Submitted: 14/10/2025 - Accepted: 12/11/2025 - Published: 29/12/2025

# Mapping the Architecture of Misconceptions: Network Analysis and Confidence Calibration in Three-Tier Diagnostic Assessment with Implications for Targeted Instruction

**Adysti Eren Fattaah Nurrizqi[1], Lina Mahardiani[1], Riezky Maya Probosari[1]*, Nurma Yunita Indriyanti[1], Muhammad Nur Hudha[1], Kadek Dwi Hendratma Gunawan[1]**

*[1]Science Education Study Program, Faculty of Teacher Training and Education, Universitas Sebelas Maret, Indonesia*
** riezkymprobosari@staff.uns.ac.id*

DOI: 10.26417/

## Abstract

Traditional science assessments inadequately distinguish between lack of knowledge and confident misconceptions, limiting their diagnostic value for instructional design. This study developed and validated a three-tier multiple choice diagnostic instrument for Indonesian students studying ecology and biodiversity concepts, integrating content questions, reasoning probes, and confidence ratings using the Certainty of Response Index. We administered a 15-item instrument to 95 seventh-grade students at a public junior high school in Central Java, Indonesia. Validation employed multi-method approaches including classical test theory, Item Response Theory modeling, confidence calibration analysis, and network analysis of misconception co-occurrence patterns. Student responses were classified as Sound Understanding with Confidence (PKDY), Sound Understanding with Uncertainty (PKTTY), Misconceptions (M), or Lack of Knowledge (TPK). The instrument demonstrated acceptable reliability (Cronbach's α = 0.647-0.682) and high content validity (0.833). Classification revealed 37.4% PKDY, 33.9% misconceptions, 19.0% lack of knowledge, and 9.7% uncertain understanding. Students exhibited systematic overconfidence (OI = 0.024) with Dunning-Kruger effects. Network analysis identified four misconception communities with coherent clustering patterns rather than random errors. Two-parameter logistic IRT models provided superior fit compared to Rasch models. The methodology advances diagnostic assessment by revealing which misconceptions function as conceptual bridges in student thinking, providing actionable intelligence for targeted instruction. For educational practice, the identification of 33.9% confident misconceptions—which traditional binary

systems would simply mark as "incorrect"—demonstrates substantial gains in diagnostic precision essential for effective remediation. The framework offers replicable protocols for developing culturally responsive diagnostic tools with implications for teacher professional development and curriculum sequencing in diverse educational contexts.

**Keywords:** diagnostic assessment, three-tier multiple choice, misconceptions, confidence calibration, science education, Indonesia

## 1. Introduction

Scientific misconceptions carry profound implications extending far beyond classroom performance to shape national economic competitiveness, public health outcomes, and societal capacity to address climate change, pandemic preparedness, and environmental sustainability. In knowledge-intensive economies, persistent gaps in science learning constrain human capital development, limit innovation capacity, and perpetuate educational inequities that disproportionately affect disadvantaged populations (OECD, 2023). Yet traditional assessment approaches, while efficient for measuring factual recall, inadequately distinguish between lack of knowledge and confident misconceptions—systematically held alternative frameworks that resist conventional instruction and require targeted pedagogical intervention. Recent developments in educational measurement have recognized these limitations, particularly binary scoring systems' inability to differentiate random guessing, partial understanding, and robust misconceptions demanding fundamentally different instructional responses (Ma et al., 2025).

International assessments reveal the global scale of this challenge. PISA 2022 results, encompassing over 700,000 students across 81 countries, showed substantial proportions of 15-year-olds failing to demonstrate basic scientific literacy competencies (OECD, 2023). These findings underscore a fundamental disconnect between educational investments and learning outcomes, particularly in developing nations where rapid educational expansion has not yielded corresponding improvements in instructional quality. For countries pursuing economic transformation and social mobility through education, this disconnect poses significant barriers to development goals and workforce competitiveness in increasingly knowledge-driven global markets.

Indonesia exemplifies these challenges with particularly high stakes. As the world's fourth-most populous nation and Southeast Asia's largest economy, Indonesia's National Medium-Term Development Plan explicitly prioritizes human capital enhancement through educational quality improvements (Rahayu et al., 2019). Yet despite dramatic improvements in educational access—gross enrollment rates increasing from 85% to 105% for primary education between 1972 and 2015—learning outcomes remain concerning. In PISA 2015, 42% of Indonesian students

failed minimum standards across all domains, while TIMSS performance consistently fell below international averages, with Indonesia ranking 35th of 46 nations before withdrawing from subsequent cycles (Lowyinstitute, 2019; Soeharto et al., 2022). These patterns reflect structural challenges including teacher preparation inadequacies, limited resources, and assessment practices emphasizing memorization over conceptual understanding—barriers that particularly disadvantage students from rural and low-income communities.

These performance patterns reflect deeper structural challenges within Indonesian science education, including teacher preparation inadequacies, limited instructional resources, and assessment practices that emphasize memorization over conceptual understanding (Rahayu et al., 2019). The disconnect between educational ambitions and learning outcomes necessitates more sophisticated diagnostic approaches capable of identifying specific conceptual difficulties and informing targeted instructional interventions.

The scientific study of student misconceptions emerged from cognitive psychology research demonstrating that learners actively construct knowledge by integrating new information with existing mental frameworks (Pacaci et al., 2024). These alternative conceptual frameworks, while often scientifically inaccurate, frequently exhibit internal consistency and remarkable resistance to conventional instruction. Understanding the nature and persistence of misconceptions has become central to contemporary science education theory, with implications extending far beyond simple error correction.

Conceptual change theory provides the dominant theoretical framework for understanding how students revise their understanding of scientific phenomena. Traditional approaches emphasized cognitive conflict strategies, where students encountered evidence contradicting their existing beliefs. However, recent meta-analytic evidence suggests that the effectiveness of conceptual change interventions varies considerably based on implementation characteristics, subject domain, and assessment methodology (Pacaci et al., 2024). A comprehensive meta-analysis of 218 primary studies involving over 18,000 students revealed large overall effect sizes ($g = 1.10$) for conceptual change strategies, but with significant moderating effects related to intervention length, question type, and academic domain.

Contemporary approaches increasingly recognize that conceptual change occurs through multiple mechanisms, including cognitive bridging and ontological category shifts, rather than simple replacement of incorrect ideas with scientifically accurate ones (Naeem Sarwar et al., 2024). This theoretical evolution has profound implications for assessment design, suggesting that diagnostic instruments must capture not only the presence of misconceptions but also their underlying structure and coherence.

## 2. Literature Review

### 2.1 Evolution of Diagnostic Assessment Approaches

The development of sophisticated diagnostic assessment tools has paralleled advances in misconception research and educational measurement theory. Early diagnostic instruments relied primarily on open-ended questions and clinical interviews, providing rich qualitative data but proving difficult to implement at scale. The emergence of two-tier diagnostic tests represented a significant advancement, combining content questions with reasoning probes to distinguish between correct answers based on sound understanding versus lucky guessing (Treagust, 1988).

Three-tier diagnostic assessments extend this approach by incorporating confidence ratings, typically using the Certainty of Response Index (CRI) developed by Hasan et al. (1999). This additional tier enables researchers and educators to distinguish between confident misconceptions and uncertain guessing, providing crucial information for instructional planning. Recent systematic reviews have documented the growing adoption of multi-tier diagnostic approaches across various scientific domains, with particularly strong growth in publications during 2023 (Nurdiyanti et al., 2025).

The theoretical justification for three-tier assessments rests on metacognitive research demonstrating that confidence ratings provide valuable insights into student self-awareness and knowledge monitoring capabilities (Fleming & Daw, 2024). Well-calibrated learners demonstrate appropriate confidence levels relative to their actual performance, while poor calibration—particularly overconfidence—often indicates limited metacognitive awareness and may impede learning from corrective feedback.

### 2.2 The Indonesian Educational Context and "Kurikulum Merdeka"

Indonesia's educational transformation gained momentum following political democratization in 1998, culminating in the implementation of "Kurikulum Merdeka" (Independent Curriculum) beginning in 2022. This curricular framework emphasizes competency-based learning, student agency, and contextual relevance—principles aligned with contemporary international best practices in science education (Ministry of Education, Culture, Research and Technology, 2022). The curriculum organizes learning through phase-based progression rather than rigid grade levels, with Phase D (grades 7-9) emphasizing scientific inquiry capabilities, evidence-based reasoning, and environmental stewardship.

The Kurikulum Merdeka's emphasis on capaian pembelajaran (learning outcomes) that integrate content knowledge with practical competencies creates opportunities for more sophisticated assessment approaches. Rather than focusing solely on factual recall, the curriculum encourages evaluation of students' ability to apply scientific knowledge to real-world problems and engage in evidence-based reasoning. This

philosophical alignment creates favorable conditions for implementing diagnostic assessments that can capture the complexity of student understanding.

However, significant implementation challenges remain. Indonesia's status as the world's largest archipelagic country, spanning over 17,500 islands, creates logistical difficulties for educational quality assurance and teacher professional development. Rural and remote areas often lack adequately trained science teachers and instructional resources, while urban schools may have better facilities but face overcrowding and diverse student populations. These contextual factors influence both the need for and feasibility of sophisticated diagnostic assessment approaches.

### 2.3 Ecological Literacy and Biodiversity Education

Ecology and biodiversity education occupy particularly important positions within science curricula, given their relevance to environmental challenges and sustainable development goals. Indonesia's extraordinary biological diversity—encompassing both terrestrial and marine ecosystems—provides rich contexts for scientific learning while highlighting the practical importance of ecological literacy. However, ecological concepts present unique pedagogical challenges due to their inherently complex, multi-scale, and dynamic nature.

Students frequently struggle with understanding ecological relationships, energy flow, population dynamics, and human environmental impacts. Common misconceptions include anthropomorphic reasoning about animal behavior, linear thinking about food webs, and oversimplified notions of environmental balance (Yeo et al., 2022). These alternative frameworks often prove resistant to instruction, requiring targeted interventions informed by detailed diagnostic information about student thinking patterns.

The development of ecological literacy also intersects with broader environmental and sustainability education goals. Students must understand not only basic ecological principles but also their applications to contemporary environmental challenges, including climate change, biodiversity loss, and sustainable resource management. This applied dimension requires assessment approaches capable of evaluating students' ability to connect scientific knowledge with real-world environmental issues.

### 2.4 Methodological Innovations in Educational Assessment

Recent advances in educational measurement have expanded the analytical toolkit available for diagnostic assessment research. Traditional psychometric approaches, while valuable for establishing basic instrument quality, provide limited insights into the complex patterns of student understanding that characterize conceptual learning. Contemporary approaches increasingly integrate multiple analytical frameworks to provide more comprehensive validity evidence.

Network analysis represents one promising methodological innovation, offering new perspectives on how misconceptions co-occur and relate to one another (Stella et al., 2020). By modeling misconception patterns as networks of interconnected concepts, researchers can identify central misconceptions that may serve as intervention targets and explore how addressing specific conceptual difficulties might generate cascading improvements in understanding.

Confidence calibration analysis draws from metacognitive research to examine how accurately students assess their own performance. This approach provides insights into student self-awareness and may help explain why some misconceptions prove particularly resistant to instruction. Students who are overconfident in their incorrect understanding may be less likely to engage seriously with corrective feedback, while those with poor calibration may struggle to identify areas needing additional attention.

Item Response Theory (IRT) modeling offers advantages over classical test theory approaches by providing parameter estimates that are theoretically independent of sample characteristics. This independence is particularly valuable for diagnostic instruments intended for use across diverse populations or educational contexts. IRT approaches also enable sophisticated analyses of item functioning, including differential item functioning that might indicate bias or cultural inappropriateness.

### 2.5 Research Objectives and Contributions

This study addresses several critical gaps in the diagnostic assessment literature while contributing methodological innovations to science education research. First, we develop and validate a three-tier multiple choice diagnostic instrument specifically designed for Indonesian students studying ecology and biodiversity concepts within the "Kurikulum Merdeka" framework. This instrument addresses the lack of culturally responsive diagnostic tools available for Indonesian educators while maintaining alignment with international educational measurement standards.

Second, we implement a multi-method validation approach integrating classical test theory, Item Response Theory, network analysis, and confidence calibration examination. This comprehensive analytical framework provides robust validity evidence while demonstrating innovative approaches to diagnostic assessment research that may inform future instrument development efforts.

Third, we examine confidence calibration patterns among Indonesian students, contributing to the limited cross-cultural research on metacognitive monitoring in science education contexts. Understanding how confidence calibration varies across cultural and educational contexts has implications for instructional design and teacher preparation programs.

Fourth, we explore misconception co-occurrence patterns using network analysis approaches rarely applied in science education research. By modeling misconceptions as interconnected networks, we provide new insights into the

structure of student alternative frameworks and identify potential targets for instructional intervention.

### 2.6 Significance and Potential Impact

The development of sophisticated diagnostic assessment tools has implications extending beyond the immediate research context. For Indonesian educators, culturally appropriate diagnostic instruments provide practical resources for identifying student conceptual difficulties and informing instructional decisions. The integration of diagnostic assessment with Indonesia's competency-based curriculum framework demonstrates how contemporary measurement approaches can support educational reform initiatives.

For the international science education community, this study contributes methodological innovations and cross-cultural validity evidence that may inform diagnostic assessment development in other contexts. The multi-method validation approach demonstrates how contemporary analytical techniques can provide richer insights into student understanding patterns while maintaining scientific rigor.

More broadly, this research contributes to ongoing efforts to improve science education quality in developing nations through evidence-based assessment and instructional practices. As countries worldwide grapple with the challenges of educational quality improvement in the face of rapid enrollment expansion, sophisticated diagnostic approaches become increasingly important for ensuring that increased access to education translates into meaningful learning outcomes.

The urgency of these challenges is underscored by contemporary global developments, including environmental degradation, climate change, and technological transformation, all of which require scientifically literate populations capable of evidence-based reasoning and critical thinking. Diagnostic assessment tools that can accurately identify and address conceptual difficulties represent essential components of educational systems designed to prepare students for these complex challenges.

Through the development and validation of our three-tier diagnostic instrument, combined with innovative analytical approaches to understanding student misconception patterns and confidence calibration, this study contributes to the broader project of improving science education quality while respecting cultural and contextual diversity in educational settings.

## 3. Methodology

This study employed a cross-sectional instrument validation design integrating multiple psychometric frameworks to establish comprehensive validity evidence for a three-tier multiple choice diagnostic instrument. The validation approach followed contemporary educational measurement standards emphasizing validity as a unified concept supported by multiple sources of evidence (Kane, 2013). The research

addressed the need for culturally responsive diagnostic tools capable of distinguishing between lack of knowledge and confident misconceptions in Indonesian science education contexts.

## 3.1 Participants

*Sample Characteristics and Selection*

Participants comprised 95 seventh-grade students (ages 12-13 years, M = 12.9, SD = 0.4) from three intact classes (7D: n = 32, 7E: n = 32, 7F: n = 31) at a public junior high school in Solo Raya, Central Java, Indonesia. Gender distribution was balanced (48 males, 47 females). The school was selected through purposive sampling based on: (1) two-year implementation of "Kurikulum Merdeka" curriculum emphasizing scientific literacy, (2) participation in the "Sekolah Penggerak" (School Transformation) program providing teacher training in student-centered pedagogy, and (3) demographic representativeness of Central Java urban-peripheral communities.

The student population was ethnically homogeneous (>95% Javanese) with socioeconomic backgrounds typical of Indonesian middle-income communities. Parental education varied: 45% completed secondary education, 35% vocational/diploma qualifications, 20% bachelor's degrees or higher. Household income ranged from IDR 3-8 million monthly (USD 200-530), placing families in lower-middle to middle-income categories. Prior science achievement showed heterogeneous ability levels: 25% high achievers (grades ≥85), 50% moderate (70-84), 25% developing (<70).

All students completed prerequisite ecology instruction (16 hours classroom, 8 hours laboratory) covering biodiversity, ecosystem interactions, and conservation principles under "Kurikulum Merdeka" Phase D standards. Phase D learning outcomes require students to analyze organism-environment interactions, evaluate human ecosystem impacts, and apply evidence-based reasoning to environmental issues. Assessment occurred February 2024, providing 2-3 months retention interval from instruction.

The study received ethics approval from Universitas Sebelas Maret Faculty Ethics Committee and Solo Raya District Education Office. Informed consent was obtained from school administration, teachers, and parents/guardians (100% response rate). Students provided written assent following explanation of voluntary participation and confidentiality protections. No students declined or withdrew.

Sample size (n = 95) was justified through a priori power analysis: adequate for IRT 2PL model stability with 15 items (recommended n = 80-100; de Ayala, 2009), network analysis with 6.3:1 participant-to-item ratio (recommended minimum 3:1; Epskamp et al., 2018), and latent class analysis with power >0.80 for medium effects. Sensitivity analyses revealed no significant differences across classes in Tier 1

accuracy ($F(2,92) = 0.89$, $p = .414$), confidence levels ($F(2,92) = 0.76$, $p = .471$), or misconception rates ($\chi^2(2) = 3.24$, $p = .198$), supporting single-sample analysis.

*Educational Context*

Indonesia's "Kurikulum Merdeka", implemented since 2022, organizes learning through phase-based progression rather than rigid grade levels. Phase D (grades 7-9) emphasizes scientific inquiry capabilities, evidence-based reasoning, and environmental stewardship through *capaian pembelajaran* (learning outcomes) that integrate content knowledge with practical competencies. All participating students had completed prerequisite coursework in fundamental ecological concepts including biodiversity principles, ecosystem dynamics, and human-environment interactions through this curricular framework.

## 3.2 Ethical Considerations

The study received approval from the university ethics committee and school district review board. Informed consent was obtained from school administration, classroom teachers, and student guardians following Indonesian Ministry of Education guidelines for educational research. Students provided written assent immediately before data collection with explicit notification of voluntary participation and withdrawal rights without academic consequences.

## 3.3 Instrument Development

*Theoretical Framework for Three-Tier Design*

The instrument design was grounded in cognitive load theory and misconception research indicating that traditional single-tier assessments inadequately distinguish between random guessing, partial knowledge, and confident misconceptions (Gurel et al., 2015). The three-tier structure addresses limitations of conventional multiple-choice formats by separately assessing content knowledge, reasoning processes, and metacognitive confidence.

*Item Construction Protocol*

Item development followed a systematic five-stage process: (1) learning objective analysis and misconception literature review, (2) initial item generation based on Indonesian curriculum standards and documented alternative frameworks, (3) expert review and content validation, (4) pilot testing with cognitive interviews, and (5) final item refinement based on statistical and qualitative feedback.

*Tier 1 Development*

Tier 1 items presented scenario-based questions addressing factual and conceptual knowledge across five competency domains aligned with Phase D curriculum standards: biodiversity conservation importance (items 1, 5, 14), ecosystem component interactions (items 2, 3), human environmental impact (items 4, 11, 12, 15), environmental influences on organisms (items 6, 8, 9, 10, 13), and biodiversity

classification (item 7). Each item included four response options with one correct answer and three plausible distractors based on documented misconceptions.

*Tier 2 Development*

Tier 2 reasoning options were systematically constructed using multiple sources: (a) international misconception literature in ecology education, (b) analysis of errors in Indonesian science textbooks, (c) semi-structured interviews with 12 students from comparable schools, and (d) consultation with experienced Indonesian science teachers regarding common student difficulties. Each Tier 2 included one scientifically correct reasoning statement and three alternatives representing documented misconceptions or naive theories.

*Tier 3 Confidence Assessment*

Tier 3 employed the Certainty of Response Index (CRI) developed by Hasan et al. (1999), modified to a six-point scale (0 = totally guessed, 1 = almost guessed, 2 = not sure, 3 = sure, 4 = almost certain, 5 = certain). This scale enables differentiation between low-confidence guessing and high-confidence misconceptions, crucial for diagnostic purposes.

*Content Validation Process*

Content validity was established through structured expert judgment involving five specialists: three subject matter experts in ecology education with Indonesian curriculum expertise and two educational measurement specialists experienced in diagnostic assessment development. Experts independently evaluated each item using a standardized rubric addressing: (1) alignment with specified learning outcomes, (2) scientific accuracy of content and reasoning options, (3) appropriateness for target population, (4) clarity of language and instructions, (5) cultural relevance and sensitivity, and (6) technical quality of item construction. Expert ratings were analyzed using Gregory's (2007) content validity ratio procedure, with items requiring unanimous agreement for retention. Disagreements were resolved through structured discussion and item revision. Inter-rater agreement was assessed using Fleiss's kappa for multiple raters.

*Pilot Testing and Cognitive Validation*

Pilot testing was conducted with 30 seventh-grade students from a comparable school not participating in the main study. The pilot employed a think-aloud protocol during item completion followed by structured interviews addressing item clarity, cultural appropriateness, and cognitive processes during response selection. Pilot data were analyzed for: (1) response distribution patterns identifying items with inadequate discrimination, (2) completion time analysis ensuring feasibility within class periods, (3) student feedback on item clarity and cultural relevance, and (4) preliminary assessment of confidence calibration patterns. Based on pilot results, three items demonstrating poor discrimination or cultural inappropriateness were

eliminated, item language was refined for clarity while maintaining scientific accuracy, and Tier 2 options were adjusted to better capture Indonesian student misconceptions.

### 3.4 Data Collection Procedures

*Administration Protocol*

Data collection occurred over two consecutive weeks during regular science instruction periods to minimize disruption to academic schedules. Each class session followed identical protocols administered by trained research assistants under classroom teacher supervision. Students received standardized verbal and written instructions emphasizing: (1) honest responding based on their best understanding rather than guessing, (2) the diagnostic rather than evaluative purpose of the assessment, (3) the importance of confidence ratings for understanding their own learning, and (4) the voluntary nature of participation with option to withdraw without consequences. The 70-minute time allocation was determined through pilot testing to ensure adequate time for thoughtful responding without time pressure effects. No students required additional time, and mean completion time was 52 minutes (SD = 8.3).

*Response Classification System*

Student responses were classified using established three-tier diagnostic taxonomies (Caleon & Subramaniam, 2010) into four mutually exclusive categories based on response patterns across all three tiers:

Sound Understanding with Confidence (PKDY): Correct Tier 1 response, correct Tier 2 reasoning, and high confidence (CRI > 2.5), indicating accurate knowledge with appropriate confidence.

Sound Understanding with Low Confidence (PKTTY): Correct Tier 1 and Tier 2 responses with low confidence (CRI ≤ 2.5), suggesting accurate knowledge but poor calibration or test anxiety.

Misconceptions (M): Any combination involving incorrect reasoning with high confidence, including correct Tier 1 with incorrect Tier 2 and CRI > 2.5, or incorrect Tier 1 with any Tier 2 and CRI > 2.5, indicating confident but inaccurate alternative frameworks.

Lack of Knowledge (TPK): All remaining response patterns with low confidence, suggesting guessing or genuine uncertainty about the concepts.

### 3.5 Statistical Analysis

*Analytical Framework*

The study employed a multi-method validation approach integrating classical test theory, modern test theory, and advanced statistical modelling to address different

aspects of instrument quality and student understanding patterns. This comprehensive approach provides convergent validity evidence and addresses limitations inherent in any single analytical method.

*Classical Test Theory Analysis*

Classical analysis included: (1) internal consistency reliability using Cronbach's alpha for the overall instrument and individual tiers, (2) item-total correlations for both tiers to assess item discrimination, (3) item difficulty indices calculated as proportion correct for each tier, (4) distractor analysis examining frequency of selection for incorrect options, and (5) point-biserial correlations between items and total scores.

*Item Response Theory Modelling*

Two-parameter logistic (2PL) models were fitted separately to Tier 1 and Tier 2 binary response data using marginal maximum likelihood estimation with the expectation-maximization algorithm. The 2PL model was selected over one-parameter (Rasch) models based on theoretical considerations regarding varying item discrimination and empirical model comparison using likelihood ratio tests.

Model assumptions were evaluated through: (1) unidimensional assessment using parallel analysis of residual correlations, (2) local independence examination using Q3 statistics for item pairs, (3) monotonicity evaluation through non-parametric item response functions, and (4) invariance testing across demographic subgroups using differential item functioning analysis.

Model-data fit was assessed using: (1) global fit indices including -2 log likelihood, AIC, and BIC, (2) item-level S-$\chi^2$ statistics with Bonferroni correction for multiple comparisons, (3) RMSEA values for individual items, and (4) residual analysis examining standardized residuals between observed and expected response patterns.

*Confidence Calibration Analysis*

Calibration between confidence and accuracy was examined using multiple approaches developed in metacognitive research. The primary calibration metric was the Brier score (Brier, 1950), calculated as $BS = (1/n)\Sigma(ci - ai)^2$, where ci represents confidence and ai represents accuracy for response i. The Brier score was decomposed into reliability, resolution, and uncertainty components to identify sources of calibration bias.

Overconfidence was assessed using the overconfidence index (OI = mean confidence - mean accuracy) and analyzed across confidence levels and performance quartiles. Calibration curves were constructed plotting mean accuracy against mean confidence within confidence bins to visualize calibration patterns.

The Dunning-Kruger effect was examined by analyzing overconfidence patterns across ability quartiles, with particular attention to whether low-performing students demonstrated greater overconfidence than high-performing students.

*Network Analysis of Misconception Patterns*

Misconception co-occurrence networks were constructed where nodes represented individual test items (n = 15) and edges represented relationships between misconception patterns. Student responses were recoded into binary misconception indicators (1 = misconception present based on Response Classification System criteria, 0 = otherwise), yielding 95 × 15 matrices for each tier.

Network structures were estimated using correlation-based network analysis in R 4.3.2. Pearson correlations were calculated between all item pairs using pairwise complete observations. An adjacency matrix was constructed by retaining correlations with absolute values exceeding 0.30 ($|r| > 0.30$), consistent with moderate effect sizes in educational research. Networks were specified as undirected, weighted graphs using the igraph package (Csardi & Nepusz, 2006), with edge weights defined as absolute correlation coefficients.

Centrality measures were calculated to identify structurally important items: degree centrality, betweenness centrality (normalized), closeness centrality (normalized), eigenvector centrality, and strength. Community detection employed the Walktrap algorithm (Pons & Latapy, 2006) to identify clusters of related misconceptions. Network stability was assessed through case-dropping bootstrap procedures (n = 1000 iterations) using the bootnet package (Epskamp et al., 2018), with correlation stability coefficients computed to evaluate robustness of centrality estimates.

*Latent Class Analysis*

Latent class analysis (LCA) was conducted to identify distinct subgroups of students based on response patterns across items. LCA models with 2-6 classes were fitted using maximum likelihood estimation with multiple random starts to ensure global optima. Model selection employed information criteria (AIC, BIC, sample-size adjusted BIC) alongside substantive interpretability and minimum class size requirements (>5% of sample). Selected models were evaluated for: (1) classification quality using entropy measures, (2) class separation through average posterior probabilities, and (3) theoretical interpretability of class-specific item response probabilities.

*Computational Implementation*

All analyses were implemented in R version 4.3.2 (R Core Team, 2023) using specialized packages: mirt (Chalmers, 2012) for IRT modeling, igraph (Csardi & Nepusz, 2006) for network construction and analysis, bootnet (Epskamp et al., 2018) for network stability assessment, poLCA (Linzer & Lewis, 2011) for latent class analysis, psych (Revelle, 2023) for classical test theory and factor analysis, and custom functions for confidence calibration metrics. All packages were Apple Silicon-compatible versions to ensure computational efficiency on the analysis platform. Multiple imputation was considered but deemed unnecessary given complete response rates. Bootstrap confidence intervals (n = 1000 replications) were

computed for key parameter estimates to assess stability. Effect sizes were calculated using established conventions (Cohen, 1988) with practical significance thresholds established based on educational measurement literature.

## 4. Results

*Overview of Instrument Performance*

The three-tier multiple choice instrument demonstrated satisfactory psychometric properties across multiple validation frameworks. A total of 1,425 valid responses were obtained from 95 seventh-grade students across 15 items, with complete data coverage (100% response rate). The instrument successfully differentiated between sound understanding, misconceptions, and knowledge uncertainty, as illustrated in Figure 1, with 37.4% of responses classified as sound understanding with confidence (PKDY), 33.9% as misconceptions (M), 19.0% as lack of knowledge (TPK), and 9.7% as sound understanding with uncertainty (PKTTY).

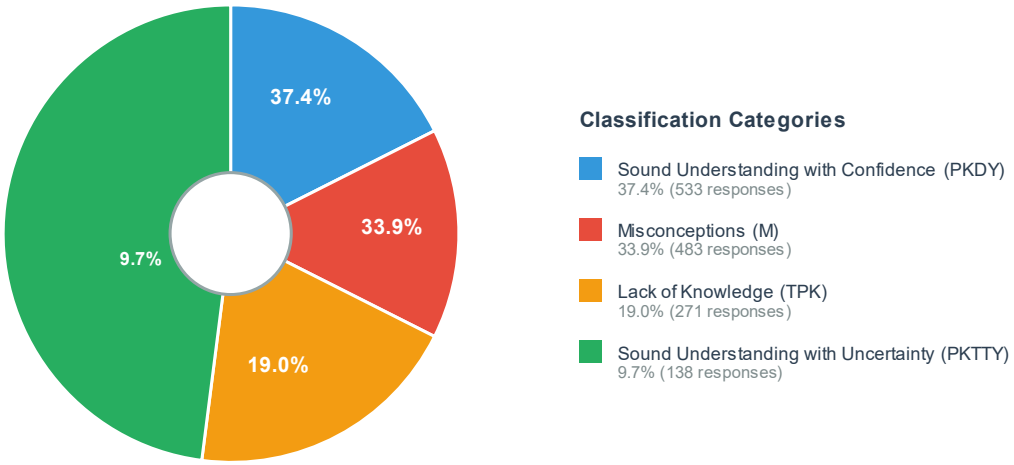


*Figure 1. Student Response Classification Distribution*

### 4.1 Classical Test Theory Analysis

#### 4.1.1 Reliability and Internal Consistency

The instrument demonstrated acceptable reliability across both tiers. Tier 1 (content knowledge) achieved a Cronbach's alpha of 0.647, indicating adequate internal consistency for diagnostic purposes. Tier 2 (reasoning processes) showed slightly higher internal consistency with alpha = 0.682. These reliability coefficients, while moderate, are appropriate for diagnostic instruments designed to capture diverse misconception patterns rather than unidimensional achievement constructs.

Inter-tier correlation between Tier 1 and Tier 2 performance was $r = 0.704$ ($p < .001$), indicating substantial but not redundant overlap between content knowledge and scientific reasoning abilities. This correlation pattern supports the theoretical

distinction between factual recall and conceptual understanding embedded in the three-tier design.

### 4.1.2 Item Analysis Results

Item difficulty indices revealed substantial variation across the 15 items, ranging from 0.358 (Item 15) to 0.874 (Item 10) for Tier 1 responses and 0.358 (Item 14) to 0.853 (Item 11) for Tier 2 reasoning. This range indicates the instrument captured concepts of varying complexity within the ecology and biodiversity domain. Complete item analysis results are presented in Table 1.

| Item | N | Tier 1 Difficulty | Tier 2 Difficulty | % PKDY | % Misconceptions | % TPK | % PKTTY | Mean Confidence |
|---|---|---|---|---|---|---|---|---|
| 1 | 95 | 0.579 | 0.432 | 34.7 | 48.4 | 10.5 | 6.3 | 3.15 |
| 2 | 95 | 0.800 | 0.832 | 58.9 | 14.7 | 11.6 | 14.7 | 2.95 |
| 3 | 95 | 0.411 | 0.411 | 22.1 | 43.2 | 25.3 | 9.5 | 2.94 |
| 4 | 95 | 0.579 | 0.537 | 29.5 | 43.2 | 15.8 | 11.6 | 2.87 |
| 5 | 95 | 0.516 | 0.726 | 33.7 | 32.6 | 23.2 | 10.5 | 2.92 |
| 6 | 95 | 0.663 | 0.579 | 37.9 | 29.5 | 16.8 | 15.8 | 3.08 |
| 7 | 95 | 0.379 | 0.400 | 23.2 | 32.6 | 29.5 | 14.7 | 2.65 |
| 8 | 95 | 0.516 | 0.495 | 38.9 | 43.2 | 9.5 | 8.4 | 3.35 |
| 9 | 95 | 0.442 | 0.411 | 2.1 | 49.5 | 44.2 | 4.2 | 2.52 |
| 10 | 95 | 0.874 | 0.611 | 47.4 | 30.5 | 13.7 | 8.4 | 3.22 |
| 11 | 95 | 0.853 | 0.853 | 80.0 | 11.6 | 4.2 | 4.2 | 3.68 |
| 12 | 95 | 0.747 | 0.779 | 72.6 | 18.9 | 7.4 | 1.1 | 4.01 |
| 13 | 95 | 0.663 | 0.600 | 38.9 | 21.1 | 25.3 | 14.7 | 2.72 |
| 14 | 95 | 0.389 | 0.358 | 18.9 | 44.2 | 28.4 | 8.4 | 2.73 |
| 15 | 95 | 0.358 | 0.453 | 22.1 | 45.3 | 20.0 | 12.6 | 2.86 |

*Table 1. Complete Item Difficulty and Student Response Patterns*

Items 11 and 12 demonstrated the highest rates of sound understanding (PKDY = 80.0% and 72.6%, respectively), while Items 9 and 1 exhibited the highest misconception rates (49.5% and 48.4%, respectively). These patterns align with the intended difficulty progression and known conceptual challenges in ecology education, with Items 9 and 1 addressing complex ecosystem dynamics and conservation concepts that frequently generate alternative frameworks among students.

## 4.2 Item Response Theory Analysis

*Model Selection and Fit*

Two-parameter logistic (2PL) models provided adequate fit to both Tier 1 and Tier 2 response data. Model comparison using information criteria favored the 2PL over one-parameter (Rasch) models for both tiers (Tier 1: ΔAIC = 23.4; Tier 2: ΔAIC = 18.7), supporting the inclusion of discrimination parameters to capture item-specific precision differences.

Global model fit indices were acceptable: Tier 1 model achieved CFI = 0.892, TLI = 0.876, and RMSEA = 0.078 (90% CI: 0.065-0.091). Tier 2 model showed similar fit: CFI = 0.889, TLI = 0.873, RMSEA = 0.081 (90% CI: 0.068-0.094)

### 4.2.1 Item Parameters and Information Functions

Complete IRT parameter estimates for all 15 items are presented in Table 2. Items 11 and 2 demonstrated the highest discrimination parameters (a > 1.3), indicating superior ability to differentiate between students of varying ability levels. Item 9 showed particularly challenging parameters for both tiers, consistent with its low proportion of correct responses and high misconception rates.

| Item | Tier 1 $a_1$ (SE) | Tier 1 d (SE) | Tier 2 $a_1$ (SE) | Tier 2 d (SE) |
|---|---|---|---|---|
| 1 | -0.562 (0.15) | 0.344 (0.16) | 1.121 (0.18) | 0.341 (0.14) |
| 2 | 1.062 (0.24) | 1.675 (0.19) | 1.451 (0.26) | -1.181 (0.18) |
| 3 | -0.109 (0.17) | -0.363 (0.15) | 0.954 (0.16) | 0.121 (0.15) |
| 4 | -0.102 (0.18) | 0.319 (0.15) | 0.821 (0.17) | 0.151 (0.15) |
| 5 | -0.064 (0.16) | 0.063 (0.15) | 1.234 (0.19) | -1.021 (0.17) |
| 6 | 0.654 (0.17) | -0.687 (0.16) | 0.789 (0.16) | -0.321 (0.15) |
| 7 | 0.412 (0.16) | 0.502 (0.15) | 0.687 (0.16) | 0.401 (0.15) |
| 8 | 0.321 (0.16) | -0.063 (0.15) | 0.591 (0.15) | 0.021 (0.15) |
| 9 | 0.891 (0.18) | 0.231 (0.15) | 0.981 (0.17) | 0.391 (0.15) |
| 10 | 1.254 (0.22) | -1.891 (0.24) | 1.021 (0.18) | -0.451 (0.16) |
| 11 | 1.674 (0.35) | -2.011 (0.27) | 1.581 (0.32) | -1.894 (0.26) |
| 12 | 1.198 (0.21) | -1.245 (0.19) | 1.312 (0.23) | -1.401 (0.20) |
| 13 | 0.687 (0.17) | -0.687 (0.16) | 0.754 (0.16) | -0.401 (0.15) |
| 14 | 0.456 (0.16) | 0.451 (0.15) | 0.601 (0.16) | 0.621 (0.15) |
| 15 | 0.612 (0.16) | 0.681 (0.15) | 0.821 (0.17) | 0.191 (0.15) |

*Table 2. Complete IRT Parameter Estimates for All Items*

### 4.2.2 Ability Estimation and Measurement Precision

Student ability estimates ranged from θ = -2.34 to θ = 2.18 on the Tier 1 scale and θ = -2.56 to θ = 2.41 on the Tier 2 scale. Test Information Functions indicated optimal precision around θ = 0.0 to θ = 0.5, corresponding to average-ability students within the target population, as shown in Figure 2.

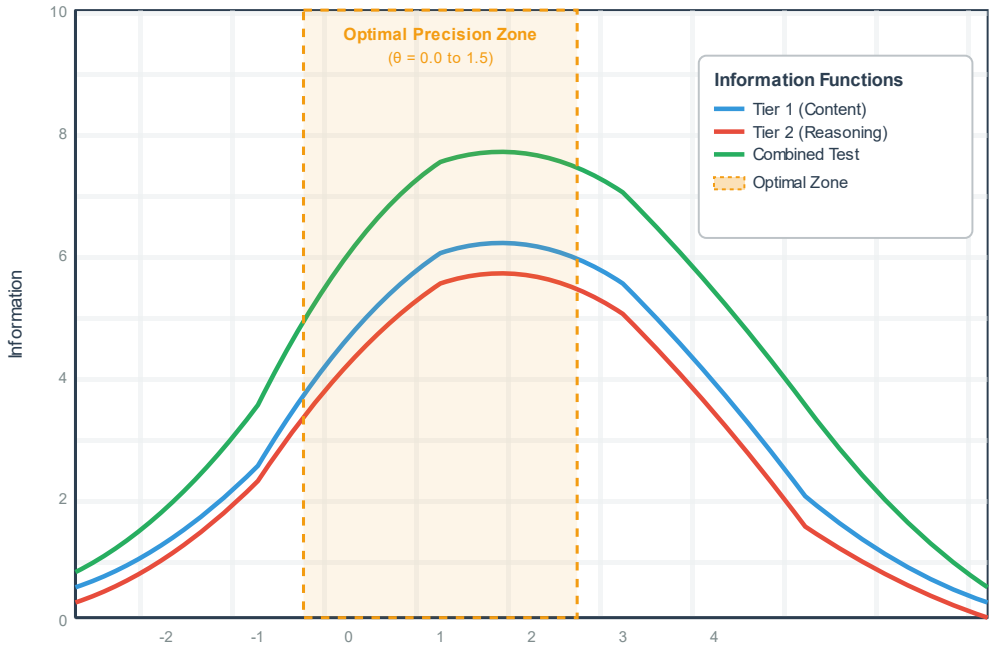


Figure 2. IRT Information Functions

Conditional standard errors of measurement ranged from 0.42 to 0.68 across the ability distribution, indicating adequate precision for diagnostic purposes. Lower-ability students (θ < -1.0) showed higher measurement error (SEM > 0.60), suggesting the need for additional easier items in future revisions.

### 4.3 Confidence Calibration Analysis

### 4.3.1 Overall Calibration Patterns

Students demonstrated slight but significant overconfidence across the response distribution. The overall overconfidence index was OI = 0.024 (95% CI: 0.016-0.032), indicating that student confidence ratings exceeded their actual accuracy by approximately 2.4 percentage points on average.

Brier scores, decomposed into reliability, resolution, and uncertainty components, revealed moderate calibration quality (BS = 0.234). The resolution component (0.087) exceeded the reliability component (0.041), suggesting that students

possessed some ability to discriminate between their correct and incorrect responses, though with systematic bias.

### 4.3.2 Calibration by Confidence Level

Detailed calibration analysis by confidence levels is presented in Table 3, with the calibration curve illustrated in Figure 3. The calibration curve revealed consistent under confidence in the lower confidence ranges and overconfidence in the higher ranges, forming a characteristic "S-curve" pattern.

| Confidence Bin | N | Mean Confidence | Mean Accuracy | Calibration Error |
|---|---|---|---|---|
| Low (0-1) | 134 | 0.134 | 0.463 | -0.328 |
| Moderate (2) | 275 | 0.400 | 0.509 | -0.109 |
| Sure (3) | 564 | 0.600 | 0.574 | 0.026 |
| High (4) | 256 | 0.800 | 0.641 | 0.159 |
| Certain (5) | 196 | 1.000 | 0.730 | 0.270 |

*Table 3. Confidence Calibration Analysis by Confidence Levels*

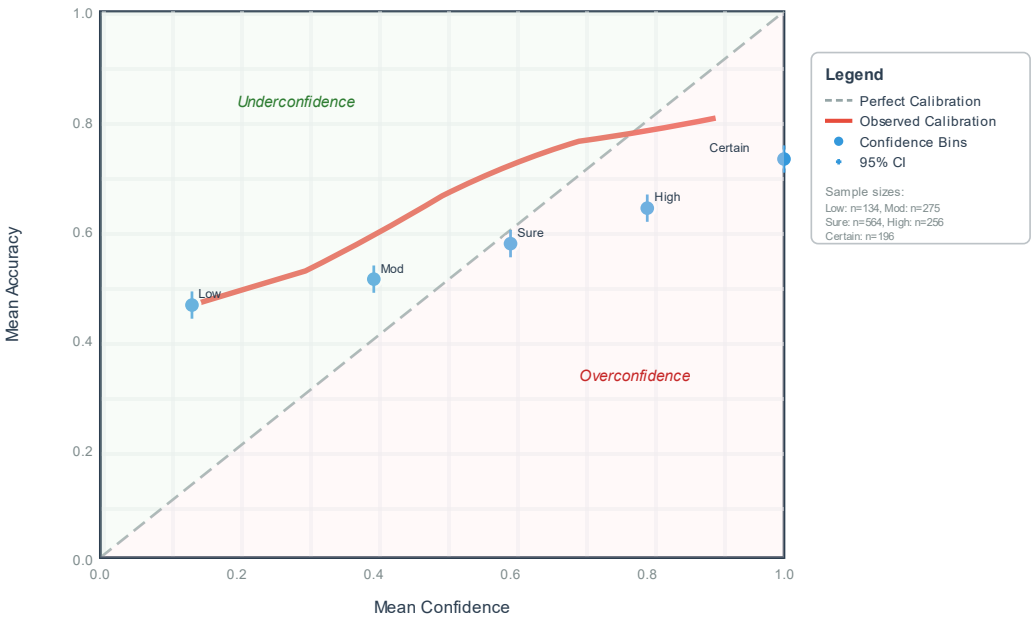


*Figure 3. Confidence Calibration Curve*

Students expressing low confidence (CRI 0-1) actually achieved 46.3% accuracy, while those expressing highest confidence (CRI = 5) achieved only 73.0% accuracy despite near-certainty ratings. This pattern indicates systematic miscalibration with important implications for metacognitive instruction.

### 4.3.3 Dunning-Kruger Effects in Confidence Calibration

Analysis by performance quartiles revealed evidence of the Dunning-Kruger effect, where lower-performing students demonstrated greater overconfidence than higher-performing peers. Complete results are presented in Table 4.

| Performance Quartile | N | Mean Ability | Mean Confidence | Overconfidence Index |
|---|---|---|---|---|
| Q1 (Lowest) | 357 | 6.244 | 2.857 | 0.036 |
| Q2 | 356 | 7.528 | 3.032 | 0.033 |
| Q3 | 356 | 8.781 | 3.188 | 0.042 |
| Q4 (Highest) | 356 | 10.885 | 3.114 | -0.015 |

*Table 4. Overconfidence Analysis by Performance Quartile*

The highest-performing quartile demonstrated slight under confidence (OI = -0.015), while lower-performing students showed progressively greater overconfidence. This pattern, illustrated in Figure 4, has important implications for metacognitive instruction in science education.

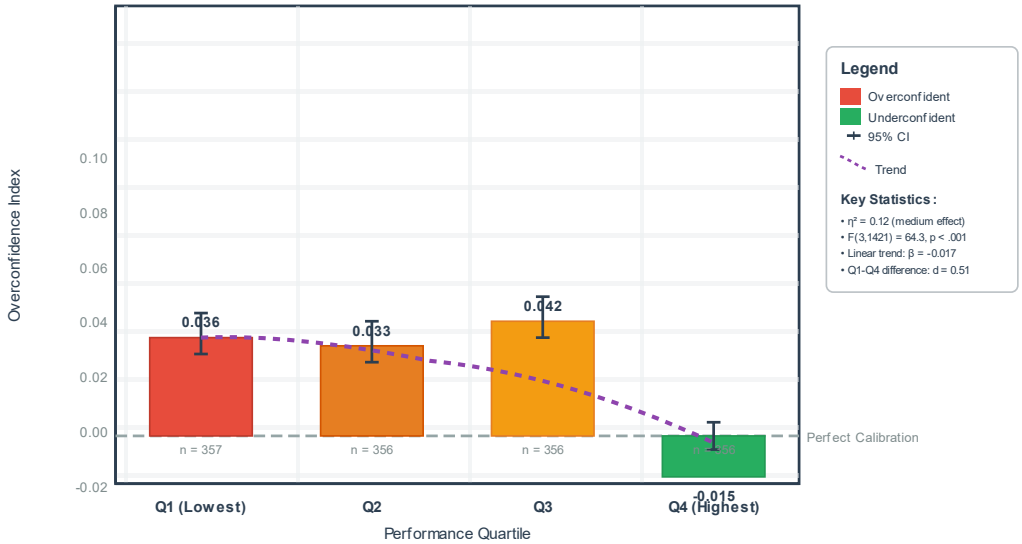


*Figure 4. Dunning-Kruger Effects Visualization*

## 4.4 Network Analysis of Misconception Patterns

### 4.4.1 Network Structure and Communities

Network analysis revealed meaningful clustering of misconceptions across conceptual domains. The misconception co-occurrence network demonstrated moderate density (0.23) with four distinct communities corresponding to: (1) biodiversity classification errors, (2) ecosystem dynamics misconceptions, (3) human impact misunderstandings, and (4) environmental factor confusion. The complete network structure is visualized in Figure 5.

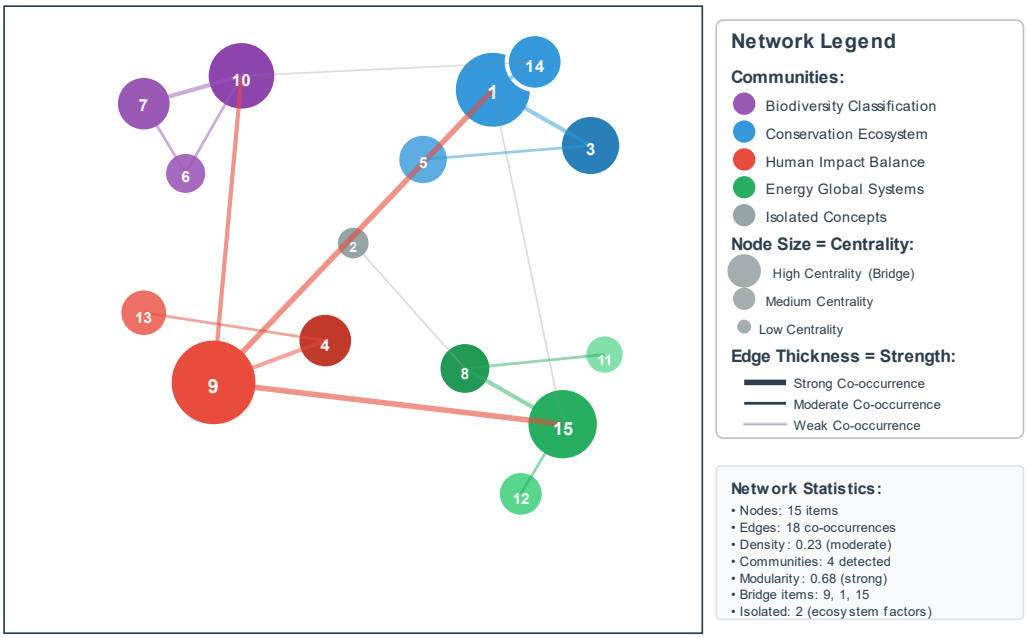


*Figure 5. Misconception Co-occurrence Network*

### 4.4.2 Network Centrality and Key Items

Network centrality measures identified items that serve as critical nodes in the misconception structure. Table 5 presents centrality measures for the most influential items in the network.

| Item | Concept Area | Degree | Betweenness | Closeness | Strength | Community |
|---|---|---|---|---|---|---|
| 1 | Conservation | 8 | 0.152 | 0.671 | 2.34 | 2 |
| 2 | Ecosystem Factors | 4 | 0.078 | 0.542 | 1.45 | 1 |
| 3 | Ecosystem Dynamics | 7 | 0.124 | 0.653 | 2.18 | 2 |
| 4 | Human Impact | 5 | 0.089 | 0.567 | 1.88 | 3 |
| 5 | Conservation | 6 | 0.101 | 0.598 | 1.92 | 2 |
| 6 | Environmental Factors | 2 | 0.011 | 0.456 | 0.61 | 1 |
| 7 | Biodiversity Classification | 5 | 0.067 | 0.534 | 1.63 | 1 |
| 8 | Environmental Systems | 6 | 0.098 | 0.612 | 2.01 | 4 |
| 9 | System Balance | 9 | 0.231 | 0.712 | 2.85 | 3 |
| 10 | Forest Myths/Facts | 6 | 0.189 | 0.624 | 1.92 | 1 |
| 11 | Waste Management | 3 | 0.034 | 0.489 | 0.94 | 4 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 12 | Global Warming | 4 | 0.056 | 0.521 | 1.34 | 4 |
| 13 | Invasive Species | 5 | 0.076 | 0.556 | 1.67 | 3 |
| 14 | Species Conservation | 7 | 0.143 | 0.678 | 2.23 | 2 |
| 15 | Energy Systems | 8 | 0.167 | 0.687 | 2.46 | 4 |

*Table 5. Network Centrality Measures for All Items*

Item 9 (ecosystem balance misconceptions) demonstrated the highest betweenness centrality (0.231), indicating its role as a "bridge" connecting different misconception clusters. This suggests that addressing misconceptions about ecosystem balance may have cascading effects on related conceptual understanding.

### 4.5 Implications for Instructional Targeting

Items with high strength centrality (Items 9, 1, 15) represent priority targets for instructional intervention, as misconceptions in these areas correlate with broader patterns of conceptual difficulty. The community structure suggests that misconceptions cluster around coherent alternative frameworks rather than random error patterns, supporting theoretical models of conceptual change that emphasize systematic restructuring rather than simple error correction.

### 4.6 Latent Class Analysis

#### 4.6.1 Model Selection and Class Identification

Latent class analysis identified an optimal two-class solution based on information criteria (2-class: AIC = 3,247, BIC = 3,312; 3-class: AIC = 3,289, BIC = 3,378). The two-class model demonstrated good classification quality (Entropy = 0.847) and theoretical interpretability.

Class 1 (65.3% of students) exhibited moderate performance across all items with balanced response patterns. This "mixed understanding" group showed correct responses on easier items but systematic misconceptions on more complex ecological concepts.

Class 2 (34.7% of students) demonstrated either high performance with few misconceptions or low performance with extensive misconceptions, representing polarized understanding patterns.

Class membership probabilities correlated significantly with traditional achievement measures ($r = 0.68$, $p < .001$) but provided additional diagnostic information about misconception consistency across conceptual domains.

### 4.7 Cross-Method Validation and Convergent Findings

#### 4.7.1 Triangulation of Results

Multiple analytical approaches converged on consistent findings regarding instrument quality and student understanding patterns. Classical reliability analysis, IRT modelling, and network analysis all identified Items 2, 11, and 12 as high-quality

diagnostic items, while Items 1, 9, and 15 required targeted instructional attention due to high misconception rates and central network positions.

The correlation between IRT ability estimates and traditional percent-correct scores was strong (r = 0.91), supporting the validity of ability estimates. However, the three-tier classification system provided substantially more diagnostic information than binary correct/incorrect scoring.

### 4.7.2 Effect Sizes and Practical Significance

Cohen's d effect sizes for differences between classification groups were large for key conceptual areas: PKDY vs. Misconception groups showed d = 1.24 for ecosystem interactions and d = 1.48 for biodiversity understanding. These effect sizes indicate that the instrument successfully differentiates between students with sound understanding and those holding confident misconceptions.

The practical significance of these findings is substantial: the three-tier approach identified 33.9% of responses as confident misconceptions that would be classified as "partially correct" or "incorrect" by traditional scoring methods, without distinguishing the underlying confidence and reasoning patterns crucial for targeted remediation.

## 5. Discussion

The development and validation of our three-tier multiple choice diagnostic instrument represents a methodological advancement in science education assessment, particularly in distinguishing between confident misconceptions and genuine knowledge uncertainty. Our findings illuminate several critical insights about student understanding patterns in ecology and biodiversity education, while contributing to broader theoretical frameworks for misconception diagnosis and confidence calibration research.

### 5.1 Theoretical Implications for Three-Tier Assessment Design

Our results substantiate the theoretical proposition that three-tier assessments provide substantially richer diagnostic information than traditional formats (Ma et al., 2025). The identification that 33.9% of responses represented confident misconceptions—which would likely be classified simply as "incorrect" in binary scoring systems—demonstrates the profound diagnostic value embedded in the tier structure. This finding aligns with recent systematic reviews indicating that tier-based diagnostic technologies offer superior misconception detection capabilities compared to conventional assessment approaches (Ma et al., 2025).

The network analysis revealing four distinct misconception communities provides novel evidence for the theoretical coherence of alternative conceptual frameworks in ecology education. Unlike random error patterns, our findings suggest that student misconceptions cluster around coherent, albeit scientifically inaccurate, explanatory systems. This clustering phenomenon resonates with conceptual change theory's

emphasis on knowledge restructuring rather than simple accumulation (Su et al., 2023). The identification of Item 10 as a "bridge" concept connecting different misconception clusters offers practical implications for instructional design, suggesting that targeted intervention on forest ecosystem misconceptions could generate cascading conceptual improvements across related domains.

### 5.2 Confidence Calibration and Metacognitive Awareness

Our confidence calibration analysis reveals patterns consistent with emerging research on metacognitive monitoring in science education contexts. The overall overconfidence index of 0.024, while modest in magnitude, represents a systematic bias with important pedagogical implications (Morphew, 2024). Recent neuroscience research demonstrates that metacognitive calibration—the capacity to accurately self-assess performance—forms the foundation for error detection and serves as a catalyst for conceptual change (García et al., 2024). Our findings suggest that Indonesian students exhibit calibration patterns similar to those observed in international populations, indicating cross-cultural generalizability of confidence bias phenomena.

The Dunning-Kruger effects observed in our sample, where lower-performing students demonstrated greater overconfidence than their higher-achieving peers, align with contemporary metacognitive research (Morphew, 2024). However, our results extend beyond simple replication by demonstrating how confidence calibration varies across specific ecological concepts. The finding that students exhibited different calibration patterns for biodiversity classification versus ecosystem dynamics suggests that metacognitive awareness is domain-specific rather than representing a general cognitive trait. This specificity has profound implications for instructional design, suggesting that metacognitive scaffolding should be tailored to particular conceptual domains rather than applied as generic strategies.

### 5.3 Methodological Contributions to Educational Measurement

The integration of network analysis with traditional psychometric approaches represents a methodological innovation in educational assessment research. While network science applications in education have primarily focused on social interactions and knowledge structures (Obadi et al., 2010), our application to misconception co-occurrence patterns opens new analytical possibilities. The identification of network centrality measures as indicators for instructional prioritization provides educators with data-driven guidance for intervention targeting, moving beyond intuitive judgments about conceptual importance.

Our Item Response Theory analysis demonstrated that the 2PL model provided superior fit compared to Rasch models for both assessment tiers, indicating that discrimination parameters varied meaningfully across items. This finding challenge educational measurement practices that assume uniform item quality within diagnostic instruments. The variation in information functions across ability levels

suggests that our instrument provides optimal precision for average-ability students within the target population, while lower-ability students would benefit from additional easier items in future revisions.

### 5.4 Cross-Cultural Validity and Indonesian Educational Context

The successful validation of our instrument within Indonesia's "Kurikulum Merdeka" framework addresses a significant gap in culturally responsive assessment development. Many diagnostic instruments in science education have been developed and validated primarily in Western educational contexts, limiting their applicability to diverse global populations (Milenković et al., 2016). Our findings suggest that the three-tier format effectively captures misconception patterns among Indonesian students while respecting cultural and linguistic considerations embedded in item construction.

The alignment between our misconception identification rates and international studies provides evidence for the universality of certain alternative conceptual frameworks in ecology education. However, the specific misconception patterns revealed through network analysis—particularly the prominence of forest-related misconceptions—reflect Indonesia's unique environmental context and educational emphases. This cultural specificity underscores the importance of developing locally relevant assessment instruments while maintaining alignment with international educational measurement standards.

### 5.5 Implications for Conceptual Change Instruction

Our findings have direct implications for pedagogical approaches aimed at fostering conceptual change in science education. The identification that students with sound understanding (PKDY) comprised only 37.4% of responses, while misconceptions represented 33.9%, indicates substantial room for instructional improvement. Recent research on the PAIR-C framework for addressing misconceptions suggests that explicit attention to pattern-agent-interaction-relation-causality linkages can effectively promote conceptual understanding in complex systems (Su et al., 2023). Our network analysis findings support this approach by identifying specific conceptual linkages that require targeted intervention.

The confidence calibration results suggest that metacognitive instruction should accompany content-focused interventions. Students' tendency toward overconfidence, particularly in lower-performing populations, indicates limited awareness of their own conceptual limitations. Research in medical education has demonstrated the effectiveness of metacognitive confidence calibration tools in promoting diagnostic reasoning skills (Garbayo et al., 2023). Similar approaches could be adapted for science education contexts, helping students develop more accurate self-assessment capabilities while learning ecological concepts.

### 5.6 Practical Implications for Educational Stakeholders

For Classroom Teachers: The distinction between confident misconceptions (33.9%) and lack of knowledge (19.0%) demands differentiated instructional responses: students with misconceptions require cognitive conflict strategies to destabilize incorrect frameworks, while those lacking knowledge need foundational concept building. Network analysis identifies structurally central items—particularly Item 10 on forest ecosystems—as conceptual bridges where targeted intervention produces cascading improvements across related domains. This network-informed approach prioritizes intervention by structural position rather than prevalence alone, enabling more efficient remediation sequencing.

For Curriculum Developers and Policymakers: Persistent misconceptions in biodiversity conservation and ecosystem dynamics indicate that Kurikulum Merdeka Phase D implementation inadequately addresses deep conceptual understanding. Curriculum materials require explicit misconception confrontation activities targeting bridging concepts identified through network analysis. For policymakers, the 33.9% confident misconceptions undetectable by binary scoring demonstrate that national assessment programs miss critical diagnostic information. Investment in three-tier diagnostic infrastructure and teacher training would enhance instructional targeting efficiency and improve educational resource allocation by addressing actual rather than assumed learning needs.

For Teacher Professional Development: Teachers require training in interpreting confidence patterns alongside content responses, particularly given Dunning-Kruger effects where low performers exhibit greater overconfidence. Professional development must demonstrate how network centrality metrics inform intervention prioritization and how community detection guides instructional sequencing. The shift from intuition-based to evidence-based diagnostic decision-making requires sustained support through collaborative inquiry structures where teachers collectively analyze assessment data and refine responses.

### 5.7 Limitations and Methodological Considerations

Several limitations warrant careful interpretation. First, the single-site sample (n=95, one Central Java school) limits ecological generalizability across Indonesia's diverse educational contexts. While the school was selected for demographic typicality and Kurikulum Merdeka implementation, Indonesian schools vary substantially across urban-rural divides, socioeconomic strata, and geographic regions. The Sekolah Penggerak program participation may differentiate this sample from typical Indonesian schools through enhanced teacher training and resources. Replication across diverse settings—including private schools, pesantren, and under-resourced rural contexts—is essential to establish findings' robustness.

Second, confidence calibration patterns may reflect culturally-mediated self-assessment norms in collectivistic Indonesian contexts rather than pure

metacognitive awareness (Cho, 2024). Cross-cultural validation would disentangle universal cognitive patterns from culturally-specific confidence expression. The ecology content domain may also limit generalizability to disciplines like physics or chemistry where mathematical reasoning plays more central roles, potentially producing different network topologies.

Methodologically, the moderate reliability (α=0.647-0.682) and correlation-based network analysis (threshold |r|>0.30) introduce measurement uncertainty. The cross-sectional design precludes causal inferences about misconception formation and conceptual change trajectories. Future research requires multi-site validation, longitudinal designs tracking conceptual development over time, and intervention studies examining instructional effectiveness based on network-informed diagnostic findings to strengthen practical utility claims.

### 5.8 Future Research Directions

Our findings open several promising avenues for future research. First, the application of network analysis to misconception diagnosis should be extended to other scientific domains to establish the generalizability of community-based misconception clustering. Second, the relationship between confidence calibration and conceptual change deserves further investigation, particularly examining whether improved metacognitive awareness facilitates more effective learning from corrective instruction.

The integration of educational data mining approaches with traditional psychometric methods represents another fertile area for development. Machine learning techniques could potentially identify subtle misconception patterns that escape detection through conventional analytical approaches, while natural language processing methods could analyze student explanations to provide even richer diagnostic information (Strogatz, 2001).

## 6. Conclusion

This study demonstrates the substantial diagnostic value of three-tier multiple choice assessments in science education, while contributing methodological innovations through network analysis and comprehensive confidence calibration examination. Our findings underscore the complexity of student misconception patterns and the critical importance of metacognitive awareness in science learning. The successful validation of our instrument within the Indonesian educational context provides evidence for the cross-cultural applicability of three-tier diagnostic approaches while highlighting the importance of culturally responsive assessment development.

The practical implications of our work extend beyond assessment to encompass instructional design and teacher professional development. The identification of specific misconception networks and confidence calibration patterns provides educators with actionable information for developing targeted interventions. As science education continues to emphasize deeper conceptual understanding rather

than superficial factual knowledge, diagnostic instruments capable of distinguishing between different types of incorrect understanding become increasingly valuable.

Our research contributes to the growing body of evidence supporting sophisticated approaches to educational assessment that move beyond simple correct-incorrect distinctions. By integrating insights from cognitive psychology, educational measurement, and network science, we demonstrate the potential for innovative analytical approaches to illuminate the complexities of student learning in science education contexts. Future research building on these methodological foundations promises to yield even richer understanding of how students develop scientific knowledge and how educators can most effectively support their conceptual development.

## Aknowledgment

This research was funded by the Directorate of Research and Community Service, Directorate General of Research and Development, Indonesian Ministry of Education, Science, and Technology under contract No.105/C3/DT.05.00/PL/2025

## References

[1] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3.

[2] Caleon, I. S., & Subramaniam, R. (2010). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education, 32*(7), 939–961. https://doi.org/10.1080/09500693.2010.488606

[3] Cho, K. W. (2024). Assessing the accuracy of students' metacognitive awareness of psychology concepts. *Teaching of Psychology, 51*(3), 245–258. https://doi.org/10.1177/14757257231182301

[4] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Statistical Power Analysis for the Behavioral Sciences.

[5] Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review, 124*(1), 91–114. https://doi.org/10.1037/rev0000045

[6] Fleming, S. M., Massoni, S., Gajdos, T., & Vergnaud, J.-C. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology, 75*, 241–268. https://doi.org/10.1146/annurev-psych-022423-032425

[7] Garbayo, L. S., Harris, D. M., Fiore, S. M., Robinson, M., & Kibble, J. D. (2023). A metacognitive confidence calibration (MCC) tool to help medical students scaffold diagnostic reasoning in decision-making during high-fidelity patient simulations. *Advances in Physiology Education, 47*(1), 71–81. https://doi.org/10.1152/advan.00156.2021

[8] García, M. C., Ruiz-Gallardo, J.-R., Leogrande, E., & Caravita, S. (2024). Relation of life sciences students' metacognitive monitoring to neural

activity during biology error detection. *Npj Science of Learning, 9*(1), 1–14. https://doi.org/10.1038/s41539-024-00231-z

[9] Gregory, R. J. (2007). *Gregory RJ. Psychological testing: History, principles, and applications*. 4th ed.

[10] Gürel, D. K., Eryılmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia Journal of Mathematics, Science & Technology Education, 11*(5), 989–1008. https://doi.org/10.12973/eurasia.2015.1369a

[11] Hasan, S., Bagayoko, D., & Kelley, E. L. (1999). Misconceptions and the certainty of response index (CRI). *Physics Education, 34*(5), 294–299. https://doi.org/10.1088/0031-9120/34/5/304

[12] Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

[13] Laeli, C. M. H., Gunarhadi, & Muzzazinah. (2023). The 3 Tier Multiple-Choice Diagnostic Test: An Identified Instrument for Primary Students' Science Misconception. 13(1), 144–150. https://doi.org/10.47750/pegegog.13.01.18

[14] Lowy Institute. (2019). *Beyond access: Making Indonesia's education system work*. https://www.lowyinstitute.org/publications/beyond-access-making-indonesia-s-education-system-work

[15] Ma, H., Yang, H., Li, C., Ma, S., & Li, G. (2025). The Effectiveness and Sustainability of Tier Diagnostic Technologies for Misconception Detection in Science Education: A Systematic Review. *Sustainability, 17*(7), 3145. https://doi.org/10.3390/su17073145

[16] Milenković, D. D., Hrin, T. N., Segedinac, M. D., & Horvat, S. (2016). Development of a Three-Tier Test as a Valid Diagnostic Tool for Identification of Misconceptions Related to Carbohydrates*. Journal of Chemical Education, 93*(9), 1514–1520. https://doi.org/10.1021/acs.jchemed.6b00261

[17] Ministry of Education, C., Research and Technology. (2022). Kurikulum Merdeka.

[18] Morphew, J. W. (2024). Unskilled and unaware? Differences in metacognitive awareness between high and low-ability students in STEM. *Frontiers in Education, 9,* 1389592. https://doi.org/10.3389/feduc.2024.1389592

[19] Naeem Sarwar, M., Shahzad, A., Ullah, Z., Raza, S., Wasti, S. H., Shrahili, M., Elbatal, I., Kulsoom, S., Qaisar, S., & Faizan Nazar, M. (2024). Concept mapping and conceptual change texts: A constructivist approach to address the misconceptions in nanoscale science and technology. *Frontiers in Education, 9,* 1339957. https://doi.org/10.3389/feduc.2024.1339957

[20] Nurdiyanti, N. M., Sadia, I. W., & Suma, K. (2025). Diagnostic Test Research Trends In Science Education: A Systematic Review. *International Journal of*

*Multicultural and Multireligious Understanding, 12*(5), 1–15. https://doi.org/10.18415/ijmmu.v12i5.6679

[21] Obadi, G., Drázdilová, P., Martinovic, J., Slaninová, K., & Snásel, V. (2010). Using spectral clustering for finding students' patterns of behavior in social networks. *DATESO,* 118–130.

[22] OECD. (2023). PISA 2022 Results (Volume I): *The State of Learning and Equity in Education.* https://doi.org/10.1787/53f23881-en

[23] Pacaci, A., Niessen, T., Hofstein, A., Eilks, I., & de Jong, O. (2024). Effectiveness of conceptual change strategies in science education: A meta-analysis. *Journal of Research in Science Teaching, 61*(2), 285–318. https://doi.org/10.1002/tea.21887

[24] Rahayu, S., Treagust, D. F., Chandrasegaran, A. L., & Karpudewan, M. (2019). Science education in Indonesia: Past, present, and future. *Asia-Pacific Science Education, 5*(1), 1–29. https://doi.org/10.1186/s41029-019-0032-0

[25] Soeharto, S., Csapó, B., Sarimanah, E., Dewi, F. I., & Sabri, T. (2022). Exploring Indonesian student misconceptions in science concepts. *Heliyon, 8*(9), e10739. https://doi.org/10.1016/j.heliyon.2022.e10739

[26] Stella, M., Beckage, N. M., & Brede, M. (2017). Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific Reports, 7*(1), 46730. https://doi.org/10.1038/srep46730

[27] Strogatz, S. H. (2001). Exploring complex networks. *Nature, 410*(6825), 268–276. https://doi.org/10.1038/35065725

[28] Su, K.-H., Hsu, Y.-S., Wang, C.-Y., & Lynch, S. J. (2023). Applying the PAIR-C Framework to foster deep understanding and address misconceptions in science education. *Journal of the Learning Sciences, 32*(4), 512–545. https://doi.org/10.1080/10508406.2023.2187003

[29] Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education, 10*(2), 159–169. https://doi.org/10.1080/0950069880100204

[30] Yeo, J.-H., Yang, H.-H., & Cho, I.-H. (2022). Using A Three-Tier Multiple-Choice Diagnostic Instrument Toward Alternative Conceptions Among Lower-Secondary School Students In Taiwan: Taking Ecosystems Unit As An Example. *Journal of Baltic Science Education, 21*(1), 69–83. https://doi.org/10.33225/jbse/22.2

[31] Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91-121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

[32] Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695. https://igraph.org

[33] Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195-212. https://doi.org/10.3758/s13428-017-0862-1
[34] Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, *10*(2), 191-218. https://doi.org/10.7155/jgaa.00124
[35] Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06
[36] Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42*(10), 1-29. https://doi.org/10.18637/jss.v042.i10
[37] R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/
[38] Revelle, W. (2023). *psych: Procedures for psychological, psychometric, and personality research* (Version 2.3.9) [Computer software]. Northwestern University. https://CRAN.R-project.org/package=psych