Available online at: http://journal.uny.ac.id/index.php/jipi

Jurnal Inovasi Pendidikan IPA, 12 (1), 2026, 554-562

Analysis of Computerized Adaptive Test to Reveal Misconceptions and Science Literacy in Science Courses

Muhammad Azzarkasyi[1*], Ani Rusilowati[2], Endang Susilaningsih[2], Ibrahim[3], Mohd Isha Awang[4]

[1]Physics Education, Faculty of Teacher Training and Education, Universitas Serambi Mekkah, Banda Aceh, Indonesia.
[2]Department of Science Education, Universitas Negeri Semarang, Semarang, 50229, Indonesia
[3]Biology Education, Faculty of Teacher Training and Education, Universitas Serambi Mekkah, Banda Aceh, Indonesia.
[4]University Utara Malaysia Kedah, Malaysia.
* Corresponding Author. E-mail: azzarkasyi@gmail.com

Received: 5 January 2026; Revised: 1 April 2026; Accepted: 21 April 2026

Abstract: The development of diagnostic instruments that accurately capable in assessing interdisciplinary science literacy and identifying misconceptions represents a strategic imperative in higher education, particularly within integrated science courses that demand concept transfer across physics, chemistry, biology, and earth sciences. The study advanced both theoretical and methodological frontiers by developing and psychometrically validating a Computerized Adaptive Test (CAT) instrument, which is grounded in a Four-Tier Diagnostic Test framework and designed to profile students' science literacy and misconception patterns in integrated science contexts. The study employed a quantitative instrument development design with 150 students of science education from three universities in Aceh, Indonesia. The study moved beyond Classical Test Theory by utilizing Item Response Theory (IRT), specifically the Rasch model and two-parameter logistic (2PL) model to evaluate item parameters essential for CAT calibration. The findings demonstrated strong psychometric properties, with Rasch infit and outfit statistics within acceptable ranges confirming unidimensional, while item difficulty parameters spanned from 2.1 to 1.8 logits, providing adequate ability continuum coverage. Critically, domain-specific analysis revealed that items requiring cross-disciplinary concept transfer, particularly those integrating physics and biology, exhibited significantly higher discrimination parameters than items confined to isolated disciplinary content, offering a novel theoretical insight that science literacy is inherently an integrative construct rather than a sum of disciplinary knowledge components. Methodologically, this study advances CAT development by demonstrating that rigorous IRT-based calibration and iterative quality control are essential for ensuring measurement accuracy. The identification of invalid items underscores that item attrition is a necessary feature of responsible test development. Generally, these findings contribute to the broader movement toward adaptive, personalized assessment in higher education, providing a replicable model for researchers and practitioners seeking to leverage CAT technologies to enhance diagnostic precision and support targeted remediation in interdisciplinary science education worldwide.

Keywords: Science Literacy, Misconceptions, Computerized Adaptive Test (CAT), Validity, Reliability.

How to Cite: Azzarkasyi, M., Rusilowati, A., Susilaningsih, E., Ibrahim, Awang, M. I., (2026) Analysis of Computerized Adaptive Test Diagnostic Instruments to Reveal Misconceptions and Science Literacy in Integrated Science Courses. *Jurnal Inovasi Pendidikan IPA, 12*(1), 45-53. doi:https://doi.org/10.21831/jipi.v12i1.94641

INTRODUCTION

Developments in the 21st century require higher education graduates not only to master content knowledge but also to possess comprehensive science literacy skills. Science literacy, as defined by the (OECD, 2023), encompasses the ability to use scientific knowledge, identify questions, and draw evidence-based conclusions to understand and make decisions about nature and the changes of humans.

In the context of higher education, particularly in science courses such as Integrated Science, science literacy is a critical foundation for prospective educators or scientists to think logically, analytically, and solve complex problems. (Hartono et al., 2023; Lestari, 2021; Ni'mah, 2019).

However, the main challenge in achieving this goal is the persistent of misconceptions among students. Misconceptions are defined as conceptual understandings that are inconsistent with accepted scientific explanations (Maison et al., 2020; Rohmadhani et al., 2021; Rusilowati et al., 2021; Widarti et al., 2024). In interdisciplinary integrated science education, such as physics, chemistry, biology, and earth and space sciences, the risk of misconceptions is higher. The uniqueness of misconceptions in this context lies in the failure of concept transfer, namely, students understand a principle in one field (e.g., biology) but fail to apply the same laws of physics or chemistry in the context of field. According to (Taqwim et al., 2022), misconceptions often occur when students are unable to link abstract concepts within a topic to the underlying scientific principles, resulting in a fragmented understanding. This inability to connect concepts from various disciplines, which have different terminologies and logical frameworks, makes misconceptions more complex and difficult to detect. Undetected and unaddressed misconceptions become cognitive obstacles that interfere with the construction of new knowledge, hinder the development of science literacy, and finally, reduce the quality of graduates (Nurhidayatulah & Prodjosantoso, 2018; Suparno, 2013; Taqwim et al., 2022).

Furthermore, in the context of pre-service teacher education, misconceptions have broader implications, known as the Cycle of Misconception. If pre-service teachers hold misconceptions during their training. Here, there is a significant risk that they will pass on these erroneous understandings to their future students. This phenomenon creates a continuous cycle that is difficult to break, making efforts to diagnose and remediate misconceptions at the university level. It is not only individual academic improvements, but a long-term investment in the generative quality of science education.

Therefore, a diagnostic instrument is needed that is not only capable of identifying the presence or absence of misconceptions but also capable of measuring science literacy more holistically. Conventional diagnostic instruments developed based on Classical Test Theory (CTT), such as standard multiple-choice tests (single-tier tests), often fail to distinguish incorrect answers, which result to misconceptions, and those resulting from a lack of knowledge or carelessness (Isnaini et al., 2025; Nasyidiah et al., 2020). The CTT relies solely on total scores and is not yet capable of distinguishing between lucky guesses and prior knowledge, so the diagnostic depth is very limited.

To overcome this weakness, the Four-Tier Diagnostic Test (FTT) emerged as a more comprehensive alternative. This instrument consists of four levels: (1) choice of answers to content questions, (2) level of confidence in those answers, (3) choice of reasons for the answers, and (4) level of confidence in the reasons chosen (Gurel et al., 2015; Kaltakci-Gurel et al., 2017). The main advantage of FTT over traditional CTT approaches lies in its ability to distinguish between lucky guesses and a lack of knowledge through a dual-confidence scale. By comparing the consistency of confidence between answers and justifications, this instrument can accurately classify whether a student has a sound understanding, holds a misconception, or merely guessing. This structure enables researchers or educators to map students' understanding profiles more accurately, distinguishing among misconceptions, lack of knowledge, and understanding.

Despite its great potential, the development and implementation of FTT-based diagnostic instruments for integrated science courses in higher education remains limited. Most FTT research still focuses on specific science subjects (such as electrical physics or cellular biology) at the secondary school level (Isnaini et al., 2025; Önder Çelikkanlı & Kızılcık, 2022). Development for integrative contexts, such as Integrated Science, requires more complex considerations because it encompasses various conceptual domains that demand the transfer of concepts across disciplines. In addition, the main requirement for the widely used of instrument and provide a meaningful data is the fulfilment the principles of high validity and reliability. Validity refers to the extent or power of the instrument to measure the intended construct (misconceptions and science literacy), while reliability relates to the consistency of measurement results (Amelia et al., 2024; Ishtiaq Ahmed & Sundas Ishtiaq, 2021; Susongko et al., 2024). The development of assessment instruments based on a scientific approach and their validity is a crucial step in evaluating students' science literacy skills (Fahmi et al., 2022).

Based on the previous description and background, there is a gap between the urgency of accurate diagnostic needs for misconceptions and science literacy in integrated science courses and the availability of psychometrically tested instruments. The development and validation of FTT-based

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

diagnostic instruments for this context is a strategic necessity. Valid and reliable instruments will be powerful tools for lecturers to conduct assessments for learning, diagnose students' learning difficulties as early, and design targeted remedial learning. For researchers, these instruments might be used to evaluate the effectiveness of innovative learning models or media in reducing misconceptions and improving science literacy.

METHOD

The study used an instrument development of a research design with a quantitative approach that focuses on analysing the psychometric characteristics of test items. The stages of the study are illustrated in the flowchart in Figure 1. The process begins with the development of initial items in the Four-Tier Diagnostic Test format, followed by expert validation and empirical testing to obtain response data. After that, it is analysed for validity and reliability. The analysis results are used for compiling a valid and reliable question bank as the foundation for the Computerized Adaptive Test (CAT) system.

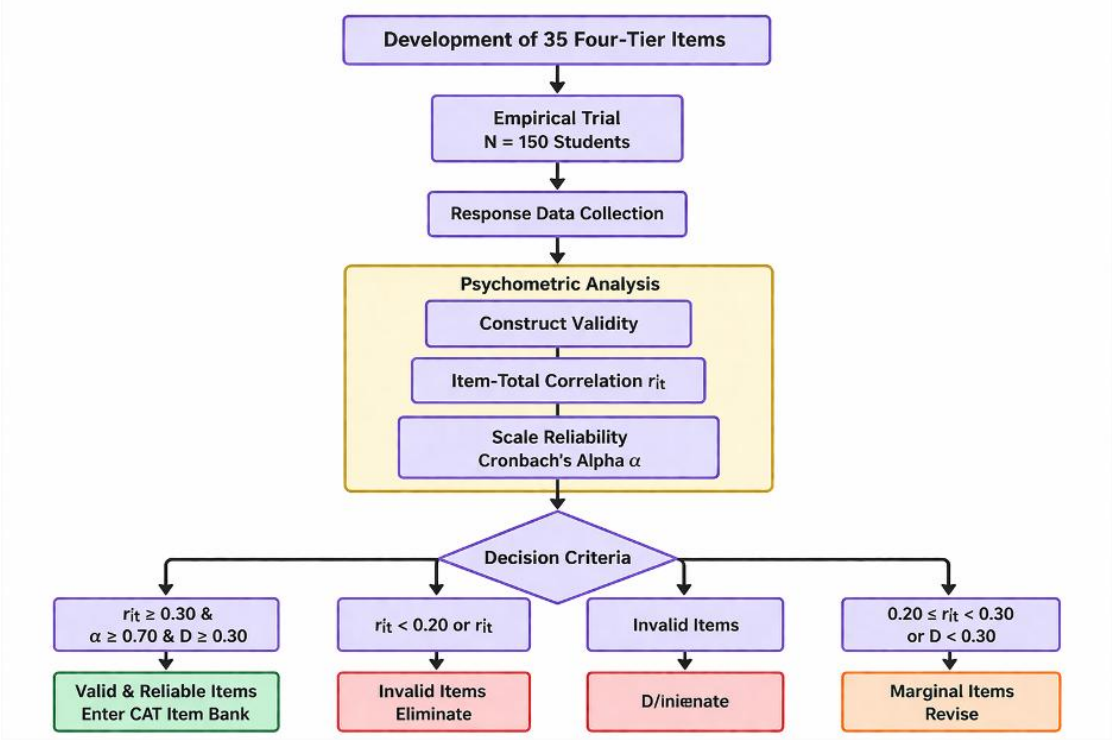

Figure 1. Research Flowchart

The research population were students majoring in science education (physics, chemistry, biology) from three universities in Aceh: Serambi Mekkah University, Syiah Kuala University, and UIN Ar-Raniry, who were in their fifth to seventh semesters. The research sample consisted of 150 students, selected purposively based on the criteria of having taken Integrated Science or integrative science courses and their willingness to participate. This number met the recommended respondent-to-item ratio for stable psychometric analysis.

The research instrument consisted of 35 integrated science diagnostic questions, developed in a Four-Tier Diagnostic Test format, referring to the framework (Gurel et al., 2015; Kaltakci-Gurel et al., 2017). This instrument measures two main constructs, namely science literacy (dimensions of content knowledge, process competence, and application context) and misconceptions (covering the domains of physics, chemistry, biology, and earth and space). Each item is designed to identify not only the correctness of the answer but also the level of confidence and the underlying reasoning. Thus, it leads to a more comprehensive diagnosis.

Student's responses were converted into binary scores (1 or 0) based on the combination of the correctness of the answer and the rationale, as well as the level of confidence. A response was categorized as "sound understanding" (score 1) if the student answered correctly on Tier 1 and Tier 3, and expressed high confidence (≥ 80%) on Tier 2 and Tier 4. Conversely, a response is categorized as a

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

misconception, lack of knowledge, or a guess (score 0) if there is a mismatch between the correctness of the answer and the reasoning, or if the level of confidence is low even though the answer is correct by chance.

Before the empirical testing, the instrument was first validated by three experts, including a science education expert, a learning assessment expert, and a science expert (in physics, chemistry, and biology). Content validity was calculated using Aiken's V coefficient to evaluate the alignment of each item with the measured indicators, covering aspects of content, construct, and language. Items with an Aiken's V value of ≥ 0.80 were deemed content-valid and feasible for testing.

The construct validity of the instrument was tested through item-total correlation analysis. The correlation coefficient was calculated using the following formula:

$$r\_it = (\sum ▒(X\_i - X^{-}\_i)(T - T^{-}))/\sqrt{(\sum ▒[\![(X\_i - X^{-}\_i)\text{^}2 \sum ▒(T - T^{-})\text{^}2]\!])}$$

where r_it is the item-total correlation, X_i is the i-th item score, T is the total score, and X^{-}_i and $T^{-}$is the average of score. Items with item-total correlation (r_it) ≥ 0,30 categorized as having good validity, while items with r_it between 0.20 and 0.29 are in the marginal category and require revision. Items with a correlation below 0.20 or a negative value are considered invalid and are eliminated from the item bank because they are indicated as not measuring the same construct or even contradicting the scale.

The overall reliability of the instrument was estimated using Cronbach's Alpha (α) coefficient. This coefficient was calculated using the following formula:

$$a = k/(k-1)\ (1 - (\sum ▒\sigma\_i\text{^}2)/(\sigma\_T\text{^}2))$$

where k is the number of grains, σ_i^2 is the variance of the i-th item score, and σ_T^2 is the variance of the total score. An α value ≥ 0.70 indicates that the instrument has adequate reliability for group measurement. If the α value reaches 0.80 or higher, the reliability is categorized as high and the instrument is eligible for use in decision making at the individual level. The analysis also includes examining the Alpha coefficient if an item is deleted (Alpha if Item Deleted) to identify items that, if removed, can actually improve the internal consistency of the scale.

Additionally, item analysis was conducted to assess the quality of each item. The item difficulty level (Difficulty Index/p) was calculated as the proportion of respondents who answered correctly, with an optimal range between 0.30 and 0.70. The item discrimination index (D) was calculated by comparing the performance of the high-ability (top 27%) and low-ability (bottom 27%) groups of respondents using the formula D=(U-L)/(N/2), where U and L each are the correct number in the upper and lower groups. N is the total number of respondents. Items with a discrimination index ≥ 0.30 are considered capable of distinguishing between the two groups effectively. Items that pass all of these psychometric criteria will be retained as anchor items in the question bank for CAT system.

The data collection procedure was conducted online through an online survey platform with randomized item order settings to minimize bias. Then, the collected response data was analysed using statistical software to produce objective decisions regarding the feasibility of each item, thereby ensuring the quality of the diagnostic instrument

RESULT AND DISCUSSION

Results of Scale Reliability Analysis

The reliability analysis of the instrument was conducted using Cronbach's Alpha (α) coefficient to measure the internal consistency of all items. Based on the statistical calculations, Cronbach's Alpha value was obtained at 0.798 with a 95% confidence interval ranging from 0.752 to 0.838. These results are presented in Table 1.

Table 1. Instrument Scale Reliability Statistics

Estimated	Cronbach's α value

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

Estimates of point	0.798
Lower limit of 95% CI	0.752
Upper limit of 95% CI	0.838
Note: CI = Confidence Interval	

The value of $\alpha = 0.798$ indicates that the instrument has a high level of reliability ($\alpha > 0.70$) based on the criteria proposed by (Nunnally & Bernstein, 1994; Ventura-León & Peña-Calero, 2020) Nunnally & Bernstein (1994). A narrow confidence interval does not include values below 0.70, indicating that this reliability estimate is stable and statistically reliable.

A Cronbach's Alpha value of 0.798 indicates that the Computerized Adaptive Test (CAT) instrument for Integrated Science has adequate to high internal consistency. This result meets the minimum standard of 0.70 for research instruments (Edelsbrunner et al., 2025; Fraenkel & Wallen, 1990) and even approaches the recommended standard of 0.80 for individual diagnostic instruments. These findings are consistent with research (Oladele & Ndlovu, 2021; Rahim et al., 2023; Wauters et al., 2010), which states that the question bank for the CAT system must have a minimum reliability of 0.75 to ensure the accuracy of individual ability estimates.

The 95% confidence interval (0.752 - 0.838), which does not include values below 0.70, indicating that this reliability estimate is robust and can be generalized to a broader population. It provides a strong psychometric basis for implementing the instrument in a CAT system, as a high reliability is a prerequisite for accurate adaptive measurement (Choi & McClenen, 2020; Embretson & Reise, 2013; Oladele & Ndlovu, 2021).

Results of Item Construct Validity Analysis

The validity of each item was analyzed through item-total correlation, which measures the value of an item contributes to the overall scale construct. The results of the analysis for 35 items are presented in Figure 2, which shows the distribution of item-total correlation coefficients.

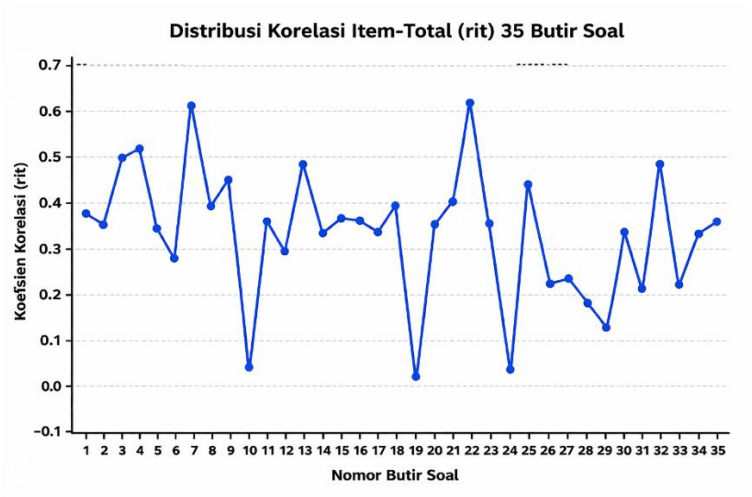

Figure 2. Distribution of Item-Total Correlation Coefficients

Based on the construct validity criteria established in the research method (rit ≥ 0.30), the analysis results show that 24 items (68.57%) meet the validity criteria well. 3 items (8.57%) are in the marginal category (0.20 ≤ rit < 0.30), and 8 items (22.86%) are invalid (rit < 0.20). The distribution of item validity categories is presented in Table 2.

Table 2. Item Validity Categorization Based on Item-Total Correlation Coefficient

Validity Category	Range rit	Number of Items	Percentage	Item Number

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

Very Good	≥ 0.50	6	17.14%	Q_3, Q_4, Q_7, Q_13, Q_22, Q_32
Good	0.30 - 0.49	18	51.43%	Q_1, Q_2, Q_5, Q_8, Q_9, Q_11, Q_14, Q_15, Q_16, Q_17, Q_18, Q_20, Q_21, Q_23, Q_25, Q_30, Q_34, Q_35
Marginal	0.20 - 0.29	3	8.57%	Q_6, Q_12, Q_27
Invalid	< 0.20	8	22.86%	Q_10, Q_19, Q_24, Q_26, Q_28, Q_29, Q_31, Q_33

Item-total correlation analysis revealed interesting patterns on item quality. A total of 24 items (68.57%) showed an adequate correlation (rit ≥ 0.30), indicating that the majority of items consistently measured the construct of science literacy and misconceptions in Integrated Science. Items Q_7 (0.621) and Q_22 (0.627) were the strongest contributors to the scale, indicating that these items had high discriminatory power and were able to clearly distinguish between students who understood the concepts and those who had misconceptions.

Domain-Specific Analysis of Item Discrimination

Further analysis of the relationship between item-total correlation and scientific content domains reveals meaningful pedagogical insights. Items with very high discrimination (rit ≥ 0.50) predominantly assess concepts requiring cross-disciplinary integration, particularly between physics and biology. For instance, Q_7 and Q_22, which address energy transfer in biological systems (e.g., photosynthesis and cellular respiration from the perspective of thermodynamic principles), demonstrate the highest discriminatory power. This finding suggests that items demanding concept transfer, where students must apply physical laws (e.g., energy conservation, entropy) to biological contexts, are particularly effective at distinguishing students with a sound understanding from those holding misconceptions.

Conversely, items classified as invalid (rit < 0.20) were predominantly those assessing standalone chemistry concepts, particularly those related to matter classification (e.g., elements, compounds, mixtures) without explicit connections to other disciplines. Items such as Q_10, Q_19, and Q_24, which focus on the particulate nature of matter and phase changes in isolated chemical contexts, showed very low or negative correlations. This pattern indicates that in an interdisciplinary diagnostic instrument, items confined to a single discipline may fail to align with the overarching construct of integrated science literacy, which inherently emphasizes connections across domains.

These domain-specific patterns provide important pedagogical insights: effective diagnosis of misconceptions in integrated science requires items that explicitly bridge disciplinary boundaries. Items that isolate concepts within a single domain may not adequately capture the integrative thinking skills that characterize science literacy in interdisciplinary contexts.

Validity Patterns in the Context of Prior Research

The identified proportion of invalid items (22.86%) warrants comparison with findings from previous research on interdisciplinary diagnostic instruments. This rate is consistent with studies by (Istiyono et al., 2023), who reported that approximately 18–25% of items developed for integrated science diagnostic tests failed to meet validity criteria in initial validation phases. Similarly, (Rahim et al., 2023) found that 21.5% of items in their four-tier diagnostic instrument for interdisciplinary physics-chemistry concepts required elimination due to low item-total correlations. Also, (Oladele & Ndlovu, 2021) documented comparable proportions (19–24%) of invalid items during the calibration phase of their CAT system for integrated mathematics and science.

The convergence of these findings suggests that developing high-quality diagnostic instruments for interdisciplinary contexts inherently involves a higher item attrition rate compared to single-discipline instruments. This phenomenon can be attributed to several factors: (1) the complexity of constructing items that accurately represent cross-disciplinary integration, (2) the increased cognitive load on respondents when navigating multiple disciplinary frameworks within a single item, and (3) the greater likelihood of ambiguous interpretations when concepts from different disciplines share similar terminologies but differ in underlying principles.

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

Implications for the Development of an Integrated Science CAT System

These findings provide clear guidance for improving the question bank for CAT system. First, six items with very high correlations (rit ≥ 0.50) used as anchor items or opening items in the adaptive algorithm, as it has excellent discriminatory power. Notably, these high-performing items predominantly feature integrated concepts requiring concept transfer across disciplines, suggesting that future item development should prioritize the integrative content.

Second, three marginal items (rit 0.20-0.29) require substantive revision before included in the question bank. Particular attention should be given to ensuring these items reflect meaningful cross-disciplinary connections rather than isolated disciplinary knowledge.

Third, eight invalid items (rit < 0.20) need to be eliminated from the question bank because they do not contribute to the measurement of the intended construct. In particular, items with negative correlations (Q_10, Q_19, Q_23) must be reviewed in depth. Possible causes consist of conceptual errors in item formulation or technical problems in scoring. In the context of the Four-Tier Test for diagnosing misconceptions, negatively correlated items may actually identify complex patterns of understanding that are not consistent with the main construct. Thus, it requires additional qualitative analysis of student's response patterns (Istiyono et al., 2023; Oladele & Ndlovu, 2021; Rahim et al., 2023).

CONCLUSION

The particular study concludes that the Computerized Adaptive Test (CAT) based diagnostic instrument for measuring science literacy and identifying misconceptions in integrated science courses has met basic psychometric requirements but requires selective refinement. The instrument demonstrated high reliability (α = 0.798), confirming that the item bank possesses the internal consistency essential for stable ability estimation in adaptive testing systems, a prerequisite that validates the feasibility of implementing CAT within interdisciplinary science contexts. The finding of the majority items exhibited sound construct validity indicates a sufficient of high-quality items to support adaptive algorithms. Yet, the identification of a nontrivial proportion of items with inadequate validity, including those with negative correlations, reveals a critical insight for the field: in interdisciplinary domains, the calibration phase of CAT systems must incorporate rigorous, iterative psychometric evaluation, as not all items survive empirical test. And, the attrition is not a limitation but rather an essential feature of responsible test development that ensures only items with proven psychometric integrity populate the item bank. These findings contribute to adaptive testing by demonstrating that CAT frameworks can be successfully adapted for integrative content provided that domain-specific validation protocols are employed. Moreover, they highlight that continuous quality monitoring and selective refinement are fundamental to maintaining the accuracy of adaptive algorithms. Thus, this study provides both a validated foundation for a CAT-based diagnostic tool in integrated science education and a methodological model for future adaptive testing development in complex interdisciplinary contexts.

REFERENCE

Amelia, R. N., Listiaji, P., Dewi, N. R., Heriyanti, A. P., Atmaja, B. D., Shoba, T. M., & Sajidi, I. (2024). Developing and Validating a Rubric for Measuring Skills in Designing Science Experiments for Prospective Science Teachers. *Jurnal Inovasi Pendidikan IPA*, *10*(1), 32–46. https://doi.org/10.21831/jipi.v10i1.65853

Choi, Y., & McClenen, C. (2020). Development of Adaptive Formative Assessment System Using Computerized Adaptive Testing and Dynamic Bayesian Networks. *Applied Sciences*, *10*(22), 8196. https://doi.org/10.3390/app10228196

Edelsbrunner, P. A., Simonsmeier, B. A., & Schneider, M. (2025). The Cronbach's Alpha of Domain-Specific Knowledge Tests Before and After Learning: A Meta-Analysis of Published Studies. *Educational Psychology Review*, *37*(1), 4. https://doi.org/10.1007/s10648-024-09982-y

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Psychology Press.

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

Fahmi, F., Chalisah, N., Istyadji, M., Irhasyuarna, Y., & Kusasi, M. (2022). Scientific literacy on the topic of light and optical instruments in the innovation of science teaching materials. *Jurnal Inovasi Pendidikan IPA*, *8*(2), 154–163. https://doi.org/10.21831/jipi.v8i2.41343

Fraenkel, J. R., & Wallen, N. E. (1990). *How to design and evaluate research in education.* ERIC.

Gurel, D. K., Eryılmaz, A., & McDermott, L. C. (2015). *A review and comparison of diagnostic instruments to identify students' misconceptions in science*.

Hartono, A., Djulia, E., Hasruddin, H., & Jayanti, U. N. A. D. (2023). Biology Students' Science Literacy Level on Genetic Concepts. *Jurnal Pendidikan IPA Indonesia*, *12*(1), 146–152. https://doi.org/10.15294/jpii.v12i1.39941

Ishtiaq Ahmed, & Sundas Ishtiaq. (2021). Reliability and Validity: Importance in medical research. *Journal of the Pakistan Medical Association*, *71*(10), 2401–2406. https://doi.org/10.47391/JPMA.06-861

Isnaini, F., Tiur, H., Silitonga, M., Musa, M., Hidayatullah, S., Sirait, J., & Afrizon, R. (2025). Diagnosing Students' Problem-Solving Challenges in Rotational Dynamics Using Two-Tier AR Flashcard Tests. *Jurnal Inovasi Pendidikan IPA*, *11*(2), 402–417. https://doi.org/10.21831/jipi.v11i2.80461

Istiyono, E., Dwandaru, W. S. B., Fenditasari, K., Ayub, M. R. S. S. N., & Saepuzaman, D. (2023). The Development of a Four-Tier Diagnostic Test Based on Modern Test Theory in Physics Education. *European Journal of Educational Research*, *volume-12-2023*(volume-12-issue-1-january-2023), 371–385. https://doi.org/10.12973/eu-jer.12.1.371

Kaltakci-Gurel, D., Eryilmaz, A., & McDermott, L. C. (2017). Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics. *Research in Science & Technological Education*, *35*(2), 238–260. https://doi.org/10.1080/02635143.2017.1310094

Lestari, S. (2021). pengaruh model pembelajaran peer led guided inquiry terhadap kompetensi literasi sains ditinjau dari kemampuan akademik. *Jurnal Inovasi Pendidikan IPA*, *7*(1). https://doi.org/10.21831/jipi.v7i1.29845

Maison, M., Lestari, N., & Widaningtyas, A. (2020). Identifikasi Miskonsepsi Siswa Pada Materi Usaha Dan Energi. *Jurnal Penelitian Pendidikan IPA*, *6*(1), 32–39. https://doi.org/10.29303/jppipa.v6i1.314

Nasyidiah, F. I., Siahaan, P., & Sasmita, D. (2020). PENGEMBANGAN INSTRUMEN FOUR-TIER DIAGNOSTIC TEST UNTUK MENDETEKSI MISKONSEPSI SISWA KELAS X PADA MATERI IMPULS. *WaPFi (Wahana Pendidikan Fisika)*, *5*(2), 31–40. https://doi.org/10.17509/wapfi.v5i2.27156

Ni'mah, F. (2019). Research trends of scientific literacy in Indonesia: Where are we? *Jurnal Inovasi Pendidikan IPA*, *5*(1), 23–30. https://doi.org/10.21831/jipi.v5i1.20862

Nunnally, J., & Bernstein, I. (1994). *Psychometric Theory 3rd edition (MacGraw-Hill, New York)*.

Nurhidayatulah, N., & Prodjosantoso, A. K. (2018). Miskonsepsi materi larutan penyangga. *Jurnal Inovasi Pendidikan IPA*, *4*(1), 41–51. https://doi.org/10.21831/jipi.v4i1.10029

OECD. (2023). *PISA 2022 Assessment and Analytical Framework, PISA* (PISA, Tran.). OECD Publishing. https://doi.org/10.1787/dfe0bf9c-en

Oladele, J. I., & Ndlovu, M. (2021). A Review of Standardised Assessment Development Procedure and Algorithms for Computer Adaptive Testing: Applications and Relevance for Fourth Industrial Revolution. *International Journal of Learning, Teaching and Educational Research*, *20*(5), 1–17. https://doi.org/10.26803/ijlter.20.5.1

Önder Çelikkanlı, N., & Kızılcık, H. (2022). A review of studies about four-tier diagnostic tests in physics education. *Journal of Turkish Science Education*, *19*(4).

Rahim, A., Hadi, S., Susilowati, D., Marlina, & Muti'ah. (2023). Developing of Computerized Adaptive Test (CAT) Based on a Learning Management System in Mathematics Final Exam for Junior High School. *International Journal of Educational Reform*. https://doi.org/10.1177/10567879231211297

Rohmadhani, I. A. N., Susilo, H., & Lestari, U. (2021). Identification misconceptions using Movement and Circulatory System Diagnostic Test (MCSD-Test) in XI class SMA/MA in East Java.

ISSN 2406-9205 (print), ISSN 2477-4820 (online)

Journal of Physics: Conference Series, *1918*(5). https://doi.org/10.1088/1742-6596/1918/5/052082

Rusilowati, A., Susanti, R., Sulistyaningsih, T., Asih, T. S. N., Fiona, E., & Aryani, A. (2021). Identify misconception with multiple choice three tier diagnostik test on newton law material. *Journal of Physics: Conference Series*, *1918*(5). https://doi.org/10.1088/1742-6596/1918/5/052058

Suparno, P. (2013). *Miskonsepsi dan perubahan konsep dalam pendidikan fisika*. Grasindo.

Susongko, P., Abdul Wahab, N. B., Arfiani, Y., & Kusuma, M. (2024). Validation and Implementation of 3-Dimensional Scientific Literacy Test (Lisa3D Test): Measuring Scientific Literacy for Senior High School Students based on Scientific Reasoning, Scientific Inquiry, and Nature of Science. *Jurnal Pendidikan IPA Indonesia*, *13*(3). https://doi.org/10.15294/591rx526

Taqwim, M. A., Sunarno, W., & Ramli, M. (2022). Remediation using SSCS model for reducing misconceptions about work and energy. *Jurnal Inovasi Pendidikan IPA*, *8*(2), 210–223. https://doi.org/10.21831/jipi.v8i2.49343

Ventura-León, J., & Peña-Calero, B. N. (2020). El mundo no debería girar alrededor del alfa de Cronbach ≥ ,70. *Adicciones*, *33*(4), 369–372. https://doi.org/10.20882/adicciones.1576

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: Possibilities and challenges. *Journal of Computer Assisted Learning*, *26*(6), 549–562.

Widarti, H. R., Nuriyanti, D., Sari, M. E. F., Wiyarsi, A., Yatimah, S., & Rokhim, D. A. (2024). Identification of learning difficulties and misconceptions of chemical bonding material: A review. *Ecletica Quimica*, *49*. https://doi.org/10.26850/1678-4618.eq.v49.2024.e1508

ISSN 2406-9205 (print), ISSN 2477-4820 (online)