International Journal of Research in Social Science and Humanities (IJRSS)

E-ISSN : 2582-6220

DOI: 10.47505/IJRSS.2026.1.2

Vol. 7 (1) January - 2026

Development of a Micro lesson using an Inquiry-Based LAS in a Flipped Classroom Setting for the Aral Program in Science 10

Janine C. Macaraya[1]& Douglas A. Salazar[2]

[1]Student, College of Education – DSME (CED)

[2] Faculty Member, College of Education – DSME (CED)

MSU-Iligan Institute of Technology

Philippines

ABSTRACT

This study examined the validity, effectiveness, and learner perceptions of a microlesson using inquiry-based Learning Activity Sheets (LAS) in a flipped classroom setting focused on natural selection. Expert validation showed that the lesson plan obtained high mean validity ratings, indicating strengths in structure, clarity, real-life application, and promotion of higher-order thinking. However, content validity analysis revealed insufficient expert consensus, with all Content Validity Index (CVI) values below acceptable thresholds, highlighting the need for substantial revisions in SMART objective formulation, learner engagement, assessment alignment, cultural relevance, and instructional clarity, despite the instrument's high reliability and consistent expert ratings. In contrast, the Learning Activity Sheet demonstrated excellent content validity, achieving unanimous expert agreement, perfect I-CVI and S-CVI/Ave values (1.00), and 100% inter-rater agreement, confirming its suitability for instructional use. The pre-test and post-test instruments were rated as highly valid and reliable in terms of organization, clarity, and relevance; nevertheless, moderate content validity suggested the need for refinement of selected items. Analysis of LAS Task 1 (HOTS–SOLO questions) indicated moderate and consistent student performance, while LAS Task 2 showed low to moderate performance with uniform difficulty across items, supporting the appropriateness of both tasks. Pretest–posttest results revealed statistically significant and educationally meaningful learning gains, with a Wilcoxon Signed-Rank Test result of $Z = 2.37$ ($p = 0.0078$) and a very large effect size ($r = 0.89$), indicating strong intervention impact. Learners reported positive perceptions of the microlesson, reflected in high ratings for concentration, value, interest, and enjoyment, supported by excellent reliability ($\alpha = 0.91$). Thematic analysis of learner reflections showed achievement of core learning objectives, high engagement, and a desire for deeper exploration, with suggestions focused mainly on increased interactivity. Overall, the findings affirm the effectiveness of the microlesson while identifying targeted areas for improvement. These results support evidence-based refinement and classroom implementation.

Key Words: Inquiry-Based Learning, Flipped Classroom Setting, HOTS-SOLO, Micro-lesson.

1. INTRODUCTION

Science education today emphasizes the development of scientific literacy and higher-order thinking skills to prepare learners for informed decision-making and real-world problem solving. As educational goals shift beyond rote memorization, there is a growing move toward student-centered and inquiry-based learning, where learners actively construct knowledge through questioning, investigation, and evidence-based reasoning. However, secondary science instruction faces persistent challenges, including limited instructional time, large class sizes, varying learner readiness, and constraints in resources and teacher training for inquiry-oriented approaches. These challenges highlight the need for innovative instructional models, such as the flipped classroom combined with inquiry-based Learner-Active Strategies (LAS), which can maximize classroom time for meaningful engagement and align with the objectives of the Aral Program in Science 10.

The Aral Program was implemented as a targeted educational intervention to support academically at-risk learners, particularly those who struggle to meet grade-level expectations due to persistent learning difficulties. These challenges call for flexible, learner-centered, and effective instructional strategies that can address varied learning needs while maximizing limited classroom time. Consequently, there is a strong need to design innovative approaches—such as an inquiry-based microlesson delivered through a flipped classroom model—to enhance engagement, support conceptual understanding, and improve learning outcomes for Science 10 learners under the Aral Program.

Systematic and meta-analytic reviews conducted between 2023 and 2025 highlight the effectiveness of microlearning and microlessons in improving learner engagement, retention, and cognitive efficiency, particularly when instructional time is limited and learner readiness varies (Monib et al., 2024; Silva et al., 2025). Microlessons, which focus on a single learning objective delivered in short, structured segments, are especially beneficial for academically at-risk learners as they reduce cognitive overload while allowing flexible access to content.

At the same time, extensive research supports the use of inquiry-based learning as a powerful approach for developing critical thinking, scientific reasoning, and conceptual understanding in science classrooms. Recent systematic reviews indicate that inquiry-based instruction is most effective when learners are provided with clear scaffolds, guiding questions, and structured activities that promote active engagement and evidence-based explanation (Sam, 2024; Urdanivia-Alarcón et al., 2023). These findings align with the principles of Learner-Active Strategies (LAS), which position students as active participants in the learning process and emphasize collaboration, reflection, and meaning-making—key competencies emphasized in secondary science education.

Furthermore, meta-analytic evidence on the flipped classroom model suggests that its effectiveness depends largely on the quality and design of pre-class learning materials. While flipped instruction can significantly increase opportunities for in-class inquiry and collaboration, poorly designed or overly lengthy pre-class content may limit its impact, particularly for learners with academic difficulties (Galindo-Domínguez & Bezanilla, 2025). In this context, integrating inquiry-based LAS microlessons within a flipped classroom setting emerges as an evidence-informed instructional strategy. Such an approach is especially relevant for science 10 learners under the Aral Program, as it addresses learning gaps, maximizes limited instructional time, and provides structured yet flexible support for academically at-risk students.

Recent systematic reviews and meta-analyses (2023–2025) reveal persistent gaps across microlesson/microlearning, inquiry-based learning, and flipped classroom research. Despite evidence supporting microlearning's effectiveness for improving knowledge retention, the literature lacks a clear and unified conceptual framework, with most studies treating microlessons as short content segments rather than as a coherent instructional design approach; moreover, empirical evidence on their impact on higher-order outcomes such as problem solving and critical thinking remains limited, particularly within integrated models like blended or flipped learning (Monib et al., 2024; Silva et al., 2025). Similarly, inquiry-based learning research confirms positive effects on engagement and reasoning but highlights challenges related to student readiness, insufficient instructional scaffolding, inadequate teacher preparation, and a lack of systematic analysis of specific inquiry phases across contexts and grade levels (Sam, 2024; Urdanivia-Alarcón et al., 2023). In the case of the flipped classroom, recent reviews note a strong concentration of studies in higher education, methodological limitations such as short intervention periods and small samples, and mixed findings on learner autonomy and deep learning, often linked to poorly designed pre-class materials and uneven teacher digital readiness (Galindo-Domínguez & Bezanilla, 2025). Collectively, these gaps underscore the need for context-sensitive studies that integrate well-designed microlessons, inquiry-based LAS, and flipped classroom models, particularly in secondary science settings and for academically at-risk learners, such as those in the Science 10 Aral Program.

Therefore, the purpose of this study is to address the identified gaps in recent systematic reviews and meta-analyses by developing an instructional approach that integrates well-designed microlesson, inquiry-based Learner-Active Strategies (LAS), and the flipped classroom model. Specifically, this study responds to the lack of conceptual and practical integration of microlessons within coherent instructional sequences, the need for structured scaffolding

DOI: 10.47505/IJRSS.2026.1.2

in inquiry-based learning, and the limited empirical evidence on the effectiveness of flipped instruction in secondary science classrooms. By focusing on academically at-risk learners in the Science 10 Aral Program, the study aims to generate empirical evidence on how an inquiry-based LAS microlesson implemented in a flipped classroom setting can support deeper conceptual understanding, enhance higher-order thinking skills, and mitigate learning gaps within a real and relevant educational context.

2. METHODS

2.1 Design of the Microlesson Framework

The microlesson framework was systematically designed based on clearly defined learning objectives, prescribed learning content, and the selected instructional approach. The learning objectives were formulated to be achieved within a 25-minute instructional period and were aligned with 21st-century skills. Specifically, the microlesson aimed to develop learners' critical thinking by enabling them to describe natural selection as the primary mechanism driving evolutionary change. Creativity was fostered through the completion of teacher-provided online exercises, while collaboration was encouraged by engaging learners in discussions and inquiry-based activities that promote appreciation of the role of natural selection. Communication skills were emphasized by allowing learners to express their understanding and values regarding the importance of evolutionary principles in addressing real-world issues. The learning content focused on the mechanisms of evolution and was aligned with the content standards stating that learners should understand that multiple theories provide evidence explaining how organisms evolve. The learning competency guiding the microlesson required learners to utilize information from secondary sources to describe natural selection as the primary mechanism of evolutionary change. To effectively achieve these objectives, the microlesson adopted an inquiry-based learning approach implemented through Learning Activity Sheets within a flipped classroom setting, wherein instructional content was accessed prior to class and in-class time was dedicated to guided inquiry and collaborative learning tasks.

2.2 Development of Inquiry-Based Learning Activity Sheets

The Inquiry-Based Learning Activity Sheets were developed following the 7E's STEM-based instructional model to systematically guide learners through the inquiry process in a flipped classroom setting. The Elicit stage was implemented as a pre-class activity, wherein learners watched an instructional video to activate prior knowledge on natural selection. This was followed by the Engage stage, also conducted pre-class, where learners responded to higher-order thinking questions aligned with the SOLO taxonomy to stimulate analytical and abstract reasoning. During in-class implementation, the Explore stage provided learners with interactive online activities from validated educational platforms to deepen conceptual understanding through self-directed exploration. The Explain stage required learners to work in pairs to articulate and justify their understanding of natural selection and evolutionary processes. The Elaborate stage encouraged collaborative brainstorming to strengthen conceptual connections between evolution and natural selection. Learning outcomes were measured during the Evaluate stage through an eight-item teacher-constructed assessment. Finally, the Extend stage introduced supplementary learning materials to reinforce concepts and prepare learners for subsequent lessons.

2.3 Expert Validation and Revision

The selection of experts for the validation of the instructional materials was guided by specific professional and academic criteria to ensure credible and informed evaluation. Experts were required to have a major in Science and hold a master's degree in Science Education, indicating advanced academic preparation and content expertise relevant to the study. In addition, a minimum of three years of teaching experience was required to ensure that the experts possessed sufficient classroom exposure and practical understanding of instructional design and learner needs. Lastly, the experts were currently employed in the Department of Education at the Junior High School (JHS) level, ensuring familiarity with the curriculum standards, learning competencies, and classroom realities relevant to the instructional context of the developed materials.

DOI: 10.47505/IJRSS.2026.1.2

Table 1. Criteria for Expert Selection

Major	Degree	Years of Teaching	Current Workplace
Science	Masters in Science Education	3 and/or more years	Department of Education (JHS)

The study employed multiple validation instruments to ensure the quality, validity, and effectiveness of the developed instructional materials. The pre-test and post-test validation instrument, adopted from Renz Ferrer, was used to establish content and construct validity through a 5-point rating scale assessing organization, clarity, and relevance of test items, with provisions for qualitative comments and suggestions. The Learning Activity Sheets were validated using an instrument adapted from DepEd Region I, which evaluated content accuracy, alignment with learning objectives and MELCs, organization, language, illustrations, and layout, ensuring that the materials were free from errors, age-appropriate, culturally relevant, and technically sound. The lesson plan was evaluated using a rating scale adapted from Go (2019) to assess instructional design, alignment, and pedagogical appropriateness. In addition, learners' experiences were measured using a Perception Questionnaire adapted from Deci and Ryan (1994) to determine levels of motivation, enjoyment, perceived choice, and value of the developed lesson through a 6-point rating scale. To complement the quantitative data, a Self-Reflection Questionnaire adapted from Capampangan (2013) was used to gather qualitative insights regarding learners' learning experiences, engagement, and suggestions for improvement. The combined use of these instruments provided a comprehensive evaluation of the instructional materials from both expert and learner perspectives.

2.3.1 Evaluation of the Lesson Plan

The mean scores for the evaluation of the lesson plan indicate the overall level of validity as perceived by the experts, while the standard deviation reflects the degree of agreement or consistency among their ratings. The indicator "Are the lesson objectives SMART?" obtained a mean score of 3.00 with a standard deviation of 0.71, which was interpreted as Moderately Valid. All remaining indicators received mean scores ranging from 3.20 to 3.60, which fall under the interpretation of Highly Valid. Indicators related to lesson structure and delivery, such as clarity of lesson presentation (M = 3.60, SD = 0.55) and provision of sufficient practice (M = 3.60, SD = 0.55), suggest that experts strongly agreed that the lesson is well-organized and provides adequate opportunities for learners to apply concepts. The moderate SD values indicate a high level of consistency in expert judgments. Indicators focusing on learner engagement, contextualization, and appropriateness of materials—including promotion of learner engagement, use of localized contexts, cultural relevance, and suitability of language—also yielded high validity ratings. These results imply that the lesson effectively considers learner diversity and real-world relevance, which are essential in enhancing motivation and comprehension. Similarly, indicators related to problem design and assessment strategies, such as the inclusion of real-life applications, allowance for multiple solutions, requirement of reasoning, clarity of problem statements, and use of scoring rubrics, were all rated as Highly Valid. This indicates that the lesson encourages higher-order thinking skills and aligns assessment practices with instructional objectives. Finally, the indicators addressing usefulness and acceptability of the lesson and strategies—specifically whether teachers would find the strategy useful and learners would find it interesting—also obtained high validity ratings. This suggests strong expert confidence in the practical applicability and instructional value of the developed lesson plan.

DOI: 10.47505/IJRSS.2026.1.2

Table 1.1. Descriptive Statistics for the Validity of the Lesson Plan

Indicator	M	SD	Interpretation
Are the lesson objectives SMART?	3	0.71	Moderately Valid
Are the steps in the lesson presentation clear?	3.6	0.55	Highly Valid
Do the activities promote learner engagement?	3.2	0.45	Highly Valid
Does the lesson provide sufficient practice?	3.6	0.55	Highly Valid
Are the assessment activities appropriate?	3.2	0.45	Highly Valid
Is the use of localized contexts appropriate?	3.6	0.55	Highly Valid
Are the materials culturally relevant?	3.2	0.45	Highly Valid
Are the words suitable for learners?	3.6	0.55	Highly Valid
Are technical terms appropriate?	3.2	0.45	Highly Valid
Do problems involve real-life application?	3.6	0.55	Highly Valid
Do problems allow multiple solutions?	3.2	0.45	Highly Valid
Do problems require reasoning?	3.6	0.55	Highly Valid
Are problems stated clearly?	3.2	0.45	Highly Valid
Do the problems use a scoring rubric?	3.6	0.55	Highly Valid
Will teachers find the strategy useful?	3.2	0.45	Highly Valid
Will learners find the strategy interesting?	3.6	0.55	Highly Valid
Are the lessons useful to teachers?	3.2	0.45	Highly Valid

The I-CVI reflects the proportion of experts who judged each indicator as valid, while the S-CVI/Ave represents the overall average content validity of the lesson plan across all indicators. The results show that the I-CVI values vary considerably across indicators, ranging from 0.20 to 0.60. Indicators such as Steps Clear, Practice, Localized Context, Language Suitability, Real-life Application, Reasoning, Scoring Rubric, and Interesting to Learners obtained an I-CVI of 0.60, indicating that a majority of the expert validators considered these aspects of the lesson plan acceptable. Although these values suggest partial agreement among experts, they fall below the commonly accepted threshold of 0.78 for five validators, implying that these components require further refinement to achieve strong content validity. On the other hand, several indicators—including Objectives SMART, Learner Engagement, Assessment, Cultural Relevance, Technical Terms, Multiple Solutions, Problem Clarity, Usefulness to Teachers, and Lesson Usefulness—received lower I-CVI values of 0.20 or 0.40. These low values indicate that only one or two experts rated these indicators as valid, suggesting notable concerns regarding the clarity, relevance, or alignment of these components with the intended instructional goals. In particular, the low I-CVI for SMART objectives and learner engagement highlights areas that need substantial revision to better meet pedagogical standards and learner needs. At the scale level, the S-CVI/Ave of 0.39 indicates that, overall, the lesson plan exhibits low content validity. This value is well below the acceptable benchmark of 0.80, signifying insufficient consensus among experts regarding

DOI: 10.47505/IJRSS.2026.1.2

the validity of the lesson plan as a whole. The findings suggest that while some elements of the lesson plan show potential, significant improvements are necessary across multiple components. Overall, the CVI results imply that the lesson plan requires systematic revision, particularly in strengthening the formulation of learning objectives, enhancing learner engagement strategies, improving assessment alignment, and ensuring cultural relevance and instructional clarity. Addressing these concerns is essential to improve expert agreement and to elevate the lesson plan's overall content validity before classroom implementation.

Table 1.2. Content Validity Index of the of the Lesson Plan

Item	I-CVI
Objectives SMART	0.4
Steps Clear	0.6
Learner Engagement	0.2
Practice	0.6
Assessment	0.2
Localized Context	0.6
Cultural Relevance	0.2
Language Suitability	0.6
Technical Terms	0.2
Real-life Application	0.6
Multiple Solutions	0.2
Reasoning	0.6
Problem Clarity	0.2
Scoring Rubric	0.6
Usefulness to Teachers	0.2
Interesting to Learners	0.6
Lesson Usefulness	0.2
Scale-Level CVI (S-CVI/Ave)	0.39

Reliability analysis was conducted to determine the internal consistency of the instrument, or the extent to which the set of items consistently measures the same construct—in this case, the validity of the lesson plan. The results show that the instrument obtained a Cronbach's alpha value of 0.97, which indicates excellent internal consistency. According to commonly accepted standards, alpha values of 0.70 and above are considered acceptable, while values 0.90 and higher reflect very high reliability. Thus, the obtained alpha value suggests that the 17 indicators in the validation instrument are highly consistent and reliably assess various aspects of the lesson plan. The reliability analysis was based on 17 items evaluated by five (5) expert validators. Despite the relatively small number of experts, the very high alpha value demonstrates that the experts responded consistently across items. This finding confirms that the validation instrument is stable, dependable, and suitable for expert evaluation, and that the results derived from it can be considered reliable for decision-making regarding lesson plan revision and improvement.

Table 1.3. Reliability of the Lesson Plan Validation Instrument

Measure	Value
Cronbach's Alpha (α)	0.97
Number of Items	17
Number of Experts	5

Friedman test and Kendall's coefficient of concordance (W), which were used to examine differences and levels of agreement among the lesson plan validity dimensions as rated by the expert validators. The results show a

DOI: 10.47505/IJRSS.2026.1.2

Friedman chi-square (χ^2) value of 2.00 with 2 degrees of freedom and a p-value of 0.368. Since the p-value is greater than the 0.05 level of significance, the result indicates that there is no statistically significant difference in the experts' ratings across the evaluated dimensions. In addition, Kendall's coefficient of concordance (W) was computed to determine the degree of agreement among expert validators. The obtained Kendall's W value of 0.20 indicates a fair level of agreement among the experts. This implies that while there is some consistency in expert judgments, there is also noticeable variation in how the validators rated the different dimensions of the lesson plan.

Table 1.4. Differences and Agreement Among Lesson Plan Validity Dimensions

Statistic	Value	df	p	Interpretation
Friedman χ^2	2	2	0.368	No significant difference
Kendall's W	0.2	—	—	Fair agreement

2.3.2 Evaluation of the Learning Activity Sheets

The frequency and percentage analysis was conducted to describe how the expert validators responded to each indicator used in evaluating the validity of the Learning Activity Sheet. A total of five (5) expert validators participated in the evaluation using an instrument composed of twenty-three (23) indicators. These indicators were grouped into three domains: Content (8 indicators), Organization, Presentation, and Language (9 indicators), and Design and Layout (6 indicators). Since each expert rated all 23 indicators, the total number of responses was computed by multiplying the number of indicators by the number of experts. Thus, 23 indicators × 5 experts resulted in 115 total responses. These responses were then tallied to determine the frequency and corresponding percentage of "Yes" and "No" ratings. This approach provides a clear descriptive summary of the overall level of agreement among experts regarding the validity of the Learning Activity Sheet.

Table 1.5. Frequency and Percentage of the Learning Activity Sheets

Response	Frequency	Percentage
Yes	115	100.00%
No	0	0.00%
Total	115	100.00%

The Content Validity Index (CVI) was used to determine the degree to which the items in the Learning Activity Sheet adequately represent the intended content domain based on expert judgment. CVI is commonly applied in instructional material validation to quantify the level of expert agreement regarding item relevance, clarity, and appropriateness. At the item level, the Item-Level Content Validity Index (I-CVI) was computed by dividing the number of experts who rated an item as valid ("Yes") by the total number of expert validators. In this study, all five (5) experts rated every one of the 23 indicators as valid. As a result, each item obtained an I-CVI value of 1.00, indicating unanimous agreement among experts for all indicators. At the scale level, the Scale-Level Content Validity Index using the average method (S-CVI/Ave) was computed by averaging the I-CVI values of all items. Since all 23 items had an I-CVI of 1.00, the S-CVI/Ave was calculated as 23 divided by 23, resulting in a value of 1.00. This finding indicates excellent content validity, exceeding the commonly accepted threshold of 0.80, and confirms that the Learning Activity Sheet is highly appropriate and suitable for its intended instructional purpose.

Table 1.6. Content Validity Index (CVI) of the Learning Activity Sheets

CVI Measure	Value
Lowest I-CVI	1
Highest I-CVI	1
S-CVI/Ave	1

DOI: 10.47505/IJRSS.2026.1.2

The inter-rater agreement analysis was conducted to determine the level of consistency among the expert validators in evaluating the Learning Activity Sheet. This measure reflects how similarly the experts judged each indicator, providing evidence of the reliability of their evaluations. Inter-rater agreement was computed using percent agreement, which is calculated by dividing the number of agreements by the total number of ratings and multiplying the result by 100. In this study, all five (5) expert validators provided the same rating ("Yes") for each of the 23 indicators, resulting in 115 agreements out of 115 total ratings. Using the formula, the percent agreement was calculated as115/115 x 100, yielding a value of 100%. This result indicates perfect agreement among the expert validators. Such a high level of agreement demonstrates very strong consistency in expert judgments and supports the reliability of the validation process, suggesting that the Learning Activity Sheet was evaluated in a clear and unambiguous manner across all criteria.

Table 1.7. Inter-Rater Agreement of the Learning Activity Sheets

Measure	Value
Number of Validators	5
Total Ratings	115
Percent Agreement	100%
Measure	Value

Across all 23 indicators and 115 total expert ratings, the Learning Activity Sheet attained a mean I-CVI of 1.00 and 100% inter-rater agreement, leading to an interpretation of excellent validity. These results strongly indicate that the developed Learning Activity Sheet meets high standards in terms of content accuracy, instructional organization, language clarity, and design quality. Consequently, the material is deemed highly suitable for classroom implementation and can effectively support teaching and learning processes in science. For the Content domain, which consisted of eight (8) indicators and generated forty (40) total ratings, the results yielded a mean Item-Level Content Validity Index (I-CVI) of 1.00 and a percent agreement of 100%. The Organization, Presentation, and Language domain, composed of nine (9) indicators and forty-five (45) total ratings, also achieved a mean I-CVI of 1.00 with 100% agreement among experts. The Design and Layout domain, which included six (6) indicators and thirty (30) total ratings, likewise obtained a mean I-CVI of 1.00 and 100% percent agreement. This reflects expert consensus that the layout of the activity sheets is visually clear and well-structured, with readable fonts, properly labeled questions, adequate spaces for learner information, appropriate use of icons, and formats that are easy to edit, store, and print.

Table 1.8. Descriptive Summary by Domain of the Learning Activity Sheets

Domain	No. of Items	Total Ratings	Mean I-CVI	Percent Agreement	Interpretation
Content	8	40	1	100%	Excellent Validity
Organization, Presentation, and Language	9	45	1	100%	Excellent Validity
Design and Layout	6	30	1	100%	Excellent Validity
Overall	23	115	1	100%	Excellent Validity

2.3.2 Evaluation of the Pre-test and Post-test

Under the Organization domain, the three indicators obtained mean scores ranging from 3.6 to 3.8, all of which were interpreted as Highly Valid. Specifically, Organization indicators 1 and 2 each recorded a mean of 3.8 with a standard deviation of 0.84, while Organization indicator 3 had a mean of 3.6 and an SD of 0.89. The moderate

DOI: 10.47505/IJRSS.2026.1.2

SD values indicate reasonable consistency among expert ratings. For the Clarity domain, indicators 4 and 5 both received high validity ratings. Clarity indicator 4 achieved the highest mean score of 4.0 with an SD of 0.71, indicating that experts found the test directions to be very clear and accurate. Clarity indicator 5 obtained a mean of 3.6 with an SD of 0.89, reflecting that the language used in the test items is generally clear and understandable. Similarly, all indicators under the Relevance domain were rated as Highly Valid, with mean scores ranging from 3.6 to 4.0. Relevance indicators 6 and 8 both obtained a mean of 4.0 with an SD of 0.71, indicating that the test items were viewed as highly appropriate for measuring the intended variables and aligned with the objectives of the study. Relevance indicator 7 had a slightly lower mean of 3.6 and an SD of 0.89, but still fell within the highly valid range.

Table 1.9. Descriptive Statistics for the Validity of the Pre-test and Post-test

Indicator	M	SD	Verbal Interpretation
Organization 1	3.8	0.84	Highly Valid
Organization 2	3.8	0.84	Highly Valid
Organization 3	3.6	0.89	Highly Valid
Clarity 4	4	0.71	Highly Valid
Clarity 5	3.6	0.89	Highly Valid
Relevance 6	4	0.71	Highly Valid
Relevance 7	3.6	0.89	Highly Valid
Relevance 8	4	0.71	Highly Valid

The CVI was used to quantify the degree of agreement among expert validators regarding the relevance and appropriateness of each test item. At the item level, the I-CVI values range from 0.40 to 0.80, indicating varying levels of expert agreement across the indicators. The indicators Clarity 4, Relevance 6, and Relevance 8 obtained the highest I-CVI value of 0.80, suggesting that four out of the five expert validators agreed that these items were valid. These results indicate relatively strong content validity for the clarity of directions and the relevance of several test items to the study objectives. In contrast, indicators Organization 3, Clarity 5, and Relevance 7 recorded lower I-CVI values of 0.40, meaning that only two of the five experts rated these items as valid. Similarly, Organization 1 and Organization 2 obtained moderate I-CVI values of 0.60, reflecting agreement from three experts. These findings suggest that while these items are acceptable to some extent, they may require revision or refinement to improve clarity, alignment with the table of specifications, or relevance to the intended learning outcomes. At the scale level, the Scale-Level CVI (S-CVI/Ave) was computed by averaging the I-CVI values across all eight indicators. The resulting S-CVI/Ave of 0.60 indicates moderate overall content validity of the pre-test and post-test instruments. However, this value falls below the commonly recommended threshold of 0.80, suggesting that the instruments would benefit from further revision to increase expert agreement and strengthen their overall validity.

Table 1.10. Content Validity Index (CVI) of the Pre-test and Post-test

Item	I-CVI
Organization 1	0.6
Organization 2	0.6
Organization 3	0.4
Clarity 4	0.8
Clarity 5	0.4
Relevance 6	0.8
Relevance 7	0.4
Relevance 8	0.8
Scale-Level CVI (S-CVI/Ave)	0.6

DOI: 10.47505/IJRSS.2026.1.2

The reliability analysis of the validation instrument was conducted to determine its internal consistency, or the extent to which the items consistently measure the same construct. As shown in the table, the validation instrument obtained a Cronbach's alpha (α) value of 0.98, which indicates excellent internal consistency. In general, Cronbach's alpha values of 0.70 or higher are considered acceptable, while values above 0.90 reflect very high reliability. Therefore, an alpha of 0.98 suggests that the items in the instrument are highly correlated and consistently interpreted by the expert validators. The reliability analysis was based on eight (8) items evaluated by five (5) expert validators. This result confirms that the validation instrument is highly reliable and suitable for evaluating the validity of the pre-test and post-test, and that the findings derived from its use can be considered trustworthy for research and instructional decision-making.

Table 1.11. Reliability Analysis of the Validation Instrument

Measure	Value
Cronbach's Alpha (α)	0.98
Number of Items	8
Number of Experts	5

The Friedman test was applied as a nonparametric statistical procedure to determine whether there were statistically significant differences in expert ratings across the evaluated validity criteria. The results yielded a Friedman chi-square (χ^2) value of 4.00 with 2 degrees of freedom and a p-value of 0.135. Since the p-value is greater than the 0.05 level of significance, the result indicates that there is no statistically significant difference among the validity criteria. In addition, Kendall's coefficient of concordance (W) was computed to assess the level of agreement among expert validators. The obtained Kendall's W value of 0.40 indicates a moderate level of agreement among the experts.

Table 1.12. Friedman Test and Kendall's Coefficient of Concordance for Expert Ratings Across Validity Criteria

Statistic	Value	df	p	Interpretation
Friedman χ^2	4	2	0.135	No significant difference among criteria
Kendall's W	0.4	—	—	Moderate agreement among experts

2.4 Implementation in a Flipped Classroom Setting

Based on the results of the study "Science 10 Academic Performance Over Three School Years: A Statistical Study at Pantao Ragat Agro-Industrial High School" by Macaraya (2025), findings revealed a small upward trend in the academic performance of Grade 10 students in Science across three consecutive school years. While this trend indicates gradual improvement, the increase was minimal, suggesting that existing instructional strategies may not be sufficient to produce substantial learning gains. This condition highlights the need to upgrade teaching approaches through learning innovations, particularly those that actively engage learners, promote higher-order thinking, and address persistent learning gaps in Science. Furthermore, the relevance of focusing on Grade 10 and the topic of evolution is supported by the findings of the study "Awareness, Knowledge, and Attitudes Toward Human Evolution Among Junior and Senior High School Students in Mindanao, Philippines" by Perez, and Macaraya et al. (2025). The study revealed that Grade 10 students demonstrated the least level of knowledge about evolution compared to other grade levels. Given that evolution is a foundational yet conceptually challenging topic in Biology, these findings justify the conduct of an intervention study targeting Grade 10 learners. Hence, implementing an innovative instructional approach on evolution at the Grade 10 level is both timely and necessary to address identified learning deficiencies and improve students' conceptual understanding in Science. Hence, implementing an innovative instructional approach on evolution at the Grade 10 level is both timely and necessary to address identified learning deficiencies and improve students' conceptual understanding in Science. In line with this objective, the present study

DOI: 10.47505/IJRSS.2026.1.2

was conducted among seven Grade 10 learners enrolled in the ARAL Program of Pantar National High School, who were purposively selected as a pilot class. This group was chosen to initially implement and evaluate the effectiveness of the developed instructional innovation, allowing for focused observation, formative assessment, and refinement of the learning materials prior to broader application.

3. RESULTS

3.1 Evaluation of Effectiveness

This part of the study presents the evaluation of the two tasks included in the Learning Activity Sheets (LAS) and the pre-test and post-test results. To provide a comprehensive assessment, descriptive statistics were computed to summarize students' scores, while the item difficulty index was used to determine the level of difficulty of each task. In addition, the Friedman test was employed to examine possible differences in student performance across the tasks, and Kendall's coefficient of concordance (W) was calculated to determine the degree of agreement or consistency in student rankings. These statistical tools collectively provide a deeper understanding of the effectiveness and appropriateness of the LAS tasks for instructional use.

3.1.1. Learning Activity Sheets Task 1: HOTS-SOLO QUESTIONS

Descriptive statistics summarize students' performance, while the item difficulty index identifies how easy or difficult each task was. The Friedman test examines differences in performance across tasks, and Kendall's W determines the level of agreement or consistency in the results. The analysis using these tools is presented in the following section.

Table 2.1. Descriptive Statistics per Item

Item	Mean	Median	SD	Min	Max
Q1	2.43	2	1.13	1	4
Q2	2.14	2	1.07	1	4
Q3	2.86	3	0.9	1	4
Q4	2.14	2	1.07	1	4

Table 2.1 presents the descriptive statistics for each item answered by the seven students. Overall, the mean scores range from 2.14 to 2.86 on a 4-point scale, indicating moderate performance across all items. Item Q3 obtained the highest mean score (M = 2.86) and median of 3, suggesting that most students performed relatively better on this question compared with the others. Its lower standard deviation (SD = 0.90) also indicates more consistent responses among students. In contrast, Items Q2 and Q4 recorded the lowest mean scores (M = 2.14), with medians of 2, implying that these items were more challenging for the students. Both items show higher variability (SD = 1.07), reflecting differences in individual performance. Item Q1 shows a mean score of 2.43 and the highest standard deviation (SD = 1.13), suggesting a wider spread of scores and varying levels of understanding among students. The minimum and maximum scores for all items ranged from 1 to 4, indicating that while some students struggled, others were able to achieve high scores.

Table 2.2. Item Difficulty Index (p)

Item	Mean	Difficulty Index (p)	Interpretation
Q1	2.43	0.61	Moderately Easy
Q2	2.14	0.54	Moderate
Q3	2.86	0.71	Easy
Q4	2.14	0.54	Moderate

Table 2.2 presents the item difficulty index (p) for each question, which indicates how easy or difficult eachitem was for the students. The difficulty index is computed by dividing the mean score of each item by the maximum possible score, with higher values indicating easier items. Item Q3 has the highest difficulty index (p =

DOI: 10.47505/IJRSS.2026.1.2

0.71), which is interpreted as Easy. This suggests that most students were able to answer this question correctly or score high on it, indicating better understanding of the concept measured by Q3. Item Q1 obtained a difficulty index of 0.61, classified as Moderately Easy, implying that while the item posed some challenge, a majority of students were able to respond satisfactorily. Items Q2 and Q4 both recorded a difficulty index of 0.54, interpreted as Moderate. These items were relatively more difficult compared with Q1 and Q3, as students showed lower average performance. This indicates that the concepts assessed in Q2 and Q4 may require further clarification or reinforcement during instruction.

Table 2.3. Friedman Test

Statistic	Value	
Friedman χ^2	6.29	
df	3	
p-value	0.099	
Kendall's W	0.3	Fair agreement

Table 2.3 presents the results of the Friedman test and Kendall's coefficient of concordance (W), which were used to examine whether there were significant differences in student performance across the four items and to assess the level of consistency among the results. The Friedman test yielded a chi-square value of 6.29 with 3 degrees of freedom and a p-value of 0.099. Since the p-value is greater than the conventional significance level of 0.05, the result indicates that there is no statistically significant difference in student performance across the four items. This suggests that the items were generally comparable in terms of difficulty and that students performed at relatively similar levels across the tasks. Kendall's W was computed at 0.30, which indicates a fair level of agreement among the item rankings. This implies moderate consistency in how students performed across the different items, although some variability in responses was still observed. Overall, these results suggest that the set of items functioned consistently, with no item standing out as significantly easier or more difficult than the others.

3.1.2. Learning Activity Sheet Task 2: The Follow Up Questions

Descriptive statistics were used to summarize students' performance, while the item difficulty index determined how easy or difficult each task was. The Friedman test assessed differences in performance across tasks, and Kendall's W measured the level of agreement or consistency among the results. The findings from these analyses are presented in the succeeding section.

2.4. Descriptive Statistics per Item

Item	Mean (M)	Median	SD	Min	Max
Q1	2.29	2	0.76	1	3
Q2	1.71	2	0.49	1	2
Q3	2.29	3	1.25	0	3
Q4	1.86	2	0.9	1	3

Table 2.4 presents the descriptive statistics for each item, including the mean, median, standard deviation, and range of scores. These values provide an overview of students' performance and the variability of responses for each item. For Q1, the mean score of 2.29 and a median of 2 indicate that most students obtained moderate scores, suggesting an average level of understanding of the concept measured by this item. The relatively low standard deviation (SD = 0.76) shows that student responses were fairly consistent, with scores ranging from 1 to 3. Q2 obtained the lowest mean score (M = 1.71) among the four items, with a median of 2. This indicates that students generally found this item more challenging compared to the others. The small standard deviation (SD = 0.49) suggests limited variation in responses, meaning most students performed similarly on this question, with scores clustered between 1 and 2. For Q3, the mean score is 2.29 and the median is 3, indicating that several students were able to achieve higher scores on this item. However, the standard deviation is relatively high (SD = 1.25), reflecting greater variability in student performance. The wide range of scores (0 to 3) suggests that while some students demonstrated strong understanding, others struggled considerably with this item. Lastly, Q4 has a mean score of 1.86 and a median

DOI: 10.47505/IJRSS.2026.1.2

of 2, indicating generally low to moderate performance. The standard deviation of 0.90 shows moderate variability in responses, and scores ranged from 1 to 3, suggesting differences in students' levels of understanding.

2.5. Item Difficulty Index (p)

Item	Mean (M)	Difficulty Index (p)	Interpretation
Q1	2.29	0.57	Moderate difficulty
Q2	1.71	0.43	Moderately difficult
Q3	2.29	0.57	Moderate difficulty
Q4	1.86	0.46	Moderately difficult

Table 2.5 presents the item difficulty index (p) for each item, which indicates the proportion of the maximum possible score achieved by students. Higher p-values reflect easier items, while lower values indicate more difficult items. Q1 obtained a difficulty index of 0.57, which falls under moderate difficulty. This suggests that students were able to answer this item reasonably well, but it still required an adequate level of understanding and was not overly easy. Q2 recorded a difficulty index of 0.43, indicating that it was moderately difficult. This implies that many students experienced challenges with this item, consistent with its lower mean score, and suggests a need for further instructional support or item refinement. Similarly, Q3 also yielded a difficulty index of 0.57, classifying it as having moderate difficulty. This indicates a balanced item that can effectively discriminate between students with higher and lower levels of understanding. Q4 showed a difficulty index of 0.46, placing it in the moderately difficult range. This suggests that the item posed some level of challenge to students, but was still appropriate for assessing the intended learning outcome.

2.6. Friedman Test (Differences Across Items)

Statistic	Value	
Friedman χ^2	5.15	
df	3	
p-value	0.161	
Kendall's W	0.25	Fair agreement

Table 2.6 presents the results of the Friedman test, which was used to determine whether there were statistically significant differences in students' performance across the four items. The Friedman chi-square value (χ^2 = 5.15) with 3 degrees of freedom yielded a p-value of 0.161. Since the p-value is greater than the 0.05 level of significance, the result indicates that there is no statistically significant difference in students' scores across the four items. This suggests that the items were comparable in terms of difficulty and that students performed relatively consistently across them. In addition, Kendall's coefficient of concordance (W = 0.25) indicates a fair level of agreement among the item rankings. This means that while there was some consistency in how students performed across the items, the agreement was not strong, which is expected given the small sample size and varying individual responses.

3.1.3. Pre-test and Post-test

2.7. Frequency and Percentage of Correct Responses (Pretest vs Posttest)

Item	Pretest (f)	Pretest (%)	Posttest (f)	Posttest (%)
Q1	1	14.29%	6	85.71%
Q2	2	28.57%	3	42.86%
Q3	1	14.29%	4	57.14%
Q4	1	14.29%	6	85.71%
Q5	2	28.57%	5	71.43%
Q6	0	0.00%	4	57.14%

DOI: 10.47505/IJRSS.2026.1.2

Q7	0	0.00%	3	42.86%
Q8	1	14.29%	5	71.43%

The results show a consistent increase in correct responses across all items after the intervention, indicating an overall improvement in students' understanding of the concepts assessed. In the pretest, correct response rates were generally low, ranging from 0.00% to 28.57%, suggesting that most students had limited prior knowledge of the topics. Items Q6 and Q7 recorded no correct responses (0.00%), highlighting these as the most challenging concepts before instruction. In contrast, the posttest results demonstrate a substantial increase in performance, with correct response rates ranging from 42.86% to 85.71%. Notably, items Q1 and Q4 showed the greatest improvement, increasing from 14.29% to 85.71%, indicating strong mastery of these concepts following the intervention. Similarly, items Q5 and Q8 improved to 71.43%, reflecting a marked enhancement in conceptual understanding. Although improvement was evident in all items, Q2 and Q7 exhibited relatively lower posttest percentages (42.86%), suggesting that these concepts may require additional instructional reinforcement or alternative teaching strategies.

2.8. Item Difficulty Index (p)

Item	Pretest p	Interpretation	Posttest p	Interpretation
Q1	0.14	Very Difficult	0.86	Easy
Q2	0.29	Difficult	0.43	Moderate
Q3	0.14	Very Difficult	0.57	Moderate
Q4	0.14	Very Difficult	0.86	Easy
Q5	0.29	Difficult	0.71	Easy
Q6	0	Extremely Difficult	0.57	Moderate
Q7	0	Extremely Difficult	0.43	Moderate
Q8	0.14	Very Difficult	0.71	Easy

In the pretest, all items were classified as difficult to extremely difficult, with p values ranging from 0.00 to 0.29. Items Q6 and Q7 were identified as extremely difficult (p = 0.00), as no students answered these items correctly. Items Q1, Q3, Q4, and Q8 were categorized as very difficult (p = 0.14), while Q2 and Q5 were considered difficult (p = 0.29). These findings suggest that students initially had limited mastery of the concepts assessed by the test items prior to the instructional intervention. In contrast, the posttest results show a substantial shift toward more favorable difficulty levels, with p values increasing to a range of 0.43 to 0.86. Items Q1 and Q4 became easy (p = 0.86), indicating strong post-instruction mastery of these concepts. Items Q5 and Q8 also reached the easy level (p = 0.71). Meanwhile, items Q2, Q3, Q6, and Q7 were classified as moderately difficult (p = 0.43–0.57), reflecting meaningful improvement while suggesting that these concepts may still benefit from further instructional reinforcement.

2.9. Gain Score and Normalized Gain

Item	Gain	Normalized Gain (g)	Interpretation
Q1	5	0.83	High gain
Q2	1	0.2	Low gain
Q3	3	0.5	Medium gain
Q4	5	0.83	High gain
Q5	3	0.6	Medium gain
Q6	4	0.57	Medium gain
Q7	3	0.43	Medium gain
Q8	4	0.67	Medium–High gain

The results show varying levels of learning gains across items. Items Q1 and Q4 recorded the highest gain scores (Gain = 5) and achieved high normalized gains (g = 0.83), indicating that students demonstrated substantial

DOI: 10.47505/IJRSS.2026.1.2

improvement and near-mastery of these concepts following instruction. These items appear to be the most effectively addressed by the intervention. Items Q3, Q5, Q6, and Q7 yielded medium normalized gains (g = 0.43–0.60), suggesting moderate yet meaningful learning progress. This indicates that while students' understanding of these concepts improved considerably, there remains room for further reinforcement or extended practice to achieve higher mastery levels. Item Q8 exhibited a medium–high normalized gain (g = 0.67), reflecting strong instructional impact, though not as pronounced as the highest-gain items. In contrast, Q2 showed the lowest gain (Gain = 1; g = 0.20), categorized as a low gain, indicating limited improvement. This suggests that the concept measured by Q2 may require alternative instructional strategies, clearer explanations, or additional learning activities.

2.10. Wilcoxon Signed-Rank Test	
Statistic	**Value**
Sample size (N)	7
Sum of positive ranks (W)	28
Z value	2.37
p-value	0.0078

The results show a sum of positive ranks (W) of 28, indicating that all observed differences favored the posttest scores over the pretest scores. The computed Z value of 2.37 reflects a substantial difference between the two testing occasions. Moreover, the p-value of 0.0078 is well below the conventional significance level of 0.05, leading to the rejection of the null hypothesis. These findings indicate that there is a statistically significant improvement in students' performance following the intervention. The consistent direction of change across all participants suggests that the observed improvement was not due to chance but rather reflects the effectiveness of the instructional strategy or learning activity implemented.

2.11 Effect Size of the Wilcoxon Signed-Rank Test (Pre-test vs Post-test Scores)	
Statistic	**Value**
Sample Size (N)	7
Z value	2.37
√N	2.65
Effect Size (r)	0.89
Magnitude of Effect	Very Large

Using the Wilcoxon test results, the effect size was calculated as $r = Z/\sqrt{N}$. With a Z value of 2.37 and a sample size of 7 ($\sqrt{N} = 2.65$), the resulting effect size (r) was 0.89. According to established benchmarks (0.10 = small, 0.30 = medium, 0.50 = large), this value represents a very large effect. This finding indicates that the intervention had a strong and meaningful impact on students' learning outcomes. The very large effect size suggests that the observed improvement in post-test scores was not only statistically significant but also educationally substantial. In practical terms, the intervention produced a pronounced enhancement in students' understanding of the subject matter.

3.2. Data Analysis and Interpretation of Effectiveness

The lesson plan was rated by experts as generally highly valid, with most indicators obtaining high mean scores (M = 3.20–3.60) and low standard deviations, indicating consistent evaluations, particularly in lesson structure, clarity, practice opportunities, real-life applications, and promotion of higher-order thinking; however, the formulation of SMART objectives was only moderately valid and requires improvement. Despite favorable descriptive ratings, the content validity results revealed limited expert consensus, as all I-CVI values (0.20–0.60) and the S-CVI/Ave (0.39) fell below acceptable standards, indicating the need for substantial revision in learner engagement, assessment alignment, cultural relevance, and instructional clarity. The validation instrument demonstrated excellent reliability (α

DOI: 10.47505/IJRSS.2026.1.2

= 0.97), confirming consistency of expert ratings, while the Friedman test showed no significant differences across validity dimensions ($p = 0.368$) and Kendall's W indicated only fair agreement ($W = 0.20$).

The frequency and percentage analysis showed unanimous expert agreement in evaluating the Learning Activity Sheet, with all five validators rating all 23 indicators as valid, resulting in 100% "Yes" responses (115/115). Consistent with this, the Content Validity Index results indicated excellent validity, as all items obtained an I-CVI of 1.00 and the S-CVI/Ave was 1.00, exceeding the acceptable standard. The inter-rater agreement further confirmed perfect consistency among experts, with a percent agreement of 100%, demonstrating highly reliable evaluations. Domain-level analysis likewise revealed excellent validity across Content, Organization/Presentation/Language, and Design/Layout, each achieving a mean I-CVI of 1.00 and full agreement.

For the pre-test and post-test, all indicators under Organization, Clarity, and Relevance were rated as Highly Valid, with mean scores ranging from 3.6 to 4.0 and moderate standard deviations, indicating generally consistent expert judgments and strong clarity, structure, and alignment of the pre-test and post-test with study objectives. However, content validity analysis revealed varying levels of expert agreement, as I-CVI values ranged from 0.40 to 0.80, resulting in a moderate S-CVI/Ave of 0.60, which falls below the recommended standard and suggests the need for refinement in some items, particularly in organization and clarity. Despite this, the validation instrument demonstrated excellent reliability ($\alpha = 0.98$), confirming consistency in expert ratings. The Friedman test showed no significant differences among validity criteria ($p = 0.135$), while Kendall's W indicated a moderate level of agreement ($W = 0.40$).

The analysis of Learning Activity Sheet Task 1 (HOTS–SOLO questions) shows that students demonstrated moderate overall performance, with mean scores ranging from 2.14 to 2.86 on a 4-point scale. Item Q3 was the easiest ($p = 0.71$) and showed more consistent responses, while Q2 and Q4 were moderately difficult ($p = 0.54$), indicating areas where further instructional support may be needed. Despite slight variations in item difficulty, the Friedman test revealed no significant differences in student performance across items ($\chi^2 = 6.29$, $p = 0.099$), suggesting that the tasks were generally comparable in difficulty. Kendall's W of 0.30 indicates a fair level of consistency in student rankings across items. Overall, the results suggest that the HOTS–SOLO tasks are appropriately designed, function consistently, and are suitable for assessing students' higher-order thinking skills, while allowing room for instructional refinement in selected items.

The analysis of Learning Activity Sheet Task 2 (Follow-Up Questions) indicates that students demonstrated low to moderate performance, with mean scores ranging from 1.71 to 2.29, reflecting varying levels of understanding across items. Item difficulty indices classified all questions as moderate to moderately difficult ($p = 0.43–0.57$), suggesting that the tasks posed an appropriate level of challenge while still allowing discrimination among learners. Although some variability in responses was observed, particularly in Q3, the Friedman test revealed no significant differences in performance across the four items ($\chi^2 = 5.15$, $p = 0.161$), indicating comparable item difficulty. Kendall's W of 0.25 reflects a fair level of consistency in student performance across items.

The pre-test and post-test results demonstrate a substantial and statistically significant improvement in students' performance following the intervention. Pretest findings showed very low correct response rates and item difficulty indices, indicating limited prior knowledge, with all items classified as difficult to extremely difficult. In contrast, posttest results revealed marked increases in correct responses and item difficulty indices shifting to moderate or easy levels, reflecting improved conceptual understanding. Gain and normalized gain analyses further showed medium to high learning gains across most items, with Q1 and Q4 exhibiting the strongest improvement, while Q2 showed comparatively lower gains, suggesting a need for further instructional support. The Wilcoxon Signed-Rank Test confirmed a significant difference between pretest and posttest scores ($Z = 2.37$, $p = 0.0078$), and the very large effect size ($r = 0.89$) indicates that the intervention had a strong and educationally meaningful impact on students' learning outcomes.

3.3. Reflection of the Learners to the Activity

This part of the paper presents the learners' perceptions and self-reflections regarding the implemented lesson and learning activities. Students' responses provide insights into their experiences, engagement, and perceived learning outcomes. To ensure a systematic and objective analysis, appropriate statistical tools were employed to

DOI: 10.47505/IJRSS.2026.1.2

further examine the data. Descriptive statistics were used to summarize response patterns, while qualitative techniques such as thematic analysis and frequency counts were applied to identify recurring themes and trends in learners' reflections. The use of these analytical tools strengthened the interpretation of the data and allowed for a clearer understanding of learners' perceptions and reflective insights.

3.3.1. Perception of the Learners

Table 3.1. Descriptive Statistics (After Reverse Scoring)

Item	Mean	SD	Min	Max
Value	3.29	1.11	1	4
Choice (I had a choice)	3.14	1.07	1	4
Enjoyment	3	0.58	2	4
Concentration	3.57	1.13	1	4
Fun	3	1.15	1	4
Importance	3.14	1.07	1	4
Interest	3.14	1.07	1	4
Willingness to repeat	3	0.82	1	4

Table 3.1 presents the descriptive statistics of learners' perception scores after reverse scoring the negatively worded items. Overall, the mean scores range from 3.00 to 3.57 on a 4-point scale, indicating generally positive perceptions toward the learning activity. The highest mean was observed for Concentration (M = 3.57, SD = 1.13), suggesting that learners perceived the activity as effective in improving their focus. Items related to Value (M = 3.29), Choice, Importance, and Interest (M = 3.14) also reflect favorable views, indicating that students recognized the usefulness and relevance of the activity and felt a reasonable sense of autonomy. Enjoyment, Fun, and Willingness to Repeat all obtained a mean of 3.00, signifying that learners generally found the activity enjoyable and were open to engaging in it again. The standard deviations, ranging from **0.58 to 1.15**, indicate moderate variability in responses, suggesting some differences in individual experiences. Overall, the results demonstrate that learners perceived the activity as valuable, engaging, and supportive of learning.

Table 3.2. Reliability Analysis (Internal Consistency)

Statistic	Value
Number of items	25
Cronbach's α	0.91
Interpretation	Excellent reliability

Table 3.2 presents the results of the reliability analysis conducted to assess the internal consistency of the perception questionnaire. The instrument consisted of 25 items and yielded a Cronbach's alpha (α) of 0.91, which indicates excellent reliability. This high alpha value suggests that the items are strongly interrelated and consistently measure the same underlying construct of learners' perceptions toward the activity. Since Cronbach's alpha values above 0.70 are generally considered acceptable and values above 0.90 reflect very high reliability, the result confirms that the questionnaire is a dependable and stable measure. Consequently, the perception data gathered using this instrument can be considered reliable and suitable for further statistical analysis and interpretation.

Table 3.3. Friedman Test (Differences Across Items)

Statistic	Value
Friedman χ^2	36.48
df	24

DOI: 10.47505/IJRSS.2026.1.2

p-value	0.071
Kendall's W	0.26
Strength of agreement	Fair

Table 3.3 presents the results of the Friedman test and Kendall's coefficient of concordance used to examine differences in learners' perception ratings across the 25 questionnaire items. The Friedman test yielded a chi-square value of 36.48 with 24 degrees of freedom and a p-value of 0.071, which is greater than the 0.05 level of significance. This indicates that there is no statistically significant difference in learners' ratings across the perception items, suggesting that students responded relatively similarly to the different aspects of the activity. Kendall's coefficient of concordance (W = 0.26) indicates a fair level of agreement among learners' responses, reflecting moderate consistency in how they rated the items.

Table 3.4. Wilcoxon Signed-Rank Test (Post-hoc Pairwise Comparisons)

Item	Wilcoxon p-value
Autonomy-related items	0.016*
Enjoyment-related items	0.078
Value-related items	0.375
Usefulness-related items	0.297

Table 3.4 presents the results of the Wilcoxon Signed-Rank Test conducted as post-hoc pairwise comparisons to further examine differences among clusters of perception items after the Friedman test. The results show that autonomy-related items yielded a statistically significant difference ($p = 0.016 < 0.05$), indicating that learners' perceptions of autonomy (e.g., having choice or control during the activity) differed meaningfully compared with other item groupings. This suggests that the activity had a particularly strong or distinct impact on students' sense of autonomy. In contrast, enjoyment-related ($p = 0.078$), value-related ($p = 0.375$), and usefulness-related items ($p = 0.297$) did not show statistically significant differences, implying that learners' ratings across these dimensions were relatively consistent.

3.3.1. Self-reflection of the Learners

Table 3.4. What is the most important thing you learned in the lesson?

Theme	Description	f	%
Understanding Natural Selection	Learning the process, theories, and mechanisms of natural selection	5	71.43
Genetic Variation as Basis of Survival	Recognizing genetic variation as essential to survival and selection	1	14.29
General Life Insight	Viewing the lesson as a broader life concept (e.g., "life is a wheel")	1	14.29

Table 3.4 shows the thematic analysis of learners' responses regarding the most important thing they learned from the lesson. The dominant theme, Understanding Natural Selection, was identified in the majority of responses (f = 5, 71.43%), indicating that most learners gained a clear understanding of the process, theories, and mechanisms underlying natural selection. This suggests that the lesson was effective in achieving its primary conceptual objective. A smaller proportion of learners (f = 1, 14.29%) emphasized Genetic Variation as the Basis of Survival, highlighting an understanding of genetic variation as a key factor in natural selection and species survival, which reflects deeper conceptual insight into evolutionary principles. Another learner (f = 1, 14.29%) expressed a General Life Insight, interpreting the lesson in a broader, metaphorical sense (e.g., "life is a wheel"), suggesting personal reflection and meaning-making beyond scientific content.

DOI: 10.47505/IJRSS.2026.1.2

Table 3.5. Do you wish you had spent more time on or done differently in the lesson? Why?

Theme	Description	f	%
Desire for More Time / Deeper Learning	Wanted more time to explore or learn more	4	57.14
No Changes Needed	Satisfied with the lesson as implemented	3	42.86

Table 3.5 presents the thematic analysis of learners' responses regarding whether they wished to spend more time on or change any aspect of the lesson. The majority of learners (f = 4, 57.14%) expressed a Desire for More Time or Deeper Learning, indicating that the lesson stimulated their interest and curiosity, leading them to want further exploration or extended discussion of the topic. This suggests that the lesson was engaging and intellectually motivating, but may benefit from additional time or enrichment activities to deepen understanding. Meanwhile, a substantial proportion of learners (f = 3, 42.86%) reported No Changes Needed, reflecting satisfaction with the lesson as implemented and suggesting that the instructional design effectively met their learning needs.

Table 3.6. What was the most enjoyable part of the lesson and why?

Theme	Description	f	%
Learning About Natural Selection	Enjoyed understanding processes, applications, and survival	4	57.14
Online / Digital Learning Experience	Enjoyed online mode or use of devices	1	14.29
Break Time / Social Aspect	Enjoyed break time and informal activities	1	14.29
Overall Lesson Experience	Enjoyed the lesson as a whole	1	14.29

Table 3.6 presents the thematic analysis of learners' responses regarding the most enjoyable part of the lesson. The majority of learners (f = 4, 57.14%) identified Learning About Natural Selection as the most enjoyable aspect, highlighting their interest in understanding the processes, real-world applications, and survival mechanisms related to the topic. This indicates that the core content of the lesson was engaging and meaningful to students. Smaller proportions of learners cited the Online/Digital Learning Experience, Break Time or Social Aspect, and the Overall Lesson Experience (each f = 1, 14.29%), suggesting that while instructional delivery, social interaction, and general classroom atmosphere contributed to enjoyment, these were secondary to content learning.

Table 3.7. What was the least enjoyable part of the lesson and why?

Theme	Description	f	%
No Least Enjoyable Part	Reported no negative experience	4	57.14
Assessment-Related Issues	Quiz or lack of review time	1	14.29
Technical Issues	Weak internet connection	1	14.29
None / Minimal Concern	Explicitly stated "nothing"	1	14.29

Table 3.7 summarizes learners' perceptions of the least enjoyable part of the lesson. Most learners (f = 4, 57.14%) reported no least enjoyable part, indicating that they did not experience any significant difficulties or negative aspects during the lesson. This suggests a generally positive learning experience. A small number of learners identified assessment-related issues (f = 1, 14.29), specifically the quiz and limited review time, implying that pacing or assessment preparation may need slight adjustment. Technical issues, such as weak internet connectivity, were also mentioned by one learner (14.29), reflecting external factors that affected the learning experience rather than the lesson design itself. Another learner explicitly stated none or minimal concern (f = 1, 14.29), further reinforcing the overall favorable perception of the lesson. Collectively, these results indicate high learner satisfaction, with only minor areas for improvement related to assessment pacing and technical support.

Table 3.8. What can you suggest for the improvement of the lesson?

Theme	Description	f	%

DOI: 10.47505/IJRSS.2026.1.2

Increase Engagement / Interactivity	Suggested animations, simulations, or more engagement	3	42.86
No Improvement Needed	Satisfied with the lesson as is	3	42.86
Improve Learning Conditions	Suggested better internet or more breaks	1	14.29

Table 3.8 presents learners' suggestions for improving the lesson. A substantial proportion of learners (f = 3, 42.86%) recommended increasing engagement and interactivity, such as incorporating animations, simulations, or more interactive activities, indicating a desire for richer and more dynamic learning experiences. An equal number of learners (f = 3, 42.86%) stated that no improvement was needed, reflecting high satisfaction with the lesson's overall design and implementation. Meanwhile, a smaller group (f = 1, 14.29%) suggested improving learning conditions, particularly through better internet connectivity or additional break time, pointing to contextual or logistical factors rather than instructional shortcomings.

4. DISCUSSIONS

Although experts generally rated the lesson plan as *highly valid* based on mean scores—indicating strengths in structure, clarity, real-life application, and promotion of higher-order thinking—the content validity analysis revealed insufficient expert consensus, with all CVI values falling below acceptable thresholds. This indicates that the lesson plan requires substantial revision, particularly in SMART objective formulation, learner engagement, assessment alignment, cultural relevance, and instructional clarity, despite the evaluation instrument itself being highly reliable and expert ratings showing no significant differences across dimensions. Although experts generally rated the lesson plan as highly valid based on mean scores—indicating strengths in structure, clarity, real-life application, and the promotion of higher-order thinking—the content validity analysis revealed insufficient expert consensus, as all I-CVI and S-CVI values fell below recommended thresholds. Previous studies emphasize that low CVI values signal the need for systematic revision of instructional materials, particularly in item clarity, relevance, and alignment with learning objectives (Lau et al., 2023; Coppi et al., 2023; Cocchi et al., 2023). This finding suggests that the lesson plan requires substantial refinement, especially in the formulation of SMART learning objectives, learner engagement strategies, assessment alignment, cultural relevance, and instructional clarity, despite the evaluation instrument demonstrating high reliability. Research further highlights that clearly articulated, action-oriented SMART objectives are essential for ensuring coherence between instruction and assessment and for strengthening content validity during instructional design revisions (Masuwai et al., 2024; Tsunami et al., 2024). Moreover, the absence of significant differences across validity dimensions and only fair expert agreement reinforces the need for targeted revisions rather than reliance solely on high mean ratings, as expert consensus is a critical indicator of content validity in educational research (Lau et al., 2023; Cocchi et al., 2023).

The Learning Activity Sheet demonstrated excellent content validity and perfect expert agreement, as evidenced by unanimous "Yes" ratings across all indicators, I-CVI and S-CVI/Ave values of 1.00, and 100% inter-rater agreement, confirming that the material is highly valid, consistently evaluated, and fully acceptable for instructional use across all domains. Comparable validation studies have also reported robust content validity with high I-CVI and S-CVI values following expert consensus (Du Wei et al., 2025; Higher Education Studies, 2025; Sapuan & Sulaiman, 2025). Methodological guidelines further underscore that rigorous CVI evaluation with expert panels ensures the instructional material's relevance and alignment with research constructs (Hakim et al., 2025). Similarly, high content validity indices have been documented in recent educational instrument development efforts, demonstrating consistent expert agreement in assessing item relevance and construct coverage (Ahmad, 2025; Noor & Sapuan, 2025).

The pre-test and post-test were perceived by experts as highly valid in terms of organization, clarity, and relevance, with consistently high mean ratings and excellent instrument reliability; however, content validity was only moderate, as reflected by sub-threshold I-CVI and S-CVI/Ave values, indicating that some test items require refinement, particularly in organization and clarity, despite overall consistency and moderate expert agreement. Content validity involves expert evaluation of item relevance and clarity, typically quantified using the Item-level

DOI: 10.47505/IJRSS.2026.1.2

Content Validity Index (I-CVI) and the Scale-level Content Validity Index (S-CVI/Ave). Established criteria recommend I-CVI values of **≥0.78** and S-CVI/Ave values of ≥0.80 to indicate robust content validity; instruments falling below these thresholds signal the need for refinement in item formulation and clarity (Hakim et al., 2025; Soriano et al., 2025). Soriano et al. (2025) found that although individual items met relevance criteria, the overall S-CVI/Ave was only 0.65, highlighting limitations in expert consensus and warranting item revision. Similarly, other studies emphasize iterative expert review to improve item clarity and construct representation when moderate CVI values are observed (International Journal of Innovative Research and Scientific Studies, 2025; Ceberio et al., 2025; Najmee et al., 2025). These findings support the interpretation that, despite strong organizational and clarity ratings, selected pre-test and post-test items require targeted refinement to enhance content validity and alignment with the construct being measured.

Learning Activity Sheet Task 1 (HOTS–SOLO questions) effectively assessed students' higher-order thinking skills, as performance was generally moderate and consistent across items, with no significant differences in difficulty; however, some items (particularly Q2 and Q4) may benefit from additional instructional support to further enhance student understanding. Learning Activity Sheet Task 1, which employed HOTS–SOLO–aligned questions, functioned effectively as an assessment of students' higher-order thinking skills, as reflected by the generally moderate mean scores and the absence of statistically significant differences in difficulty across items. This pattern suggests that the questions were appropriately calibrated to a similar cognitive level and were consistent in measuring analytical, relational, and extended abstract thinking, which are core outcomes of the SOLO taxonomy (Biggs & Tang, 2023; Hattie, 2023). The consistency in performance across items, further supported by a non-significant Friedman test and fair Kendall's W, aligns with recent findings indicating that well-designed HOTS assessments tend to produce comparable difficulty levels while still discriminating learners' depth of understanding (Koh et al., 2024; Saido et al., 2023). However, the moderately lower performance observed in specific items (notably Q2 and Q4) suggests that certain concepts or cognitive demands within these questions may require additional instructional scaffolding or clearer contextualization. Prior research emphasizes that even within well-constructed higher-order tasks, variations in item performance often signal the need for targeted instructional reinforcement rather than flaws in overall task design (Brookhart, 2024; Widodo et al., 2023). Consequently, while Task 1 is pedagogically sound and suitable for assessing higher-order thinking, refining instructional support and providing guided practice for selected items may further enhance students' conceptual understanding and performance consistency.

Learning Activity Sheet Task 2 presented an appropriate and consistent level of challenge, as students showed low to moderate performance across items with no significant differences in difficulty, indicating that the follow-up questions functioned uniformly while allowing reasonable discrimination of students' understanding. Learning Activity Sheet Task 2, which consisted of follow-up questions designed to reinforce and extend conceptual understanding, presented an appropriate and consistent level of challenge to learners. The observed low to moderate mean scores across items suggest that the questions were sufficiently demanding to promote cognitive engagement without being excessively difficult, a balance that is considered essential for effective formative and summative assessment (Brookhart, 2024). The absence of statistically significant differences in item difficulty, as indicated by the Friedman test, further demonstrates that the follow-up questions were well aligned in terms of cognitive demand and content focus, enabling a fair comparison of student performance across items. This consistency supports recent research emphasizing that well-designed follow-up or application-level questions should function uniformly while still allowing discrimination among varying levels of student understanding (Koh et al., 2024; Widodo et al., 2023). Moreover, the fair level of agreement reflected by Kendall's coefficient of concordance suggests moderate consistency in student rankings, which is expected when tasks are designed to capture individual differences in comprehension rather than produce identical response patterns (Hattie, 2023). Collectively, these findings indicate that Learning Activity Sheet Task 2 effectively fulfilled its instructional purpose by providing a coherent set of questions that consistently assessed learners' understanding while allowing meaningful differentiation in performance, thereby supporting its suitability for classroom use with potential for further refinement to enhance depth of learning.

DOI: 10.47505/IJRSS.2026.1.2

The pre-test and post-test results demonstrate a substantial and statistically significant improvement in students' performance following the intervention. Pretest findings showed very low correct response rates and item difficulty indices, indicating limited prior knowledge, with all items classified as difficult to extremely difficult. In contrast, posttest results revealed marked increases in correct responses and item difficulty indices shifting to moderate or easy levels, reflecting improved conceptual understanding. Gain and normalized gain analyses further showed medium to high learning gains across most items, with Q1 and Q4 exhibiting the strongest improvement, while Q2 showed comparatively lower gains, suggesting a need for further instructional support. The Wilcoxon Signed-Rank Test confirmed a significant difference between pretest and posttest scores ($Z = 2.37$, $p = 0.0078$), and the very large effect size ($r = 0.89$) indicates that the intervention had a strong and educationally meaningful impact on students' learning outcomes. The pre-test and post-test findings indicate a substantial and statistically significant improvement in students' learning outcomes following the instructional intervention. The very low correct response rates and item difficulty indices observed in the pretest suggest limited prior knowledge, a pattern commonly reported in studies assessing learners' baseline understanding of complex scientific concepts such as natural selection (Hattie, 2023). In contrast, the marked increase in posttest correct responses and the shift of item difficulty indices from difficult or extremely difficult to moderate or easy levels reflect meaningful gains in conceptual understanding, consistent with research showing that well-designed inquiry-based and SSI-integrated instruction can effectively enhance learning (Koh et al., 2024; Widodo et al., 2023). Further evidence of instructional effectiveness is provided by the gain and normalized gain analyses, which revealed medium to high learning gains across most items. The particularly strong improvements in Q1 and Q4 suggest that these concepts were well supported by the instructional design, while the comparatively lower gains in Q2 indicate the need for targeted reinforcement, a finding aligned with assessment literature emphasizing the diagnostic value of item-level gain analysis (Brookhart, 2024). The Wilcoxon Signed-Rank Test confirmed that the improvement from pretest to posttest was statistically significant ($Z = 2.37$, $p = 0.0078$), supporting the conclusion that the observed gains were not due to chance. Moreover, the very large effect size ($r = 0.89$) demonstrates that the intervention produced not only statistically significant but also educationally meaningful improvements, as effect sizes of this magnitude are indicative of strong instructional impact in educational research (Hattie, 2023). Collectively, these results provide robust evidence that the intervention substantially enhanced students' understanding and learning outcomes.

The learners demonstrated positive perceptions of the Microlesson Using an Inquiry-Based Learning Activity Sheets in a Flipped Classroom Setting, with mean ratings indicating high concentration, perceived value, interest, and enjoyment, supported by an excellent reliability of the perception instrument ($\alpha = 0.91$). Learners' responses across perception items were generally consistent, with no significant differences observed, although autonomy emerged as a distinct and meaningful dimension. Thematic analysis of self-reflections revealed that most learners achieved the core learning objective of understanding natural selection, found the lesson engaging and enjoyable, and expressed a desire for deeper learning or extended time. Few negative experiences were reported, and suggestions focused mainly on enhancing interactivity rather than revising content, indicating high satisfaction, strong engagement, and effective learning outcomes, with minor opportunities for enrichment and improved learning conditions. Learners' positive perceptions of the microlesson using inquiry-based Learning Activity Sheets in a flipped classroom setting indicate a high level of engagement and instructional effectiveness. The high mean ratings for concentration, perceived value, interest, and enjoyment, together with the excellent reliability of the perception instrument ($\alpha = 0.91$), suggest that the learning experience was both engaging and consistently perceived across respondents. This aligns with recent research showing that flipped and inquiry-based instructional designs promote learner engagement, motivation, and positive affect toward science learning by allowing students to actively construct knowledge and interact meaningfully with content (Lo & Hew, 2023; van Alten et al., 2024). The consistency of learners' responses across perception dimensions further indicates that the instructional design provided a coherent learning experience; however, the emergence of autonomy as a distinct dimension reflects the importance of learner control and self-directed learning, which has been identified as a key affordance of flipped classroom environments (Chen et al., 2023). The thematic analysis of learners' self-reflections strengthens these quantitative findings by providing qualitative evidence that most students achieved the core learning objective of understanding natural selection and perceived the lesson as engaging

DOI: 10.47505/IJRSS.2026.1.2

and enjoyable. Similar studies have shown that reflective activities in inquiry-based and SSI-oriented instruction support deeper conceptual understanding and help learners articulate their learning experiences more meaningfully (Zeidler & Sadler, 2023). Learners' expressed desire for extended time or deeper exploration suggests cognitive engagement beyond surface-level learning, indicating that the microlesson successfully stimulated curiosity and sustained interest (Hwang et al., 2024). The minimal negative feedback and the focus of suggestions on enhancing interactivity rather than revising content further imply high satisfaction with the instructional materials. Overall, these findings corroborate prior evidence that inquiry-based, flipped learning designs can foster positive learner perceptions, strong engagement, and effective learning outcomes, while still allowing room for enrichment through increased interactivity and improved learning conditions.

7. CONCLUSIONS

The findings of this study demonstrate that the Microlesson Using an Inquiry-Based Learning Activity Sheets in a Flipped Classroom Setting are generally effective and pedagogically sound, while also identifying specific areas for refinement. Although the lesson plan received high mean validity ratings from experts—highlighting strengths in structure, clarity, real-life relevance, and promotion of higher-order thinking—the low content validity indices indicate insufficient expert consensus, underscoring the need for substantial revision, particularly in the formulation of SMART objectives, learner engagement strategies, assessment alignment, cultural relevance, and instructional clarity. In contrast, the Learning Activity Sheet exhibited excellent content validity and perfect expert agreement, confirming its strong instructional quality and readiness for classroom implementation. The pre-test and post-test instruments were perceived as highly valid and reliable; however, moderate content validity suggests that selected items require refinement to improve clarity and organization. Student performance results further support the effectiveness of the intervention, as both LAS tasks functioned consistently and appropriately in assessing higher-order thinking and conceptual understanding. Most importantly, the pretest–posttest analysis revealed statistically significant and educationally meaningful learning gains, with a very large effect size, demonstrating the strong impact of the intervention on students' understanding of natural selection. Learners' perceptions and self-reflections corroborated these outcomes, indicating high engagement, positive learning experiences, and achievement of core learning objectives, alongside a desire for deeper exploration and enhanced interactivity. Overall, the results affirm the effectiveness of the SSI-integrated approach while highlighting targeted areas for improvement to further strengthen instructional quality and learning outcomes.

Recommendations

Based on the findings of the study, several recommendations are proposed to further strengthen the instructional quality and effectiveness of the microlesson using inquiry-based Learning Activity Sheets in a flipped classroom setting. First, the lesson plan should undergo substantial revision to improve content validity and expert consensus. Particular attention should be given to refining SMART learning objectives to ensure they are specific, measurable, aligned with assessment tasks, and clearly communicated to learners. Enhancing learner engagement strategies, strengthening alignment between objectives, activities, and assessments, and integrating culturally relevant examples will help improve instructional clarity and contextual relevance. Second, while the Learning Activity Sheet demonstrated excellent validity and can be retained for classroom implementation, its strengths should be leveraged to inform revisions of other instructional components, particularly the lesson plan and assessment tools. The LAS may serve as a model for consistency, clarity, and alignment across materials. Third, the pre-test and post-test instruments should be refined to improve content validity. Selected items, especially those with lower CVI values, should be revised for clearer wording, improved organization, and closer alignment with targeted learning outcomes. Item analysis results suggest that questions with lower gains, such as Q2, may require rephrasing or the inclusion of supporting instructional content. Fourth, additional instructional scaffolding should be provided for selected LAS Task 1 items (notably Q2 and Q4) to further support students' higher-order thinking and conceptual understanding. For LAS Task 2, enrichment activities or guided follow-up discussions may be incorporated to deepen learning while maintaining the appropriate level of challenge. Fifth, given the statistically significant and large learning gains observed, the microlesson approach should be sustained and potentially expanded to other biology topics. However,

DOI: 10.47505/IJRSS.2026.1.2

extended learning time, supplemental resources, or optional enrichment tasks are recommended to address learners' expressed desire for deeper exploration. Finally, learner feedback highlights the value of increased interactivity. Incorporating animations, simulations, collaborative tasks, and improved learning conditions—such as reliable internet access—can further enhance engagement, autonomy, and overall learning experiences.

ACKNOWLEDGMENT

The researcher would like to express sincere gratitude to Sir Douglas A. Salazar for his invaluable guidance, expertise, and continuous support throughout the conduct of this study. His insights, encouragement, and constructive feedback greatly contributed to the successful completion and refinement of this research. Heartfelt appreciation is also extended to the Pantar National High School ARAL Program students, who served as the respondents of this study. Their active participation, cooperation, and honest responses were essential in generating meaningful data and in achieving the objectives of the research. To all who contributed, directly or indirectly, to the realization of this study, the researcher is deeply thankful.

REFERENCES

Ahmad, Z. (2025). Measuring students' AI competence: Development and validation of the AI Competence Scale (AICS). Computers & Education (Elsevier).

Biggs, J., & Tang, C. (2023). Teaching for quality learning at university (5th ed.). McGraw-Hill Education.

Brookhart, S. M. (2024). How to assess higher-order thinking skills in the classroom. ASCD.

Ceberio, I., et al. (2025). Content and face validity in virtual reality assessment tools with children. Frontiers in Virtual Reality.

Chen, Y., Wang, Y., & Chen, N. S. (2023). Learner autonomy and engagement in flipped classrooms: A systematic review. Educational Research Review, 39, 100520. https://doi.org/10.1016/j.edurev.2023.100520

Cocchi, S., Cipolat Mis, C., Mazzocut, M., Barbieri, I., Bassi, M. C., Cavuto, S., Di Leo, S., Miraglia Raineri, A., Cafaro, V., & ETHIC Validation Group. (2023). Content and face validity of the Evaluation Tool of Health Information for Consumers (ETHIC): Getting health information accessible to patients and citizens. Healthcare, 11(8), 1154. https://doi.org/10.3390/healthcare11081154

Coppi, M., Fialho, I., & Cid, M. (2023). Developing a scientific literacy assessment instrument for Portuguese 3rd cycle students. Education Sciences, 13(9), 941. https://doi.org/10.3390/educsci13090941

Du Wei, Alias, B. S., & Abdul Wahab, J. L. (2025). Content validation and content validity index calculation for teacher innovativeness instruments among Henan private universities in China. Communications on Applied Nonlinear Analysis, 32(10s), 175–186.

Galindo-Domínguez, H., & Bezanilla, M.-J. (2025). A critical systematic review of the impact of the flipped classroom methodology on university students' autonomy. Trends in Higher Education, 4(2), 22. https://doi.org/10.3390/higheredu4020022

Hakim, N. A. M. L., Pairan, M. R., & Zakaria, M. I. (2025). Step-by-step guide to calculating the Content Validity Index (CVI) for single constructs using Excel. International Journal for Research and Innovation in Social Science, 8(…), 1717-1726.

Hattie, J. (2023). Visible learning: The sequel. Routledge.

DOI: 10.47505/IJRSS.2026.1.2

Higher Education Studies. (2025). Content and face validity assessment of an educational questionnaire. Higher Education Studies, 15(4).

Hwang, G. J., Chang, C. Y., & Chen, P. Y. (2024). Effects of inquiry-based learning on students' engagement, motivation, and conceptual understanding in science education. Computers & Education, 202, 104839. https://doi.org/10.1016/j.compedu.2023.104839

Koh, K. H., Tan, C., & Ng, P. T. (2024). Designing and validating higher-order thinking assessments in secondary science classrooms. Assessment in Education: Principles, Policy & Practice, 31(1), 1–18. https://doi.org/10.1080/0969594X.2023.2289456

Lau Yen Yen, E., Hashim, H., & Md Yunus, M. (2023). Development and validation of an instrument to evaluate technology-enhanced learning and teaching sustainability in teaching spelling. Sustainability, 15(5), 4256. https://doi.org/10.3390/su15054256

Lo, C. K., & Hew, K. F. (2023). A critical review of flipped classroom challenges and opportunities in K–12 and higher education. Educational Technology & Society, 26(1), 1–15.

Masuwai, A., Zulkifli, H., & Hamzah, M. I. (2024). Evaluation of content validity and face validity of Secondary School Islamic Education Teacher Self-Assessment Instrument (SSIET-SAI). Cogent Education (Taylor & Francis). https://doi.org/10.1080/2331186X.2024.2308410

Monib, W. K., Qazi, A., & Apong, R. A. (2024). Microlearning beyond boundaries: A systematic review and a novel framework for improving learning outcomes. Heliyon. https://doi.org/10.1016/j.heliyon.2024.e41413

Najmee, N. A. A., et al. (2025). Development and content validation of a needs assessment questionnaire. International Journal of the Introduction to Measurement and Methodology.

Noor, N. A. M., & Sapuan, W. N. M. (2025). Development and validation of an instrument to evaluate constructs for integrating AI into ELT. International Journal of Academic Research in Business and Social Sciences, 15(4).

Nur Ain Md Lukmanul Hakim, N. A., Pairan, M. R., & Zakaria, M. I. (2025). Step-by-step guide to calculating Content Validity Index (CVI) for single constructs using Excel. International Journal of Research and Innovation in Social Science, 9(3), 1717–1726.

Sam, R. (2024). Systematic review of inquiry-based learning: Assessing impact and best practices in education. F1000Research, 13, 1045. https://doi.org/10.12688/f1000research.155367

Silva, E. S., Costa, W. P. D., Lima, J. C. D., & Ferreira, J. C. (2025). Contribution of microlearning in basic education: A systematic review. Education Sciences, 15(3), 302. https://doi.org/10.3390/educsci15030302

Soriano, K., Nakatani, S., Onishi, K., Ito, H., Osaka, K., Betriana, F., & Yasahura, Y. (2025). Development and content validation of a Person-Centered Care Instrument for healthcare providers. Nursing Reports, 15(10), 355. https://doi.org/10.3390/nursrep15100355

Tsunami, C. K., Henríquez-Trujillo, A. R., Ferreira-Meyers, K., et al. (2024). Guidelines for integrating actionable

DOI: 10.47505/IJRSS.2026.1.2

A-SMART learning outcomes into the backward design process. MedEdPublish, 14. https://doi.org/10.12688/mep.20606

Urdanivia-Alarcón, D. A., Talavera-Mendoza, F., Rucano-Paucar, F. H., Cayani-Caceres, K. S., & Machaca Viza, R. (2023). Science and inquiry-based teaching and learning: A systematic review. Frontiers in Education, 8, 1170487. https://doi.org/10.3389/feduc.2023.1170487

van Alten, D. C. D., Phielix, C., Janssen, J., & Kester, L. (2024). Effects of flipped classrooms on learning outcomes and learner satisfaction: An updated meta-analysis. Educational Psychology Review, 36(1), 45–72. https://doi.org/10.1007/s10648-023-09759-9

Widodo, A., Riandi, R., & Surakusumah, W. (2023). Assessing students' higher-order thinking skills using SOLO-based tasks in science learning. Journal of Research in Science Teaching, 60(7), 1012–1030. https://doi.org/10.1002/tea.21823

Zeidler, D. L., & Sadler, T. D. (2023). Socio-scientific issues and reflective learning: Promoting engagement and deeper understanding in science education. Journal of Research in Science Teaching, 60(5), 645–664. https://doi.org/10.1002/tea.21802

DOI: 10.47505/IJRSS.2026.1.2