes were identified as eligible based on their participation in Biology instruction relevant to the assessment context. Class-level random sampling was employed to select two classes from the five eligible classes. The selected classes were XI.1, consisting of 34 students, and XI.2, consisting of 33 students, resulting in a total sample of 67 students. Both classes followed the same Biology curriculum coverage concerning the Human Respiratory System. The selection of intact classes was intended to accommodate the existing organization of classroom instruction while obtaining student responses for preliminary item analysis. Because the participants were drawn from a single school, the resulting item characteristics were interpreted within the context of the participating sample rather than generalized to a broader student population.

Two constructed-response assessment instruments were employed to measure Engineering Design Process Skills and creativity. The EDPS instrument consisted of 14 essay-type items representing six indicators: identifying problems, identifying needs and constraints, generating ideas, designing solutions, evaluating outcomes, and making improvements. The creativity instrument comprised eight essay-type items representing four indicators: fluency, flexibility, originality, and elaboration. Both instruments were developed within the context of the Human Respiratory System, allowing students to demonstrate their understanding through written responses involving biological problems, alternative ideas, proposed solutions, and evaluative reasoning. Each item was scored using a scale ranging from 0 to 4, resulting in maximum possible scores of 56 for the EDPS instrument and 32 for the creativity instrument. The instruments were analyzed separately because EDPS and creativity represent conceptually distinguishable competencies, although both involve higher-order cognitive processes relevant to problem-solving and design activities. The distribution of assessment items across their respective indicators is presented in Table 1.

Table 1. Instrument Blueprint for Engineering Design Process Skills and Creativity

Construct	Indicator	Item Code	Item Number
EDPS	Identifying Problems	EDPS_1, EDPS_2	2
EDPS	Identifying Needs and Constraints	EDPS_3, EDPS_4	2
EDPS	Generating Ideas	EDPS_5, EDPS_6	2
EDPS	Designing Solutions	EDPS_7, EDPS_8, EDPS_9	3
EDPS	Evaluating Outcomes	EDPS_10, EDPS_11, EDPS_12	3
EDPS	Making Improvements	EDPS_13, EDPS_14	2
Creativity	Fluency	Creativity_1, Creativity_2	2
Creativity	Flexibility	Creativity_3, Creativity_4	2
Creativity	Originality	Creativity_5, Creativity_6	2
Creativity	Elaboration	Creativity_7, Creativity_8	2
Total			22

The EDPS instrument was intended to assess students' ability to identify biological problems, recognize relevant requirements and constraints, generate alternative ideas, design potential solutions, evaluate proposed outcomes, and suggest improvements. The creativity instrument was designed to capture students' abilities to produce multiple relevant ideas, consider different approaches, generate original responses, and elaborate their proposed ideas. The constructed-response format was selected to provide opportunities for students to express reasoning and develop responses beyond predetermined answer options. Both instruments employed a common biological context to maintain content relevance while preserving the conceptual distinction between the two measured constructs.

Data were collected by administering the EDPS and creativity instruments to all 67 participating students. Students completed the assessment items individually, and their written responses were evaluated according to the scoring criteria associated with the respective instruments. Each item was assigned a score ranging from 0 to 4, with higher scores indicating stronger performance in relation to the intended assessment criteria. Student responses were evaluated independently by two trained raters, consisting of the primary researcher and a senior Biology teacher. Before the scoring process, both raters participated in a four-hour calibration session to establish a shared understanding of the assessment indicators and scoring criteria. The resulting item scores were used for subsequent analyses of item characteristics and instrument reliability. EDPS and creativity scores were processed separately to ensure that the empirical evaluation reflected the intended construct of each instrument.

Data analysis was conducted using descriptive statistics and item analysis procedures, with statistical processing performed using IBM SPSS Statistics version 24. Descriptive statistics included mean scores, standard deviations, minimum scores, and maximum scores to summarize students' performance on the two instruments. Total instrument scores were

obtained by summing the scores of the respective items, while mean percentage scores were determined relative to the maximum possible scores. Item-level analysis was performed to examine difficulty, corrected item-total correlation, and discrimination power. These analyses were used to evaluate the empirical characteristics of each assessment item and support decisions regarding its suitability for further use.

Item difficulty was determined by comparing the mean score obtained by students on each item with the maximum possible item score of four. The resulting proportion represented the relative level of student performance on the item, with higher proportions indicating greater ease and lower proportions indicating greater difficulty. Based on the classification criteria adopted in the study, items with difficulty proportions below 0.30 were categorized as difficult, those ranging from 0.30 to 0.70 were classified as moderate, and those exceeding 0.70 were categorized as easy (Riduwan, 2010). These categories were used to describe the distribution of item difficulty across the EDPS and creativity instruments.

Item-level validity classification was examined using corrected item-total correlation, which represents the relationship between an individual item score and the total score of the remaining items within the same instrument. Excluding the focal item from the total score reduces the artificial inflation of correlations that may occur when an item is included in the total score against which it is correlated. A corrected item-total correlation coefficient greater than 0.30 was adopted as the criterion for classifying an item as meeting the empirical item-validity threshold. Items with coefficients at or below 0.30 were identified as requiring further examination or revision. This criterion was used as an operational indicator of item functioning and was not interpreted as sufficient evidence of comprehensive construct validity.

Item discrimination was examined to assess the extent to which individual items differentiated students according to their performance on the measured competency. Higher discrimination coefficients indicated a stronger capacity to distinguish students with different levels of performance, whereas lower coefficients suggested limited discriminatory capacity. The discrimination coefficients were interpreted using the classification criteria adapted from the framework applied in the original study (Arikunto, 2013). Items with discrimination coefficients ranging from 0.40 to 1.00 were categorized as good, those ranging from 0.30 to 0.39 were considered acceptable, those ranging from 0.20 to 0.29 were identified as requiring revision, and those ranging from 0.00 to 0.19 were classified as candidates for rejection or substantial redesign. The criteria used to interpret item discrimination are presented in Table 2.

Table 2. Classification Criteria for Item Discrimination Power

Discrimination Index (D)	Interpretation	Decision
0.40–1.00	Good	Retain
0.30–0.39	Acceptable	Retain with minor revision if needed
0.20–0.29	Low discrimination	Revise
0.00–0.19	Very low discrimination	Reject or redesign

Instrument reliability was evaluated using Cronbach's alpha coefficient to examine the internal consistency of item scores within each instrument. Reliability analysis was conducted separately for the 14-item EDPS instrument and the eight-item creativity instrument. Cronbach's alpha was selected because it provides an estimate of the extent to which items within an instrument produce internally consistent scores (Ekolu & Quainoo, 2019; Kimberlin & Winterstein, 2008). Higher coefficients were interpreted as indicating stronger internal consistency, although internal consistency alone was not regarded as sufficient evidence of instrument validity. Because the assessments employed constructed-response items evaluated by two raters, consistency between raters was recognized as a separate consideration from the internal consistency estimated through Cronbach's alpha.

The interpretation of instrument quality was based on the combined consideration of item difficulty, corrected item-total correlation, discrimination power, and internal consistency. Items demonstrating satisfactory empirical characteristics were considered candidates for

retention, whereas items with weak item-total relationships or limited discrimination were identified for revision, redesign, or further evaluation. The analysis emphasized that an appropriate level of item difficulty or a high instrument-level reliability coefficient does not necessarily indicate satisfactory functioning of every individual item. Consequently, item-level and instrument-level evidence were considered complementary when formulating recommendations for instrument refinement. The findings were interpreted as preliminary, context-specific evidence of assessment quality, recognizing that further evaluation involving additional validity evidence, scoring consistency, and more diverse participant samples would be necessary before broader application of the instruments.

RESULTS AND DISCUSSION

The empirical evaluation of the Engineering Design Process Skills (EDPS) and creativity instruments was conducted using assessment responses from 67 eleventh-grade students. The analyses examined item difficulty, corrected item-total correlation, discrimination power, item-level validity classification, and instrument reliability. The findings are presented separately for the two instruments, followed by descriptive statistics and an integrated interpretation of instrument quality. This approach provides information about the functioning of individual items and identifies areas requiring further refinement before subsequent assessment applications.

Item Analysis of the Engineering Design Process Skills (EDPS) Instrument

The item analysis of the EDPS instrument examined the empirical characteristics of 14 constructed-response items representing six engineering design indicators. The analysis focused on students' mean scores, difficulty proportions, corrected item-total correlations, and discrimination indices. These parameters were evaluated to determine whether individual items demonstrated satisfactory characteristics for assessing engineering design processes in the context of the Human Respiratory System. The results are presented in Table 3.

Table 3. Results of the Item Analysis of the Engineering Design Process Skills Instrument

Item	Mean	Difficulty Proportion	Difficulty Level	Item-Total Correlation	Validity	Discrimination Power	Decision
EDPS_1	3.150	0.787	Easy	0.824	Valid	0.597	Retain
EDPS_2	2.750	0.687	Moderate	0.822	Valid	0.653	Retain
EDPS_3	2.540	0.634	Moderate	0.772	Valid	0.653	Retain
EDPS_4	1.970	0.493	Moderate	0.778	Valid	0.694	Retain
EDPS_5	1.880	0.470	Moderate	0.782	Valid	0.653	Retain
EDPS_6	1.760	0.440	Moderate	0.810	Valid	0.653	Retain
EDPS_7	1.700	0.425	Moderate	0.682	Valid	0.472	Retain
EDPS_8	1.400	0.351	Moderate	0.794	Valid	0.597	Retain
EDPS_9	1.970	0.493	Moderate	0.332	Valid	0.292	Revise
EDPS_10	1.910	0.478	Moderate	0.812	Valid	0.639	Retain
EDPS_11	1.640	0.410	Moderate	0.729	Valid	0.625	Retain
EDPS_12	1.330	0.332	Moderate	0.683	Valid	0.472	Retain
EDPS_13	1.990	0.496	Moderate	0.141	Invalid	0.097	Reject/Redesign
EDPS_14	1.910	0.478	Moderate	0.831	Valid	0.653	Retain

Note. Item validity was classified using a corrected item-total correlation threshold greater than 0.30. Discrimination decisions followed the criteria established in the Methods section. Decisions are preliminary and should be considered alongside item content and scoring quality.

As presented in Table 3, 13 of the 14 EDPS items met the corrected item-total correlation criterion, whereas one item, EDPS_13, failed to meet the established threshold. The correlations ranged from 0.141 to 0.831, indicating considerable differences in the relationships between individual item scores and the remaining instrument scores. The discrimination indices ranged from 0.097 to 0.694. Based on the combined validity and discrimination criteria, 12 items demonstrated satisfactory characteristics for retention, EDPS_9 required revision because of its relatively low discrimination index, and EDPS_13 required substantial redesign or replacement. These findings indicate that the EDPS instrument contained a majority of items

with acceptable empirical characteristics, although not all items performed equally well.

The difficulty analysis showed that 13 EDPS items were classified as moderately difficult, whereas EDPS_1 was categorized as easy. The difficulty proportions ranged from 0.332 to 0.787. This distribution indicates that most items elicited intermediate levels of performance among the participating students. Such a distribution may be useful in educational assessment because items of moderate difficulty can provide opportunities to observe variation in students' achievement. Nevertheless, moderate difficulty should not automatically be interpreted as evidence of satisfactory measurement quality. An item may exhibit an intermediate mean score while contributing relatively little information for distinguishing students with different levels of the measured competency. Consequently, item difficulty should be interpreted in relation to other empirical characteristics, particularly corrected item-total correlation and discrimination power (Maulina & Novirianthy, 2020; Nasikha & Irma, 2026).

EDPS_1 obtained the highest mean score of 3.150 and a difficulty proportion of 0.787, placing it in the easy category. Despite its relatively high achievement level, this item demonstrated a corrected item-total correlation of 0.824 and a discrimination index of 0.597. These findings suggest that EDPS_1 remained strongly associated with performance on the other EDPS items while differentiating students according to their measured performance. Thus, an item categorized as easy does not necessarily have limited assessment value. Its suitability depends not only on the proportion of attainable scores but also on its relationship with the intended construct and its ability to distinguish differences in student performance.

Among the remaining items, EDPS_14 demonstrated the highest corrected item-total correlation of 0.831, whereas EDPS_4 obtained the highest discrimination index of 0.694. The strong correlation observed for EDPS_14 indicates that performance on this item was closely associated with performance on the remaining EDPS items. Similarly, the relatively high discrimination coefficient for EDPS_4 suggests that the item differentiated students effectively according to the discrimination procedure applied in the study. Other items, including EDPS_2, EDPS_3, EDPS_5, EDPS_6, EDPS_7, EDPS_8, EDPS_10, EDPS_11, and EDPS_12, also demonstrated corrected item-total correlations and discrimination indices above the established criteria. These findings provide preliminary empirical support for retaining these items, subject to continued consideration of their content alignment and scoring characteristics.

EDPS_8 and EDPS_12 are particularly relevant for understanding the relationship between item difficulty and discrimination. EDPS_8 obtained a mean score of 1.400 and a difficulty proportion of 0.351, while EDPS_12 demonstrated the lowest mean score of 1.330 and a difficulty proportion of 0.332. Although these items were relatively challenging compared with the other EDPS items, they remained within the moderate difficulty category. Their corrected item-total correlations were 0.794 and 0.683, respectively, with discrimination indices of 0.597 and 0.472. These characteristics indicate that relatively low mean scores did not prevent the items from demonstrating acceptable empirical functioning. Their retention may therefore be justified by the combined evidence of difficulty, corrected item-total correlation, and discrimination power rather than achievement level alone.

In contrast, EDPS_9 demonstrated a corrected item-total correlation of 0.332 and a discrimination index of 0.292. Although its correlation exceeded the adopted validity threshold of 0.30, its discrimination index fell within the category requiring revision. This distinction is important because meeting the criterion for item-total correlation does not necessarily imply that an item has satisfactory characteristics across all measurement parameters. EDPS_9 was associated with the remaining EDPS scores but demonstrated weaker discriminatory capacity than most other items. Therefore, the item should not be classified as invalid according to the corrected item-total correlation criterion; rather, it should be identified as a valid item requiring revision based on its discrimination performance.

EDPS_13 exhibited the weakest empirical characteristics among the EDPS items.

Although its mean score of 1.990 and difficulty proportion of 0.496 indicated moderate difficulty, the corrected item-total correlation was only 0.141, accompanied by a discrimination index of 0.097. Both coefficients fell below the respective criteria adopted in this study. These results suggest that EDPS_13 contributed relatively little to the consistency of performance represented by the remaining EDPS items and demonstrated limited discriminatory capacity. The item's moderate difficulty level therefore did not compensate for its weak relationships with the other measurement indicators. Its empirical characteristics warrant substantial examination before the item is included in subsequent assessment applications.

The weaknesses observed in EDPS_9 and EDPS_13 should be interpreted in relation to their intended engineering design indicators. Based on the instrument blueprint, EDPS_9 represents designing solutions, whereas EDPS_13 represents making improvements. These processes may require students to integrate biological knowledge, identify relevant design considerations, justify decisions, and evaluate possible modifications. However, the present statistical findings cannot establish whether the weaker performance of these items resulted from unclear instructions, insufficient contextual information, inappropriate cognitive demands, inconsistent scoring, or inadequate alignment with the intended indicators. Accordingly, content review and examination of students' responses are necessary to identify the substantive causes of their weaker empirical characteristics. Combining statistical item analysis with content-oriented evaluation is important for making defensible decisions about item revision and retention (Maulina & Noviriany, 2020; Setiawan et al., 2024).

The findings also have implications for assessing engineering design processes in Biology learning. EDPS involves several related but distinguishable cognitive activities, including problem identification, consideration of constraints, idea generation, solution development, outcome evaluation, and improvement of proposed designs. Previous investigations have highlighted the relevance of engineering design activities for supporting problem-solving and creative thinking in science education (Lin et al., 2021; Ratri et al., 2024; Thammaariyasakun et al., 2025). However, the educational relevance of these processes does not automatically establish the quality of the instruments used to assess them. The present item-level findings emphasize the importance of examining how assessment tasks function empirically before their scores are used to interpret students' engineering design competencies.

Taken together, the EDPS findings indicate that most items demonstrated satisfactory preliminary empirical characteristics, while two items required differentiated follow-up decisions. EDPS_9 should undergo targeted revision because of its relatively weak discrimination, whereas EDPS_13 requires more substantial redesign or replacement because of its low corrected item-total correlation and discrimination index. These decisions should be followed by further administration and analysis to evaluate whether the revised items demonstrate improved measurement characteristics.

Item Analysis of the Creativity Instrument

The creativity instrument was evaluated using the responses of the same 67 students to eight constructed-response items representing fluency, flexibility, originality, and elaboration. The analysis examined item difficulty, corrected item-total correlation, and discrimination power to identify the empirical characteristics of each item. These parameters provided a basis for determining whether individual creativity items demonstrated acceptable functioning within the Human Respiratory System assessment context. The results are presented in Table 4.

Table 4. Results of the Item Analysis of the Creativity Instrument

Item	Mean	Difficulty Proportion	Difficulty Level	Item-Total Correlation	Validity	Discrimination Power	Decision
Creat._1	2.730	0.683	Moderate	0.832	Valid	0.722	Retain
Creat._2	2.460	0.616	Moderate	0.826	Valid	0.708	Retain
Creat._3	2.180	0.545	Moderate	0.828	Valid	0.653	Retain
Creat._4	1.880	0.470	Moderate	0.752	Valid	0.583	Retain

Item	Mean	Difficulty Proportion	Difficulty Level	Item-Total Correlation	Validity	Discrimination Power	Decision
Creat._5	1.900	0.474	Moderate	0.772	Valid	0.472	Retain
Creat._6	1.550	0.388	Moderate	0.703	Valid	0.458	Retain
Creat._7	1.480	0.369	Moderate	0.697	Valid	0.528	Retain
Creat._8	1.970	0.493	Moderate	0.247	Invalid	0.125	Reject/Redesign

Note. The eight creativity items were evaluated using the same validity and discrimination criteria as the EDPS items. Item decisions reflect preliminary empirical performance.

Table 4 shows that seven of the eight creativity items met the adopted corrected item-total correlation criterion, whereas Creativity_8 did not. The item-total correlations ranged from 0.247 to 0.832, while the discrimination indices ranged from 0.125 to 0.722. All eight items were classified as having moderate difficulty, with difficulty proportions ranging from 0.369 to 0.683. The seven items meeting the validity threshold also demonstrated discrimination coefficients within the good category. These findings indicate that most creativity items exhibited satisfactory preliminary empirical characteristics, although one item showed substantially weaker performance and required further evaluation.

The moderate difficulty classification across all creativity items suggests that the assessment tasks elicited intermediate levels of achievement among the participating students. This distribution may be relevant for a constructed-response instrument because creativity assessment requires opportunities for students to demonstrate differences in the quantity, diversity, novelty, and elaboration of their ideas. However, the observed difficulty distribution should not be interpreted as sufficient evidence that all four creativity indicators were adequately represented. Item difficulty primarily describes the relative level of achievement and does not independently establish whether the responses reflect the intended aspects of creativity. Consequently, the appropriateness of the items must also be considered in relation to their empirical associations with the remaining instrument scores and their capacity to discriminate between students.

Creativity_1 demonstrated the highest mean score of 2.730, a difficulty proportion of 0.683, and a corrected item-total correlation of 0.832. It also obtained the highest discrimination index of 0.722. These results indicate that Creativity_1 was strongly associated with performance on the remaining creativity items and exhibited comparatively strong discriminatory characteristics. Creativity_2 demonstrated a similarly favorable pattern, with a mean score of 2.460, a corrected item-total correlation of 0.826, and a discrimination index of 0.708. Both items represented the fluency indicator, suggesting that the corresponding tasks demonstrated satisfactory empirical functioning within the instrument. Nevertheless, these findings should not be interpreted as establishing the independent reliability or validity of the fluency dimension because the present analyses were conducted primarily at the overall instrument and individual-item levels.

Creativity_3 and Creativity_4, representing flexibility, also demonstrated satisfactory characteristics. Their corrected item-total correlations were 0.828 and 0.752, with discrimination indices of 0.653 and 0.583, respectively. Creativity_5 and Creativity_6, representing originality, obtained corrected item-total correlations of 0.772 and 0.703 and discrimination indices of 0.472 and 0.458. These coefficients exceeded the adopted criteria, indicating acceptable empirical performance. Creativity_7, representing elaboration, also met the established thresholds, with a corrected item-total correlation of 0.697 and a discrimination index of 0.528. Thus, seven items distributed across the four creativity indicators demonstrated relationships with the remaining instrument scores and discrimination values consistent with the criteria adopted in this study.

Creativity_8 differed from the other items in several important respects. Although it demonstrated moderate difficulty, with a mean score of 1.970 and a difficulty proportion of 0.493, its corrected item-total correlation was only 0.247 and its discrimination index was 0.125. Both coefficients failed to meet the established criteria. These results indicate that the

item showed a relatively weak relationship with the remaining creativity items and limited ability to differentiate students according to their measured performance. The findings reinforce the importance of examining multiple item characteristics because an item may exhibit a satisfactory difficulty level while failing to demonstrate adequate empirical functioning.

According to the assessment blueprint, Creativity_8 represents elaboration, which involves developing ideas through additional explanation, details, and refinement. Weak empirical functioning on this item may warrant examination of whether the task provides sufficiently clear opportunities for students to elaborate their ideas. Other considerations include the clarity of the question stimulus, alignment between the prompt and the intended indicator, differences in students' interpretation of the task, and the appropriateness of the scoring criteria. However, the available quantitative results do not establish which of these factors explains the observed coefficients. Further qualitative examination of the item and students' written responses would therefore be necessary before determining the most appropriate form of revision.

The interpretation of creativity assessment requires particular attention to the conceptual characteristics of creative thinking. Creativity involves generating ideas that vary in quantity, approach, novelty, and degree of development. Prior research has emphasized that creativity assessment involves methodological challenges related to the definition of creativity, the design of assessment tasks, and the interpretation of resulting scores (Long et al., 2022). Studies of STEAM and engineering design learning have also demonstrated the educational relevance of creative thinking in activities requiring students to explore alternatives and develop solutions (Chang et al., 2022; Erol et al., 2023; Ozkan & Topsakal, 2021). Nevertheless, instruments intended to assess creativity require direct empirical examination rather than relying solely on the theoretical relevance of the learning activities from which their items are derived.

The findings indicate that seven creativity items can be considered candidates for retention, whereas Creativity_8 requires substantial revision or replacement. The latter should undergo content review and subsequent empirical testing to determine whether modifications improve its item-total relationship and discrimination characteristics. Although the creativity instrument demonstrated generally favorable item-level results, additional evidence would be necessary to establish how adequately its four indicators represent the broader construct of creativity in Biology learning.

Reliability Analysis of the Instruments

Instrument reliability was examined to determine the internal consistency of the EDPS and creativity item scores. Cronbach's alpha coefficients were calculated separately for the two instruments using the complete sets of administered items. This analysis complemented the item-level findings by providing an overall estimate of score consistency within each instrument. The reliability results are presented in Table 5.

Table 5. Results of the Reliability Analysis of the Instruments

Instrument	Number of Items	Cronbach's Alpha	Interpretation
EDPS	14	0.933	High
Creativity	8	0.869	High

As shown in Table 5, the EDPS instrument obtained a Cronbach's alpha coefficient of 0.933, while the creativity instrument obtained a coefficient of 0.869. Both coefficients indicated high internal consistency according to the interpretation applied in this study. The EDPS coefficient was higher than that of the creativity instrument, although this difference should not be interpreted as evidence that one construct was measured more accurately than the other. The instruments differed in item number, conceptual content, and assessment characteristics, all of which may influence internal consistency estimates.

The high internal consistency coefficients suggest that the item scores within each

instrument generally covaried in a consistent manner. This finding is relevant because reliability represents an important consideration when assessment scores are used to support interpretations of student performance. Previous methodological research has emphasized that reliability should be evaluated as part of instrument quality assessment and interpreted alongside other sources of measurement evidence (Ekolu & Quainoo, 2019; Kimberlin & Winterstein, 2008). However, Cronbach's alpha does not independently establish whether the items comprehensively represent the intended constructs or whether each individual item performs satisfactorily.

The distinction between instrument-level reliability and item-level functioning is evident in the present findings. Despite the high Cronbach's alpha coefficient for EDPS, EDPS_13 demonstrated a corrected item-total correlation of only 0.141 and a discrimination index of 0.097. Similarly, Creativity_8 showed a corrected item-total correlation of 0.247 and a discrimination index of 0.125, despite the high internal consistency coefficient of the creativity instrument. EDPS_9 also demonstrated comparatively weak discrimination, although its corrected item-total correlation exceeded the adopted threshold. These findings indicate that high overall internal consistency can coexist with individual items requiring revision or further investigation.

The reliability results should also be interpreted in relation to the characteristics of the constructed-response assessment format. Because the instruments required students to produce written responses evaluated by two raters, scoring consistency represents an additional measurement consideration. Cronbach's alpha estimates internal consistency among item scores but does not directly evaluate agreement or consistency between different raters. Accordingly, inter-rater reliability would provide complementary evidence regarding the dependability of the scoring process. The present results should therefore be regarded as evidence of internal consistency rather than a comprehensive evaluation of measurement reliability.

The reliability analysis provides preliminary support for the internal consistency of both instruments, while the item-level findings identify specific weaknesses requiring attention. Further instrument refinement should focus on reviewing the weaker items and evaluating their performance after revision. Internal consistency should also be re-examined in subsequent administrations because changes in item composition, scoring procedures, or participant characteristics may influence the resulting coefficients.

Descriptive Statistics of EDPS and Creativity

Descriptive statistics were examined to summarize the distribution of students' scores on the two assessment instruments. The analysis included mean total scores, standard deviations, observed minimum and maximum scores, and mean percentage scores. These statistics provided contextual information about students' performance and complemented the item-level evaluation by describing the overall distribution of assessment outcomes. The results are presented in Table 6.

Table 6. Descriptive Statistics of EDPS and Creativity Scores

Variable	Mean Total Score	Std. Deviation	Min.	Max.	Mean Percentage Score
EDPS	27.900	12.470	0	51	49.81
Creativity	16.150	6.920	2	28	50.47

Note. The maximum possible total scores were 56 for EDPS and 32 for creativity. Mean percentage scores were calculated relative to the respective maximum possible scores.

As presented in Table 6, the EDPS instrument produced a mean total score of 27.900, with a standard deviation of 12.470. The observed scores ranged from 0 to 51, while the mean percentage score was 49.81%. For the creativity instrument, the mean total score was 16.150, with a standard deviation of 6.920. The observed scores ranged from 2 to 28, corresponding to a mean percentage score of 50.47%. These results indicate that the average observed performance on both instruments was approximately half of the respective maximum possible

scores.

The relatively wide observed score ranges suggest substantial variation in students' performance on the two instruments. Such variation provides a relevant context for item analysis because differences in overall performance can contribute to examining whether individual items differentiate students. However, the score ranges and standard deviations should not independently be interpreted as evidence of satisfactory instrument validity or discrimination. These characteristics describe the distribution of observed performance, whereas item-total correlations and discrimination indices provide more direct evidence of individual item functioning.

The mean scores also require interpretation in relation to the purpose of the study. Because the investigation focused on preliminary empirical item quality, the descriptive statistics were not intended to classify students into predetermined levels of EDPS or creativity. No proficiency thresholds were established for categorizing individual students as demonstrating low, moderate, or high competency. Consequently, the mean percentage scores should be understood as descriptive indicators of observed performance rather than categorical judgments about students' engineering design or creative abilities.

The Human Respiratory System provided a common context in which students responded to assessment tasks intended to represent engineering design processes and creativity. Biological problems associated with respiratory structures, mechanisms, and health conditions can provide opportunities for students to apply scientific understanding, consider alternative solutions, and elaborate their reasoning. Such assessment contexts are conceptually compatible with engineering design and STEAM-oriented learning, in which students are encouraged to connect disciplinary knowledge with problem-solving and design-related activities (Ozkan & Topsakal, 2021; Ratri et al., 2024). However, the descriptive scores alone cannot establish the extent to which students successfully performed each engineering design stage or creativity dimension.

The descriptive findings should therefore be interpreted alongside the item-level and reliability results. Although both instruments captured variation in student scores, the weak empirical characteristics identified for EDPS_13 and Creativity_8, together with the relatively low discrimination of EDPS_9, indicate that further refinement remains necessary. This interpretation emphasizes that the presence of score variation does not eliminate the need to examine individual item quality.

Summary of Instrument Quality and Implications for Instrument Development

An integrated interpretation of the findings was conducted to summarize the empirical quality of the EDPS and creativity instruments. The synthesis considered the number of items meeting the corrected item-total correlation criterion, their discrimination characteristics, and the internal consistency of each instrument. The purpose was to distinguish between item-level validity classification and broader decisions regarding item retention or revision. A summary of the instrument characteristics is presented in Table 7.

Table 7. Summary of Empirical Instrument Quality

Instrument	Total Items	Valid	Invalid	Need Revision	Cronbach's α
EDPS	14	13	1	1	0.933
Creativity	8	7	1	0	0.869
Total	22	20	2	1	—

Note. EDPS_9 met the corrected item-total correlation criterion but required revision because of its discrimination index. EDPS_13 and Creativity_8 did not meet the validity criterion and also demonstrated very low discrimination.

Table 7 indicates that 20 of the 22 assessment items, equivalent to 90.91%, met the adopted corrected item-total correlation criterion. The remaining two items, representing 9.09%, failed to meet this criterion. However, the classification based on corrected item-total correlation should be distinguished from the combined evaluation of validity and

discrimination. Among the 20 items meeting the validity criterion, EDPS_9 exhibited discrimination below the established acceptable range and therefore required revision. Consequently, 19 items, equivalent to 86.36% of the total, met both the adopted item-total correlation and satisfactory discrimination criteria without being identified for revision. These distinctions provide a more accurate representation of instrument quality than treating all statistically valid items as equally suitable for immediate retention.

The findings reveal three categories of item decisions. First, 19 items demonstrated acceptable empirical characteristics under the criteria adopted in the study and can be considered candidates for retention. Second, EDPS_9 met the corrected item-total correlation criterion but required targeted revision because of relatively weak discrimination. Third, EDPS_13 and Creativity_8 failed to meet the item-total correlation criterion and demonstrated very low discrimination, indicating the need for substantial redesign or replacement. These decisions illustrate the importance of evaluating item quality through multiple complementary parameters rather than relying exclusively on a single validity threshold.

The results also indicate that item difficulty was not consistently associated with item quality. Both instruments contained items categorized as moderately difficult, but not all items within this category demonstrated adequate correlations or discrimination indices. EDPS_13 and Creativity_8 provide examples of items with moderate difficulty but weak empirical functioning. Conversely, EDPS_1 was categorized as easy while maintaining relatively strong item-total correlation and discrimination coefficients. These findings support a multidimensional approach to item evaluation in which difficulty, discrimination, item-total correlation, and content alignment are considered simultaneously (Maulina & Novirianthy, 2020; Mustaqim & Sulisti, 2024).

The results have implications for the further development of assessment instruments intended to measure complex competencies. EDPS and creativity involve multiple cognitive processes that may differ in their response requirements and scoring characteristics. Engineering design items should provide opportunities to demonstrate problem identification, consideration of constraints, idea generation, solution development, outcome evaluation, and design improvement. Creativity items should likewise elicit responses reflecting fluency, flexibility, originality, and elaboration. Because these processes are conceptually distinguishable, the empirical functioning of an individual item should be examined in relation to both its statistical characteristics and the indicator it is intended to represent.

The contextual characteristics of the assessment instruments are also important. The Human Respiratory System offers opportunities to connect biological understanding with authentic problems that can require students to propose and evaluate solutions. Previous research has examined the relevance of engineering design and STEAM learning for developing problem-solving and creativity-related competencies (Erol et al., 2023; Lin et al., 2021; Ozkan & Topsakal, 2021). Nevertheless, the potential educational value of such learning contexts does not guarantee that assessment items adequately represent the intended competencies. Instrument development therefore requires a continuing process of empirical examination and conceptual review.

Several limitations should be considered when interpreting the present findings. First, the sample consisted of 67 students from two classes at a single school, limiting the generalizability of the observed item characteristics. Second, the analyses were based primarily on classical test theory indices, which provide useful preliminary information but do not establish the complete dimensional structure of the instruments. Third, although student responses were evaluated by two raters, evidence of inter-rater reliability was not reported, leaving the consistency of scoring across raters insufficiently established. Furthermore, the reported analyses did not include independent evidence of content validity or a factor-analytic examination of the proposed indicators. These limitations suggest that the present findings should be interpreted as preliminary empirical evidence rather than comprehensive psychometric validation.

Future instrument development should therefore prioritize targeted item revision and additional empirical evaluation. EDPS_9 should be reviewed to improve its discriminatory capacity, whereas EDPS_13 and Creativity_8 require more substantial examination of their content, construction, cognitive demands, and scoring criteria. Subsequent administrations should evaluate whether revised items demonstrate improved item-total correlations and discrimination indices. Additional studies involving larger and more diverse student samples, expert review, scoring consistency analysis, and appropriate construct-related validity evidence would further strengthen the interpretation and potential application of these instruments.

The combined findings indicate that the EDPS and creativity instruments demonstrate promising preliminary empirical characteristics, supported by satisfactory performance on most individual items and high internal consistency coefficients. Nevertheless, the identification of three items requiring different levels of revision underscores the importance of interpreting instrument quality at both item and instrument levels. The results provide a basis for targeted refinement and further investigation of constructed-response assessments intended to measure engineering design processes and creativity in secondary Biology learning.

CONCLUSION

This study examined the empirical quality of Engineering Design Process Skills (EDPS) and creativity assessment instruments in Biology learning within the context of the Human Respiratory System. The analysis of 22 constructed-response items involving 67 eleventh-grade students demonstrated that 20 items (90.91%) met the adopted corrected item-total correlation criterion, comprising 13 of the 14 EDPS items and seven of the eight creativity items. The difficulty analysis indicated that 21 items were categorized as moderately difficult, while one EDPS item was classified as easy. These findings provide preliminary empirical evidence that most assessment items demonstrated acceptable characteristics for measuring engineering design processes and creativity within the participating student sample.

The combined evaluation of item-total correlations and discrimination power identified 19 items (86.36%) that satisfied both criteria without requiring revision. However, three items demonstrated characteristics requiring further attention. EDPS_9 met the corrected item-total correlation criterion but showed relatively weak discrimination (0.292), indicating the need for targeted revision. In contrast, EDPS_13 and Creativity_8 failed to meet the adopted item-validity criterion and demonstrated very low discrimination, requiring substantial redesign or replacement. At the instrument level, the EDPS and creativity instruments demonstrated high internal consistency, with Cronbach's alpha coefficients of 0.933 and 0.869, respectively. Nevertheless, these reliability coefficients did not eliminate the weaknesses identified in individual items, emphasizing the importance of combining item-level analysis with instrument-level reliability evaluation.

The findings contribute preliminary, context-specific evidence for evaluating constructed-response instruments intended to assess engineering design processes and creativity in secondary Biology education. The results demonstrate that satisfactory item difficulty and high internal consistency should not be interpreted independently as sufficient evidence of comprehensive instrument quality. Rather, meaningful decisions regarding item retention and refinement require consideration of item-total relationships, discrimination characteristics, and conceptual alignment with the intended assessment indicators.

LIMITATION

This study has several methodological limitations that should be considered when interpreting its findings. The sample comprised 67 eleventh-grade students from two intact classes at a single secondary school, limiting the representativeness and generalizability of the observed item characteristics. The analysis relied on classical test theory indicators, including item difficulty, corrected item-total correlation, discrimination power, and Cronbach's alpha, without examining the dimensional structure or measurement invariance of the instruments.

Although student responses were evaluated by two raters, inter-rater reliability coefficients were not reported, leaving the consistency of scoring between raters insufficiently established. Furthermore, independent evidence of content validity and construct-related validity was not presented, and the assessment tasks were restricted to the Human Respiratory System context. Consequently, the findings should be interpreted as preliminary evidence of empirical item quality rather than comprehensive psychometric validation of the EDPS and creativity instruments.

RECOMMENDATION

Based on the findings, subsequent instrument development should prioritize targeted revision of EDPS_9 and substantial redesign or replacement of EDPS_13 and Creativity_8, with particular attention to construct alignment, question clarity, cognitive demands, contextual relevance, and scoring criteria. The revised items should undergo expert review and further empirical testing to determine whether their corrected item-total correlations and discrimination indices improve. Future research should involve larger and more diverse student samples across multiple schools and biological topics to examine the stability and generalizability of item characteristics. Additional psychometric investigations should incorporate inter-rater reliability analysis, systematic content validation, and appropriate construct-related validity evidence to strengthen the interpretation of assessment scores. For educational practice, the items demonstrating satisfactory preliminary characteristics may serve as candidates for further assessment development, provided that their content alignment and scoring consistency are adequately established before broader implementation.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to SMA Negeri 1 Mijen for providing the opportunity and support to conduct this research. The authors also extend their appreciation to the Biology teachers and the eleventh-grade students who participated in the data collection process. Their cooperation, participation, and valuable contributions were essential to the completion of this study. The authors also acknowledge the support and guidance provided by colleagues and academic supervisors throughout the research process.

FUNDING INFORMATION

This research received no external funding

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mulikah	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Ary Susatyo Nugroho		✓				✓		✓	✓	✓	✓	✓		
Endah Rita Sulistya Dewi	✓		✓	✓			✓			✓	✓		✓	✓

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

The research was conducted in accordance with the ethical standards of the Institutional Review Board and complied with national regulations for research involving human subjects. Permission and ethical approval to conduct the study were granted by the school authority of SMA Negeri 1 Mijen.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

- Adegoke, B. A. (2014). The role of item analysis in detecting and improving faulty physics objective test items. *Journal of Education and Practice*, 5(21), 110.
- Aini, S. N., Munahefi, D. N., Pramasdyahsari, A. S., & Setyowati, R. D. (2024). Engineering design process STEM: Proyek miniatur gazebo Joglo. *PRISMA, Prosiding Seminar*

Nasional Matematika, 37–43.

- Arikunto, S. (2013). *Dasar-dasar evaluasi pendidikan* (2nd ed.). Bumi Aksara.
- Baumert, J., Lüdtke, O., Trautwein, U., & Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: Evidence in support of the distinction between intelligence and student achievement. *Educational Research Review*, 4(3), 165–176. <https://doi.org/10.1016/j.edurev.2009.04.002>
- Bender-Salazar, R. (2023). Design thinking as an effective method for problem-setting and needfinding for entrepreneurial teams addressing wicked problems. *Journal of Innovation and Entrepreneurship*, 12(1), Article 24. <https://doi.org/10.1186/s13731-023-00291-2>
- Biantoro, P. P., & Pertiwi, K. R. (2024). Development of IBSC (investigation based science collaborative) based student worksheet on human respiratory system to improve science literacy and collaboration skills of Grade XI students. *Jurnal Penelitian Pendidikan IPA*, 10(9), 6473–6489. <https://doi.org/10.29303/jppipa.v10i9.8560>
- Caliciuri, R. (2026). *Scientific reasoning: Theoretical foundations, measurement models, and validation studies*. UNITesi.
- Chang, T.-S., Wang, H.-C., Haynes, A. M., Song, M.-M., Lai, S.-Y., & Hsieh, S.-H. (2022). Enhancing student creativity through an interdisciplinary, project-oriented problem-based learning undergraduate curriculum. *Thinking Skills and Creativity*, 46, Article 101173. <https://doi.org/10.1016/j.tsc.2022.101173>
- Daulay, A., Ramadhani, S., Rafli, M., Zulpan, Z., Husein, A., & Wulandari, M. N. (2026). Pemahaman konsep validitas serta pengujiannya dalam penelitian. *Teaching and Learning Research Journal*, 2(1), 173–178. <https://doi.org/10.66770/tlrj.v2i1.97>
- DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., Savoy, S. M., & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship*, 39(2), 155–164. <https://doi.org/10.1111/j.1547-5069.2007.00161.x>
- Ekolu, S. O., & Quainoo, H. (2019). Reliability of assessments in engineering education using Cronbach's alpha, KR and split-half methods. *Global Journal of Engineering Education*, 21(1), 24–29.
- Erol, A., Erol, M., & Başaran, M. (2023). The effect of STEAM education with tales on problem solving and creativity skills. *European Early Childhood Education Research Journal*, 31(2), 243–258. <https://doi.org/10.1080/1350293X.2022.2081347>
- Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, 19(1), 72–91. <https://doi.org/10.1037/a0032138>
- Ghorbiy, B., Parno, P., & Taufiq, A. (2026). Item analysis of a creative thinking skills test on dynamic fluids material. *Momentum: Physics Education Journal*, 10(1), 27–36. <https://doi.org/10.21067/mpej.v10i1.12277>
- Haddad, M., & Sharma, S. (2026). Physiology, lung. In *StatPearls*. StatPearls Publishing. <https://www.ncbi.nlm.nih.gov/books/NBK545177/>
- Handayani, A. W. T., Mardiyana, & Fitriana, L. (2026). Student creativity in STEAM learning through a pop-up book project based on the engineering design process. *SOSIOEDUKASI: Jurnal Ilmiah Ilmu Pendidikan dan Sosial*, 15(2), 3353–3364. <https://doi.org/10.36526/sosioedukasi.v15i2.8505>
- Herrington, J., Reeves, T. C., & Oliver, R. (2014). Authentic learning environments. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 401–412). Springer. https://doi.org/10.1007/978-1-4614-3185-5_32
- Jannah, Z. F., Suciati, S., Fakhrudin, I. A., Silvita, S., Prasetyo, O., & Rachman, H. T. (2024). Identification of scientific knowledge of the process making palm sugar as local wisdom

- of Banyumas and its potential for empowering scientific thinking skills. *Jurnal Pendidikan Sains Indonesia*, 12(4), 935–960. <https://doi.org/10.24815/jpsi.v12i4.40224>
- Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, 65(23), 2276–2284. <https://doi.org/10.2146/ajhp070364>
- Lin, K.-Y., Wu, Y.-T., Hsu, Y.-T., & Williams, P. J. (2021). Effects of infusing the engineering design process into STEM project-based learning to develop preservice technology teachers' engineering design thinking. *International Journal of STEM Education*, 8(1), Article 1. <https://doi.org/10.1186/s40594-020-00258-9>
- Long, H., Kerr, B. A., Emler, T. E., & Birdnow, M. (2022). A critical review of assessments of creativity in education. *Review of Research in Education*, 46(1), 288–323. <https://doi.org/10.3102/0091732X221084326>
- Mardhiya, J., Setiowati, H., & Putri, A. E. (2025). Characterizing 4C skills of pre-service chemistry teachers through inquiry-based learning with multiple representations (IBL-MR): Rasch model analysis. *Journal of the Indonesian Society of Integrated Chemistry*, 17(2), 121–136. <https://doi.org/10.22437/jisic.v17i2.50173>
- Mardhiyah, R. H., Aldriani, S. N. F., Chitta, F., & Zulfikar, M. R. (2021). Pentingnya keterampilan belajar di abad 21 sebagai tuntutan dalam pengembangan sumber daya manusia. *Lectura: Jurnal Pendidikan*, 12(1), 29–40. <https://doi.org/10.31849/lectura.v12i1.5813>
- Maulina, N., & Novirianthy, R. (2020). Item analysis and peer-review evaluation of specific health problems and applied research block examination. *Jurnal Pendidikan Kedokteran Indonesia: The Indonesian Journal of Medical Education*, 9(2). <https://garuda.kemdiktisaintek.go.id/documents/detail/2098590>
- Mohamad, M. M., Sulaiman, N. L., Sern, L. C., & Salleh, K. M. (2015). Measuring the validity and reliability of research instruments. *Procedia - Social and Behavioral Sciences*, 204, 164–171. <https://doi.org/10.1016/j.sbspro.2015.08.129>
- Murray, J. K., Studer, J. A., Daly, S. R., McKilligan, S., & Seifert, C. M. (2019). Design by taking perspectives: How engineers explore problems. *Journal of Engineering Education*, 108(2), 248–275. <https://doi.org/10.1002/jee.20263>
- Mustaqim, M., & Sulisti, H. (2024). Analisis butir soal PAS matematika peminatan: Daya pembeda, tingkat kesukaran, dan kualitas pengecoh. *Al-'Adad: Jurnal Tadris Matematika*, 3(1), 44–56. <https://doi.org/10.24260/add.v3i1.3011>
- Nasikha, L., & Irma, C. N. (2026). Tingkat kesukaran dan daya beda butir soal ulangan harian teks negosiasi pada kelas X di SMK Semesta Bumiayu tahun pelajaran 2024/2025. *Hasta Wiyata*, 9(1), 93–110. <https://doi.org/10.21776/ub.hastawiyata.2025.009.01.07>
- Ozkan, G., & Topsakal, U. U. (2021). Exploring the effectiveness of STEAM design processes on middle school students' creativity. *International Journal of Technology and Design Education*, 31, 95–116. <https://doi.org/10.1007/s10798-019-09547-z>
- Peligro, V. C. (2023). The role of item analysis in detecting and improving faulty multiple-choice test. *Teaching and Teacher Education*, 132.
- Quigley, C. F., Herro, D., & Jamil, F. M. (2017). Developing a conceptual model of STEAM teaching practices. *School Science and Mathematics*, 117(1–2), 1–12. <https://doi.org/10.1111/ssm.12201>
- Ratri, A. Y., Putra, P. D. A., Rusdianto, R., & Nuha, U. (2024). Pengembangan modul berbasis engineering design process (EDP) dalam meningkatkan keterampilan berpikir kreatif siswa SMP pada pembelajaran IPA. *Jurnal Sains dan Edukasi Sains*, 7(2), 105–111. <https://doi.org/10.24246/juses.v7i2p105-111>
- Riduwan, R. (2010). *Metode dan teknik menyusun tesis*. Alfabeta.
- Safitri, E. R., & Prasetyono, H. (2026). Analisis soal HOTS ranah kognitif pada pembelajaran biologi untuk meningkatkan kemampuan berpikir kritis siswa SMA. *Indonesian Journal*

- of *Innovation Multidisipliner Research*, 4(2), 1846–1854.
<https://doi.org/10.69693/ijim.v4i2.674>
- Saufiatun, S. (2024). *Pengembangan modul ajar biologi dengan pendekatan STEAM berbasis engineering design process (EDP) untuk meningkatkan kemampuan berpikir kreatif siswa* [Undergraduate thesis, Universitas Hamzanwadi].
<https://eprints.hamzanwadi.ac.id/6271/>
- Setiawan, R., Fuad, A., & Hilmiyati, F. (2024). Urgensi dan analisis butir instrumen penilaian tes. *Jurnal Paris Langkis*, 5(1), 344–352. <https://doi.org/10.37304/paris.v5i1.17461>
- Sujarweni, W. (2020). *SPSS untuk penelitian*. Pustaka Baru Press.
- Thammaariyasakun, P., Napapongs, W., Tansaku, J., & Inkaew, C. (2025). Development of a virtual learning environment with the engineering design process to enhance students' creative thinking skills. *International Journal of Information and Education Technology*, 15(1), 137–147. <https://doi.org/10.18178/ijiet.2025.15.1.2226>
- Thornhill-Miller, B., Camarda, A., Mercier, M., Burkhardt, J.-M., Morisseau, T., Bourgeois-Bougrine, S., Vinchon, F., El Hayek, S., Augereau-Landais, M., Mourey, F., Feybesse, C., Sundquist, D., & Lubart, T. (2023). Creativity, critical thinking, communication, and collaboration: Assessment, certification, and promotion of 21st century skills for the future of work and education. *Journal of Intelligence*, 11(3), Article 54. <https://doi.org/10.3390/jintelligence11030054>
- Wulandari, I. A., Mu'min, M. B., & Firdaus, M. G. (2021). Peningkatan keterampilan berpikir kritis (KBKr) melalui pembelajaran biologi berbasis keterampilan proses sains. *Bioeduin*, 11(11). <https://doi.org/10.15575/BIOEDUIN.V11I1.12081>