Jurnal Humaniora dan Ilmu Pendidikan (JAHIDIK)
ISSN: 2797-7803, Vol 5, No 2, 2026, 15-26 https://doi.org/10.35912/jahidik.v5i2.7036

Development of a Biology Learning Outcome Instrument on Ecosystem and Its Interactions Using an Inquiry

Lathifah Turrohmah[1*], Vikto Agus Wibowo[2], Tago Alam[3]
Universitas Muhammadiyah Metro, Indonesia[1,2,3]
lathifahtrr@gmail.com[1*], viktoagus25@gmail.com[2], tagokld@gmail.com[3]

Article History
Received on 12 November 2025
1st Revision on 27 November 2025
2nd Revision on 9 December 2025
3rd Revision on 21 December 2025
Accepted on 3 January 2026

Abstract

Purpose: This study aims to develop and examine the psychometric quality of a biology learning outcome instrument covering the cognitive, affective, and psychomotor domains on the topic of ecosystems and their interactions through an inquiry learning approach for tenth-grade senior high school students.

Methodology: The study employed a research and development design using a simplified developmental model comprising instrument planning, blueprint development, expert validation, field testing, and empirical analysis. The instrument was validated by two evaluation experts and subsequently tested on 30 tenth-grade students at a public Islamic senior high school in Metro City, Lampung, using a 30-item multiple-choice cognitive test, an affective observation sheet, and a psychomotor rubric.

Results: Expert validation produced a mean feasibility score of 84.5 percent, categorized as good. Empirical testing showed that ten of the thirty cognitive items were valid, with a KR-21 reliability coefficient of 0.731, indicating high reliability, while the affective and psychomotor instruments produced Cronbach alpha coefficients of 0.891 and 0.724 respectively. Item difficulty and discrimination analyses indicated a predominance of moderate-difficulty items with good to very good discrimination power.

Conclusions: The developed instrument is feasible for measuring biology learning outcomes through an inquiry approach, although the cognitive component requires item revision to raise its validity rate.

Limitations: The study was restricted to a single class and a single topic, limiting generalizability.

Contributions: This study contributes a validated, ready-to-use multidimensional assessment instrument and empirical evidence that can guide biology teachers and instrument developers in designing inquiry-based evaluation tools for ecosystem-related content.

Keywords: *Assessment Instrument, Biology Education, Ecosystem, Inquiry Learning, Item Validity*

How to Cite: Turrohmah, L., Wibowo, V. A., & Alam, T. (2026). Development of a Biology Learning Outcome Instrument on Ecosystem and Its Interactions Using an Inquiry. *Jurnal Humaniora dan Ilmu Pendidikan*, 5(2), 15-26.

1. Introduction

Education under the independent curriculum framework in Indonesia places strong emphasis on differentiated learning, a model designed to accommodate the diverse learning needs, readiness levels, and interests of individual students so that each learner can optimally construct an understanding of subject matter, particularly in biology. Differentiated instruction in biology is inseparable from the

dual demand of building conceptual understanding while simultaneously cultivating practical skills, since biological phenomena are best understood through direct observation, experimentation, and inquiry rather than rote memorization alone. In many schools, however, this differentiated approach has not been fully realized because teaching practice tends to remain teacher centered, with lesson plans functioning as administrative formalities rather than as genuine blueprints for instructional decision making regarding models, methods, media, and assessment strategies suited to the material being taught.

A persistent consequence of this gap is that students are frequently exposed to content enrichment activities that emphasize information delivery over conceptual engagement, leaving them with limited opportunity to internalize the underlying scientific ideas being taught. When instruction narrows to content transfer, the cognitive development that should accompany biology learning, namely the capacity to analyze, evaluate, and apply biological concepts, tends to stagnate. Because conceptual mastery is best revealed through students' ability to respond to well-constructed test items, evaluation becomes the primary lens through which teachers can judge whether instructional objectives have been met. Kurniawati and Fajaryati (2022) argue that instrument-based evaluation is central to determining whether biology learning has produced measurable competence, underscoring that no curriculum reform can be verified as effective without valid measurement tools accompanying it.

Assessment in biology education is conventionally organized around three interrelated domains, namely cognitive mastery of concepts, affective disposition toward scientific practice, and psychomotor skill in conducting biological procedures. Evaluating these three domains through examinations administered before, during, or at the end of a semester allows teachers to track the trajectory of student progress, yet such evaluation is only meaningful when the instruments used possess demonstrated validity and reliability. An instrument that has not undergone empirical scrutiny risks measuring constructs other than the ones intended, thereby producing misleading conclusions about student achievement. For this reason, teachers are increasingly called upon not merely to administer tests but to systematically develop and validate the very instruments they use, a responsibility that requires methodological competence beyond routine item writing.

Inquiry learning has been widely promoted as a pedagogical approach capable of bridging the gap between passive content transfer and active conceptual construction, since it invites students to formulate questions, investigate phenomena, discuss findings collaboratively, and reflect upon their conclusions. The stages typically associated with inquiry, namely orientation, problem formulation, investigation and discussion, and reflective conclusion, mirror the authentic practices of scientific inquiry and are considered particularly suited to biology topics that demand empirical verification, such as the study of ecosystems and the interactions among their biotic and abiotic components (Zhang et al., 2023). Because inquiry learning reorganizes the classroom experience around student-driven investigation, it correspondingly demands an assessment instrument calibrated to capture not only factual recall but also the process-oriented and attitudinal outcomes that this pedagogy is designed to cultivate.

Despite the conceptual appeal of inquiry-based instruction, empirical evidence on its effectiveness remains dependent on the quality of the measurement tools used to evaluate it. Numerous studies confirm that inquiry-based learning enhances conceptual understanding, engagement, and higher-order thinking in science classrooms (Oliver, McConney, & Woods-McConney, 2021; Fan et al., 2021), yet fewer studies have specifically addressed the psychometric development of assessment instruments tailored to inquiry-based ecosystem instruction at the senior high school level. This creates a methodological void in which teachers implementing inquiry approaches to ecosystem material often rely on generic, unvalidated test banks that were never intended to measure the specific competencies fostered by inquiry pedagogy. Addressing this void requires a study that not only designs a multidimensional instrument but also subjects it to rigorous expert and empirical validation.

Building on this rationale, the present study was designed to develop and evaluate a biology learning outcome instrument covering the cognitive, affective, and psychomotor domains for the topic of

ecosystems and their interactions, situated within an inquiry learning approach implemented in a tenth-grade senior high school classroom. The study poses three central questions, namely how the instrument for measuring cognitive, affective, and psychomotor outcomes on ecosystem material through inquiry learning should be developed, how valid and reliable the resulting instrument proves to be, and how the developed cognitive items perform in terms of difficulty level and discriminating power. Addressing these questions is intended to yield both a practically usable instrument for classroom teachers and a broader methodological contribution to the literature on assessment development within inquiry-based biology instruction.

2. Literature Review

2.1 Biology Learning and the Demand for Authentic Assessment

Education is fundamentally a deliberate and planned effort to cultivate an environment in which learners can actively develop their intellectual, spiritual, and social potential, a definition long enshrined in Indonesian education law and reaffirmed in subsequent curriculum policy documents. Within this broad mandate, biology instruction occupies a distinctive position because it requires learners not merely to memorize scientific findings but to internalize biological concepts and cultivate a scientific disposition toward observation and reasoning. When biology teaching is reduced to knowledge transmission, the twin objectives of conceptual mastery and scientific attitude formation are rarely achieved, a pattern documented in classroom-based studies showing that overly didactic instruction constrains students' capacity to apply biological reasoning to novel situations (Arisandy, Fitriani, & Ghassani, 2018).

The shift toward independent or emancipated curriculum policy in Indonesia, formally introduced at the end of 2019 and expanded through subsequent ministerial decrees, has intensified the expectation that biology instruction moves away from teacher-centered delivery toward differentiated, student-centered practice. Hadi, Marniati, Ngindana, Kurdi, and Fauziah (2023) describe this policy shift as a new paradigm that grants teachers latitude to design instruction responsive to local context and student readiness, while Wahyuni, Fitriyah, and Hasanah (2023) note that successful implementation depends heavily on teachers' capacity to translate curricular flexibility into concrete pedagogical strategies rather than superficial administrative compliance. Intiana, Prihartini, Handayani, and Faridi (2023) further argue that the curriculum's emphasis on competency-based outcomes obliges subject teachers, including biology teachers, to reconsider not only what is taught but how learning is assessed, since a differentiated curriculum without differentiated and valid assessment risks reverting to uniform, low-fidelity measurement practices.

Mulyadin, Khoiron, Ginanto, and Putra (2023) similarly caution that curriculum reform initiatives frequently falter at the level of classroom implementation because teachers lack structured guidance on translating policy aspirations into workable instructional tools, a concern echoed in Prakoso, Ramdani, and Rahmah (2021) survey of teacher perceptions toward the independent learning policy, which found that many educators supported the policy's intent yet felt underprepared to operationalize it through concrete assessment design. These findings collectively suggest that curriculum reform and instrument development are inseparable undertakings, since a pedagogical innovation such as inquiry learning cannot be meaningfully evaluated, refined, or scaled without an accompanying instrument capable of capturing its intended learning outcomes.

2.2 Inquiry Learning as a Pedagogical Framework for Ecosystem Instruction

Inquiry learning is grounded in the premise that students construct scientific understanding most effectively when they actively investigate authentic questions rather than passively receive predetermined answers. The pedagogical sequence commonly associated with inquiry, comprising orientation, problem formulation, investigation, and reflective conclusion, positions learners as active agents who generate hypotheses, gather evidence, and negotiate meaning through discussion, a process considered particularly well matched to biological topics requiring empirical verification such as ecosystem dynamics, energy transfer, and population interaction (Nunaki, Damopolii, Kandowangko, & Nusantari, 2019; Wahyuni, 2023). Empirical reviews synthesizing inquiry-based science instruction across multiple national contexts report consistent, if moderate, positive

associations between inquiry exposure and measures of conceptual achievement, scientific literacy, and engagement, though the strength of these associations varies according to implementation fidelity and the quality of instructional support (Oliver, McConney, & Woods-McConney, 2021; Agustanti, 2012).

Beyond conceptual gains, inquiry-based instruction has been associated with improved learner engagement across cognitive, affective, and psychomotor dimensions of participation, since the collaborative and investigative nature of inquiry activities naturally elicits emotional investment and procedural skill practice alongside factual learning (Fan, Chen, Nair, Garg, Yeom, Kregor, Yang, & Wang, 2021). This multidimensional impact reinforces the rationale for developing assessment instruments that likewise span the three domains rather than restricting measurement to cognitive recall, since an inquiry classroom that succeeds in building positive scientific attitudes or procedural competence would remain invisible to an evaluation system that measures cognition alone. Environmental and ecological topics, including ecosystem material, further benefit from this multidimensional framing because environmental literacy itself has been conceptualized in the literature as an integrated construct combining knowledge, attitude, and skill components necessary for learners to interpret and respond to ecological phenomena responsibly (Angreani, Saefudin, & Solihat, 2022; Sainab, & Herna, 2022).

2.3 Content Validity and Expert Judgment

Before an instrument is administered to students, its content must be scrutinized by subject-matter and measurement experts to ensure that the items adequately represent the construct they are intended to measure, a step commonly referred to as content validation. Aiken's V coefficient has become one of the most widely used quantitative methods for summarizing expert judgment into a single validity index, since it converts ordinal expert ratings into a standardized value bounded between zero and one, with higher values denoting stronger agreement on an item's relevance and clarity (Merino-Soto, 2023). Kania, Kusumah, Dahlan, Nurlaelah, Gürbüz, and Bonyah (2024) demonstrate the application of this coefficient in constructing a mathematics problem-solving instrument, showing that expert panels drawn from complementary disciplinary backgrounds can strengthen the credibility of content validity claims, a principle equally applicable to biology instruments that combine content expertise with measurement expertise.

Several recent studies extend this content validation logic to educational contexts closely analogous to the present study. Masuwai, Zulkifli, and Hamzah (2024) validated a self-assessment instrument for teachers using expert panels and reported strong content validity indices, while Wahyuni (2023) applied Aiken index analysis to project assessment instruments within an independent curriculum setting, concluding that structured expert review substantially improves item quality prior to field testing. Juryatina and Amrin (2021) likewise emphasize those expert-reviewed instruments demonstrate stronger content representativeness than instruments developed without formal expert scrutiny, reinforcing the methodological rationale for incorporating expert validation as an obligatory stage preceding empirical try-out in instrument development studies such as the present one (Amalia, Dianingati, & Annisaa, 2022).

2.4 Empirical Validity, Reliability, and Item Analysis

Following expert validation, an instrument must undergo empirical testing to establish item-level validity, typically through point-biserial or product-moment correlation between item scores and total scores, and internal consistency reliability, commonly estimated through the Kuder-Richardson formula 21 for dichotomously scored cognitive items or Cronbach's alpha for polytomously scored affective and psychomotor items. Rezigalla et al. (2024) demonstrate that item validity in multiple-choice tests is closely tied to distractor functioning, showing that poorly constructed distractors depress both difficulty and discrimination indices regardless of how carefully the item stem is written, a finding with direct relevance for interpreting invalid items in a newly developed cognitive test. Similarly, Srisomsak, Sitticharoon, Keadkraichaiwat, Meethes, and Inpaen (2025) report that item flaws frequently persist undetected when instrument developers rely solely on a single statistical threshold rather than triangulating difficulty, discrimination, and qualitative item review.

Item difficulty and discrimination indices remain central diagnostic tools in classical test theory for judging whether individual items function as intended. Salim (2022) found that senior high school examinations frequently contain items with difficulty values below the ideal range, indicating items that are too difficult for the tested population, yet still retain acceptable discrimination values above the conventional threshold of 0.20, suggesting that difficulty and discrimination do not always move in tandem. Abdellatif and Al-Shahrani (2019) similarly report that blueprinting practices, meaning the systematic mapping of items to cognitive levels and content areas prior to test construction, exert a measurable influence on the resulting difficulty and reliability profile of an examination, underscoring those instrument developers benefit from deliberate test blueprint design rather than ad hoc item writing.

Reliability estimation using KR-21 provides a conservative lower-bound estimate of internal consistency for tests composed of dichotomously scored items, and coefficients above 0.70 are conventionally regarded as indicating high reliability suitable for classroom-level decision making (Salih et al., 2020). For instruments measuring affective and psychomotor domains, where item responses are typically scored on ordinal rating scales rather than as correct or incorrect, Cronbach's alpha remains the dominant reliability index, with coefficients above 0.60 generally accepted as adequate and coefficients above 0.80 regarded as good in applied educational measurement (Yundarini, Sudana, & Astawan, 2022). These established benchmarks provide the interpretive framework against which the empirical results of the present instrument are evaluated in the results and discussion section (Schulz, Hultin, & Gyllencreutz, 2025).

2.5 Research Proposition

Taken together, the literature indicates that inquiry learning is theoretically well suited to ecosystem instruction because it mirrors the empirical, investigative character of ecological inquiry, that multidimensional assessment spanning cognitive, affective, and psychomotor domains is necessary to capture the full range of outcomes inquiry pedagogy is designed to produce, and that rigorous expert and empirical validation procedures are indispensable for establishing instrument credibility. Building on this synthesis, the present study advances the proposition that a systematically developed and expert-validated biology learning outcome instrument for ecosystem material will demonstrate acceptable to high reliability across its three domains, even if a substantial proportion of initially drafted cognitive items require revision or elimination following empirical item analysis, consistent with patterns reported in comparable instrument development studies.

3. Research Methodology

3.1 Research Design

This study employed a research and development design intended to produce and empirically validate a biology learning outcome instrument rather than to test a causal instructional intervention. The developmental procedure followed five sequential stages, namely need analysis and blueprint construction, expert content validation, instrument revision, field try-out, and empirical psychometric analysis. This sequence follows established practice in instrument development research, in which expert judgment and empirical testing are treated as complementary rather than substitutable sources of validity evidence (Kania et al., 2024).

3.2 Research Setting and Participants

The developed instrument was trialed with thirty tenth-grade students enrolled in the natural science stream at a public Islamic senior high school in Metro City, Lampung Province, Indonesia, all of whom had received instruction on ecosystem material delivered through an inquiry learning approach prior to the assessment. Content validation was conducted separately by two evaluation experts holding relevant qualifications in education and biology teaching, one affiliated with a madrasah institution and the other an experienced biology teacher at the participating senior high school, both of whom rated the instrument using structured evaluation sheets covering construction, content, and language criteria.

3.3 Instrument Development Procedure

Instrument construction began with the formulation of a test blueprint that mapped thirty multiple-choice cognitive items onto the expected learning outcome for the ecosystem topic within the national curriculum framework, distributing items across the cognitive levels of application, analysis, and evaluation according to a revised Bloom's taxonomy classification. In parallel, an affective observation instrument comprising six rating items and a psychomotor performance rubric comprising three rating items were constructed to measure students' scientific attitude and practical skill respectively during ecosystem-related inquiry activities such as identifying ecological components and constructing food chains from field observation data.

Following blueprint construction, the draft instrument was submitted to the two evaluation experts, who rated each item on a four-point scale addressing item construction clarity, absence of clues to the correct answer, absence of negative phrasing, and appropriateness of accompanying visual material where relevant. Expert ratings were aggregated into a percentage feasibility score, with scores above 80 percent conventionally interpreted as indicating that an instrument is suitable for empirical try-out without requiring further major revision (Masuwai, Zulkifli, & Hamzah, 2024).

3.4 Data Collection and Analytical Techniques

Following expert validation, the cognitive instrument was administered to the thirty participating students, and responses were scored dichotomously as correct or incorrect for each of the thirty items. Item validity was calculated using the point-biserial correlation coefficient, comparing each item's correlation with total test score against the critical r-table value at a 0.05 significance level for a sample of thirty respondents, namely 0.361. Items yielding a calculated coefficient equal to or greater than this critical value were classified as valid, while items falling below it were classified as invalid and excluded from the final reliability computation.

$$r11 = (k/(k-1))x(1-(M(k-M))/(kxVt)) \quad 1$$

Formula 1 show the reliability of the cognitive instrument was estimated using the Kuder-Richardson f-21 (KR-21), computed as shown, where k denotes the number of items, M denotes the mean total score, and Vt denotes the total score variance.

For the affective and psychomotor instruments, whose items were scored on ordinal rating scales rather than dichotomously, internal consistency reliability was estimated using Cronbach's alpha, computed as the ratio of item variance to total variance adjusted by the number of items, following the standard formulation widely used in applied educational measurement (Yundarini, Sudana, & Astawan, 2022). Item difficulty was calculated as the proportion of examinees who answered each valid item correctly, with values interpreted according to conventional categories of easy, moderate, and difficult, while item discrimination was calculated by comparing the proportion of correct responses between upper and lower scoring groups, with higher discrimination values indicating stronger capacity to differentiate high- from low-performing students (Rezigalla et al., 2024). All statistical computations were performed using spreadsheet software following standard classical test theory formulas.

4. Results and Discussions

4.1 Expert Validation Results

Expert validation of the instrument construction, conducted by the first evaluation expert, produced an aggregate feasibility score of 84.5 percent, placing the instrument in the good category and indicating that the item construction, absence of answer clues, and clarity of accompanying visual material met acceptable standards for classroom use. The second evaluation expert, a practicing biology teacher at the trial school, similarly rated the observational instrument for measuring student competence in identifying individual organisms and populations within an ecosystem as adequate for field implementation, with only minor recommendations concerning wording clarity. These findings are

consistent with prior instrument development studies reporting that expert panels combining pedagogical and subject-matter expertise tend to converge on high feasibility ratings once instruments have undergone at least one revision cycle (Kania et al., 2024).

4.2 Cognitive Instrument Validity

Empirical validity testing of the thirty cognitive items, conducted with thirty tenth-grade students, showed that ten items achieved a calculated correlation coefficient exceeding the critical r-table value of 0.361 and were therefore classified as valid, while the remaining twenty items fell below this threshold and were classified as invalid.

Table 1. Summarizes the item-level validity results for the ten items that met the validity criterion

Item Number	r-table	r-calculated	Category
1	0.361	0.390	Weak
2	0.361	0.458	Moderate
3	0.361	0.459	Moderate
14	0.361	0.463	Moderate
15	0.361	0.424	Moderate
20	0.361	0.509	Moderate
21	0.361	0.444	Moderate
24	0.361	0.525	Moderate
27	0.361	0.478	Moderate
30	0.361	0.374	Weak

Table 1 describes the ten valid cognitive items retained from the original pool of thirty, showing the calculated correlation coefficient for each item alongside the fixed critical value used as the validity threshold and the resulting validity category. Eight of the ten valid items fell into the moderate validity category, while two items were categorized as weak, together confirming that approximately one third of the original item pool possessed acceptable discriminating relationship with total test performance.

The relatively high proportion of invalid items, amounting to twenty of thirty items or roughly 67 percent of the initial pool, is not unusual in first-round instrument development and mirrors findings reported in comparable item analysis studies. Rezigalla et al. (2024) attribute similarly elevated invalidity rates to weak distractor construction, in which implausible or overly obvious distractors reduce an item's capacity to discriminate between high- and low-performing examinees regardless of the clarity of the item stem. Srisomsak et al. (2025) further note that items testing lower-order recall rather than applied reasoning tend disproportionately toward invalidity when administered to classes exposed to activity-based instruction such as inquiry learning, since students who have engaged in hands-on investigation may perform inconsistently on items that reward memorized terminology rather than conceptual application. Internal item review in the present study likewise suggested that several invalid items relied on recall of ecosystem terminology rather than application of ecological reasoning, a pattern consistent with this explanation (Thuneberg, Salmi, & Bogner, 2018).

4.3 Cognitive Instrument Reliability

Reliability analysis of the retained cognitive instrument using the KR-21 formula produced a coefficient of 0.731, exceeding the conventional threshold of 0.70 and thereby indicating that the instrument possesses high reliability suitable for measuring student cognitive achievement on ecosystem material. This result aligns with reliability benchmarks reported in comparable classroom-based instrument studies, where KR-21 coefficients between 0.70 and 0.80 are typically interpreted as demonstrating dependable internal consistency for locally developed achievement tests (Salih et al., 2020). The relatively high reliability observed despite the exclusion of twenty invalid items suggests that the ten retained items, though numerically few, functioned as a coherent and internally consistent measure of the intended cognitive construct.

4.4 Item Difficulty and Discrimination

Difficulty index analysis of the ten valid cognitive items, presented in Table 2, revealed a predominance of moderate-difficulty items alongside a smaller number of easy and difficult items, indicating a generally balanced spread of item difficulty across the retained item set.

Table 2. Difficulty index classification of the valid cognitive test items

Item Number	Difficulty Index	Category
1	0.933	Easy
2	0.833	Moderate
3	0.833	Moderate
14	0.766	Moderate
15	0.566	Difficult
20	0.633	Moderate
21	0.666	Moderate
24	0.733	Moderate
27	0.600	Moderate
30	0.533	Difficult

Table 2 shows that seven of the ten retained items, or approximately 70% of the set, fell into the moderate difficulty category, while one item was classified as easy and two items as difficult, a distribution generally regarded as favorable in classical test theory because moderate-difficulty items maximize the information a test provides about differences in student ability (Salim, 2022).

Discrimination index analysis, summarized narratively rather than in tabular form for brevity, indicated that six of the ten retained items achieved very good discrimination power, three items achieved good discrimination power, and one item achieved weak discrimination power, with no item falling into the unacceptable range. This pattern indicates that, notwithstanding the substantial number of items eliminated during validity screening, the surviving items are collectively well suited to distinguishing higher- from lower-achieving students, corroborating the earlier finding that difficulty and discrimination do not always covary in a simple linear fashion, since even the two items classified as difficult retained acceptable discrimination values (Abdellatif & Al-Shahrani, 2019).

4.5 Reliability of the Affective and Psychomotor Instruments

The affective instrument, comprising six rating items intended to measure students' scientific attitude during ecosystem-related inquiry activities, produced a Cronbach alpha coefficient of 0.891, exceeding the minimum threshold of 0.60 by a considerable margin and indicating good internal consistency. The psychomotor instrument, comprising three rating items intended to measure students' practical skill in identifying organisms and constructing ecological relationships, produced a Cronbach alpha coefficient of 0.724, likewise exceeding the acceptable threshold and confirming adequate reliability.

Table 3. Summarizes the reliability coefficients obtained across the three assessed domains

Domain	Reliability Statistic	Coefficient	Interpretation
Cognitive	KR-21	0.731	High
Affective	Cronbach Alpha	0.891	Good
Psychomotor	Cronbach Alpha	0.724	Adequate

Table 3 consolidates the reliability evidence across all three measured domains, demonstrating that despite differences in item count and scoring method, each component of the instrument achieved reliability coefficients within an acceptable to good range according to established benchmarks in educational measurement (Yundarini, Sudana, & Astawan, 2022). The comparatively higher reliability of the affective instrument relative to the psychomotor instrument may reflect the smaller number of psychomotor items, since reliability coefficients calculated from short scales tend to be more sensitive

to individual item variance than coefficients calculated from longer scales, a statistical property well documented in classical test theory.

4.6 Student Learning Outcomes Across the Three Domains

Application of the validated instrument to the thirty participating students indicated that, on the cognitive dimension, the majority of students achieved scores within the good category, while a smaller subset achieved scores within the very good category, with no student falling below the passing threshold established for the ecosystem topic. On the affective dimension, twenty-two of thirty students, or 73.3%, achieved ratings within the very good category, while the remaining eight students achieved ratings within the good category. On the psychomotor dimension, ten students, or 40%, achieved ratings within the very good category, while twenty students, or 60%, achieved ratings within the good category, with no student falling into the adequate or poor categories on either the affective or psychomotor dimension.

Considered jointly across the three domains, eleven students, representing 36.6% of the class, achieved an overall composite rating of very good, while nineteen students, representing 63.3%, achieved an overall composite rating of good, indicating that all thirty students successfully met the minimum mastery expectation for the ecosystem topic when assessed through the newly developed instrument. This uniformly passing outcome is consistent with prior findings that inquiry-based instruction tends to elevate not only cognitive achievement but also affective engagement and procedural competence simultaneously, since the collaborative, investigative structure of inquiry activities provides multiple avenues for students to demonstrate mastery beyond written test performance alone (Fan et al., 2021; Nunaki, Damopolii, Kandowangko, & Nusantari, 2019).

The comparatively stronger performance observed on the affective dimension relative to the psychomotor dimension may be explained by the nature of the ecosystem topic itself, which affords ample opportunity for students to express curiosity, cooperation, and environmental concern through classroom discussion, yet offers comparatively fewer opportunities for fine-grained procedural skill practice compared to laboratory-intensive biology topics such as microscopy or dissection. This interpretation aligns with the broader environmental literacy literature, which frames ecosystem understanding as an integrated construct in which attitudinal engagement often develops ahead of procedural skill mastery, particularly when instructional time devoted to hands-on ecological fieldwork remains limited relative to classroom discussion time (Angreani, Saefudin, & Solihat, 2022).

Taken together, the results provide converging evidence supporting the proposition advanced in the literature review, namely that a systematically developed and expert-validated instrument for ecosystem material would demonstrate acceptable to high reliability across its three domains even where a substantial share of initially drafted cognitive items required elimination following empirical analysis. The findings additionally reinforce the broader methodological argument that instrument development for inquiry-based biology instruction cannot be treated as a one-step exercise in item writing, but instead requires iterative expert review, empirical validity screening, and domain-specific reliability estimation before an instrument can be considered ready for routine classroom use ((Sonia, Alberida, Arsih, & Selaras, 2023); (Marselina, Lasmawan, & Dantes, 2021)).

5. Conclusions

5.1 Conclusion

This study developed and empirically validated a multidimensional biology learning outcome instrument for the topic of ecosystems and their interactions, implemented within an inquiry learning approach for tenth-grade senior high school students. Expert validation confirmed that the instrument construction met an acceptable feasibility standard, while empirical try-out demonstrated that ten of thirty cognitive items achieved acceptable validity, yielding a high KR-21 reliability coefficient of 0.731 for the retained cognitive instrument. The accompanying affective and psychomotor instruments likewise achieved good and adequate Cronbach alpha reliability coefficients respectively. Item difficulty and discrimination analysis further showed that the retained cognitive items were

predominantly of moderate difficulty with good to very good discriminating power. Overall, the developed instrument can be considered feasible for measuring biology learning outcomes on ecosystem material through an inquiry learning approach, provided that the cognitive component undergoes further revision to improve its overall validity rate in future applications.

5.2 Research Limitations

This study is subject to several limitations that constrain the generalizability of its findings. First, the empirical try-out involved a single class of thirty students at one senior high school, a sample size and setting too narrow to support broad generalization of the reported validity and reliability coefficients to other schools or student populations. Second, the study focused exclusively on the ecosystem topic within the biology curriculum, meaning that the psychometric performance of the instrument on other biology topics remains untested. Third, content validation relied on the judgment of only two evaluation experts, a panel size smaller than typically recommended for robust Aiken's V-type content validity analysis. Finally, the study did not employ more advanced psychometric approaches such as item response theory or confirmatory factor analysis, relying instead on classical test theory indices that, while widely used and interpretable, offer less precision than modern latent trait models.

5.3 Directions and Future Study

Future research should extend the empirical try-out of the developed instrument to larger and more diverse student samples across multiple schools to establish more generalizable validity and reliability estimates. Subsequent studies could also expand the content validation panel to include additional subject-matter and measurement experts, enabling the calculation of a formal content validity coefficient such as Aiken's V rather than relying solely on percentage-based feasibility scoring. Further work might likewise apply item response theory or Rasch modeling to the cognitive item set to obtain more precise estimates of item difficulty and discrimination independent of sample-specific score distributions. Finally, researchers are encouraged to adapt the multidimensional instrument design demonstrated in this study to other biology topics beyond ecosystems, thereby testing whether the cognitive item revision strategies identified here generalize across different content domains within an inquiry learning framework.

Acknowledgement

The authors would like to thank the evaluation experts and the biology teacher who participated in the content validation process, as well as the school administration and students who took part in the instrument try-out, for their valuable cooperation throughout this study.

Author Contributions

LT conceived the study, coordinated the overall research process, and led the writing of the manuscript. VAW contributed to instrument construction, data collection, and statistical analysis. TA contributed to expert validation coordination, data interpretation, and manuscript revision. All authors read and approved the final manuscript.

References

Abdellatif, H., & Al-Shahrani, A. M. (2019). Effect of blueprinting methods on test difficulty, discrimination, and reliability indices: Cross-sectional study in an integrated learning program. *Advances in Medical Education and Practice, 10, 23-30*. https://doi.org/10.2147/AMEP.S190827

Agustanti, T. H. (2012). Implementasi metode inquiry untuk meningkatkan hasil belajar biologi. *Jurnal Pendidikan IPA Indonesia*, *1*(1), 16-20. https://doi.org/10.15294/jpii.v1i1.2007

Amalia, R. N., Dianingati, R. S., & Annisaa, E. (2022). Pengaruh jumlah responden terhadap hasil uji validitas dan reliabilitas kuesioner pengetahuan dan perilaku swamedikasi. *Generics: Journal of Research in Pharmacy*, *2*(1), 9-15. https://doi.org/10.14710/genres.v2i1.12271

Angreani, A., Saefudin, S., & Solihat, R. (2022). Virtual laboratory based online learning: Improving environmental literacy in high school students. *JPBI (Jurnal Pendidikan Biologi Indonesia)*, *8*(1), 10-21. https://doi.org/10.22219/jpbi.v8i1.19343

Arisandy, D. A., Fitriani, L., & Ghassani, F. (2018). Efektivitas model pembelajaran inquiry terhadap hasil pembelajaran biologi siswa kelas X SMA Negeri 6 Lubuklinggau. *BIOEDUSAINS: Jurnal Pendidikan Biologi dan Sains*, *1*(1), 71-79. https://doi.org/10.31539/bioedusains.v1i1.267

Fan, S., Chen, L., Nair, M., Garg, S., Yeom, S., Kregor, G., Yang, Y., & Wang, Y. (2021). Revealing impact factors on student engagement: Learning analytics adoption in online and blended courses in higher education. *Education Sciences*, *11*(10), 608. https://doi.org/10.3390/educsci11100608

Hadi, A., Marniati, M., Ngindana, R., Kurdi, M. S., & Fauziah, F. (2023). New paradigm of merdeka belajar curriculum in schools. *Al-Ishlah: Jurnal Pendidikan*, *15*(2), 1497-1510. https://doi.org/10.35445/alishlah.v15i2.3126

Intiana, S. R. H., Prihartini, A. A., Handayani, F., Mar'i, M. I., & Faridi, K. (2023). Independent curriculum and the Indonesian language education throughout the era of society 5.0: A literature review. *Al-Ishlah: Jurnal Pendidikan*, *15*(1), 911-921. https://doi.org/10.35445/alishlah.v15i1.3140

Juryatina, J., & Amrin, A. (2021). Students' interest in learning Arabic: The role of teachers. *Journal of Management and Instruction of Education (JEMIN)*, *1*(1), 1-10. https://doi.org/10.22515/jemin.v1i1.3459

Kania, N., Kusumah, Y. S., Dahlan, J. A., Nurlaelah, E., Gürbüz, F., & Bonyah, E. (2024). Constructing and providing content validity evidence through the Aiken's V index based on the experts' judgments of the instrument to measure mathematical problem-solving skills. *REID (Research and Evaluation in Education)*, *10*(1), 64-79. https://doi.org/10.21831/reid.v10i1.71032

Kurniawati, D., & Fajaryati, N. (2022). Pengembangan instrumen evaluasi pembelajaran berbasis kompetensi. *Jurnal Penelitian dan Evaluasi Pendidikan*, *26*(1), 45-58. https://doi.org/10.21831/pep.v26i1.45210

Marselina, K. T., Lasmawan, I. W., & Dantes, N. (2021). Pengembangan instrumen kemampuan berpikir kritis dan hasil belajar IPS pada siswa kelas V SD. *Jurnal Penelitian dan Evaluasi Pendidikan Indonesia*, *11*(2), 118-129. https://doi.org/10.23887/jpepi.v11i2.620

Masuwai, A., Zulkifli, H., & Hamzah, M. I. (2024). Evaluation of the validity of the content and facial validity of the self-assessment instrument for junior high school Islamic education teachers. *Cogent Education*, *11*(1). https://doi.org/10.1080/2331186X.2024.2308410

Merino-Soto, C. (2023). Aiken's V coefficient: Differences in content validity assessments. *MHSalud*, *20*(1). https://doi.org/10.15359/mhs.20-1.3

Mulyadin, T., Khoiron, M., Ginanto, D., & Putra, K. A. (2023). Workshop on Kurikulum Merdeka (Freedom Curriculum): Dismantling theories and practices. Paper presented at BEMAS: Jurnal Bermasyarakat.

Nunaki, J. H., Damopolii, I., Kandowangko, N. Y., & Nusantari, E. (2019). The effectiveness of inquiry-based learning to train the students' metacognitive skills based on gender differences. *International Journal of Instruction*, *12*(2), 505-516. https://doi.org/10.29333/iji.2019.12232a

Oliver, M., McConney, A., & Woods-McConney, A. (2021). The efficacy of inquiry-based instruction in science: A comparative analysis of six countries using PISA 2015. *Research in Science Education, 51(Suppl 2), 595-616*. https://doi.org/10.1007/s11165-019-09901-0

Prakoso, B. H., Ramdani, Z., & Rahmah, B. (2021). Teacher's perception on Merdeka Belajar policy. *Indonesian Journal of Educational Assessment*, *3*(2), 65-77. https://doi.org/10.26499/ijea.v3i2.84

Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., ALGhamdi, M. A., Al Ameer, A. Y., Yahia, A. I. O., Mohammed, O. A., & Elkhalifa Adam, M. I. (2024). Item analysis: The impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education, 24, 445*. https://doi.org/10.1186/s12909-024-05433-y

Sainab, S., & Herna, H. (2022). Perangkat pembelajaran biologi bernuansa karakter melalui inquiry dalam pembelajaran daring. *Jurnal Biolokus*, *5*(1), 42-51. https://doi.org/10.30821/biolokus.v5i1.1274

Salih, K. E. M., Jibo, A., Ishaq, M., Khan, S., Mohammed, O. A., Al-Shahrani, A. M., & Abbas, M. (2020). Psychometric analysis of multiple-choice questions in an innovative curriculum in Kingdom of Saudi Arabia. *Journal of Family Medicine and Primary Care*, *9*(7), 3663-3668. https://doi.org/10.4103/jfmpc.jfmpc_358_20

Salim, J. I. C. (2022). Discrimination and difficulty indices of a senior high school examination. *Psychology and Education: A Multidisciplinary Journal, 6, 499-504*. https://doi.org/10.5281/zenodo.7486865

Schulz, F., Hultin, M., & Gyllencreutz, L. (2025). Self-assessment of learning outcomes in prehospital disaster response skills: Instrument development and validation for mass casualty incident training. *BMJ Open*, *15*(3). https://doi.org/10.1136/bmjopen-2024-098284

Sonia, T., Alberida, H., Arsih, F., & Selaras, G. (2023). Pengaruh model pembelajaran inkuiri terbimbing terhadap keterampilan berpikir kritis peserta didik pada pembelajaran biologi. *Bioilmi: Jurnal Pendidikan*, *9*(1), 78-86. https://doi.org/10.19109/bioilmi.v9i1.14081

Srisomsak, V., Sitticharoon, C., Keadkraichaiwat, I., Meethes, S., & Inpaen, I. (2025). Detection of flawed multiple-choice questions in preclinical medical education using item difficulty and discrimination indices: A six-year analysis. *BMC Medical Education, 25*. https://doi.org/10.1186/s12909-025-08204-5

Thuneberg, H. M., Salmi, H. S., & Bogner, F. X. (2018). How creativity, autonomy and visual reasoning contribute to cognitive learning in a STEAM hands-on inquiry-based math module. *Thinking Skills and Creativity, 29, 153-160*. https://doi.org/10.1016/j.tsc.2018.07.003

Wahyuni, S., Fitriyah, I., & Hasanah, I. (2023). *The implementation of Merdeka Belajar Curriculum at English department of Indonesian universities* (Report No. 10).

Wahyuni, V. (2023). Aiken index analysis to determine the content of project assessment instruments in Islamic religious learning and character in the independent curriculum. *Journal of Quality Assessment of Continuing Education*, *6*(2), 88-99. https://doi.org/10.32923/kjmp.v6i2.4043

Yundarini, N. K. S., Sudana, D. N., & Astawan, I. G. (2022). Affective assessment instrument to assess student attitudes towards Science, Technology, Engineering and Mathematics. *Journal of Education Research and Evaluation*, *6*(4), 615-624. https://doi.org/10.23887/jere.v6i4.47681

Zhang, L., Liu, X., & Feng, H. (2023). Development and validation of an instrument for assessing scientific literacy from junior to senior high school. *Disciplinary and Interdisciplinary Science Education Research, 5, 21*. https://doi.org/10.1186/s43031-023-00093-2