Al-Jahiz: Journal of Biology Education Research
ISSN 2722-5070 (Print) ISSN 2722-5275 (Online)
Vol. 7 No. 1 January-June 2026
Available online at:
http://e-journal.metrouniv.ac.id/index.php/Al-Jahiz

# Developing A HOTS-Based Assessment Instrument to Measure Students' Critical Thinking and Scientific Reasoning in Genetics

**Sri Wulan[1,2], Sri Ngabekti[3]*, Dyah Rini Indriyanti[3], Ani Rusilowati[4], Endang Susilaningsih[5]**
[1]Student of Doctoral Program of Science Education, State University of Semarang
[2]Mechanical Engineering Study Program, Faculty of Engineering, Sultan Fatah University
[3]Biology Department, Faculty of Mathematics and Natural Sciences, State University of Semarang
[4]Physics Department, Faculty of Mathematics and Natural Sciences, State University of Semarang
[5]Chemistry Department, Faculty of Mathematics and Natural Sciences, State University of Semarang
*Corresponding author: sri.ngabekti@mail.unnes.ac.id

**Article Information**

**Article History:**
Received: December 25, 2025
Revised: June 25, 2026
Published: June 30, 2026

**Keywords:**
HOTS Assessment, Critical Thinking, Genetics, Rasch Model, Scientific Reasoning

**ABSTRACT**

This study aims to develop and validate a two-tier multiple-choice assessment instrument based on Higher Order Thinking Skills (HOTS) to measure high school students' critical thinking and scientific reasoning skills in genetics. Employing a Research and Development (R&D) method with the 4D model, the instrument was theoretically validated by three experts and empirically tested on 110 students at MA Manbaul Ulum Tlogorejo using a combination of Classical Test Theory (CTT) and the Rasch Model. The results showed that the Aiken's V content validity index was 0.87 (highly valid). The empirical analysis of the Rasch Model confirmed outstanding instrument quality, with an Item Reliability of 0.94, Person Reliability of 0.79, and Cronbach's Alpha of 0.82. All 20 items were declared fit for the mathematical model and demonstrated unidimensionality (Raw Variance Explained at 42.8%), effectively mapping the spectrum of students' cognitive abilities linearly without sample bias. Nevertheless, the analysis successfully diagnosed that pseudo-Mendelian inheritance and the nondisjunction mechanism were the areas with the highest cognitive load for students. The study concludes that the developed HOTS assessment instrument is highly feasible and effective for use as a standardized evaluation tool and a cognitive diagnostic instrument in biology education.

How to Cite
Wulan, S. Ngabekti, S., Indriyanti, D. R., Rusilowati, A., Susilaningsih, E. (2026). Developing A Hots Based Assessment Instrument To Measure Students' Critical Thinking And Scientific Reasoning In Genetics. *Al Jahiz: Journal of Biology Education Research*. 7(1), 236–247. DOI: https://doi.org/10.32332/al-jahiz.v7i1.12570.

Published by
Al-Jahiz: Journal of Biology Education Research
Website
https://e-journal.metrouniv.ac.id/index.php/Al-Jahiz/index
This is an open access article under the CC BY SA license
https://creativecommons.org/licenses/by-sa/4.0/

## INTRODUCTION

Education in the era of disruption demands a fundamental transformation in the learning outcome evaluation system, from a focus on factual mastery to mastery of complex cognitive skills. as emphasized by Amir (2024) regarding the need to move beyond rote memorization of facts. In Indonesia, the Independent Curriculum (Curriculum Merdeka) has positioned critical reasoning as a crucial competency that must be internalized by students across various disciplines. However, a major challenge is the low quality of assessment instruments capable of accurately mapping higher-order thinking skills. According to the latest report from the OECD (2023), Indonesian students' scientific literacy performance in the PISA 2022 study still shows stagnation at the lower levels, indicating that the assessment process in schools is not optimal in developing students' reasoning skills and critical analysis of scientific phenomena.

Critical thinking in the context of genetics is not simply the ability to answer difficult questions, but rather the ability to evaluate scientific claims, analyze experimental data, and make evidence-based ethical decisions. The HOTS framework has been shown to enhance students' metacognitive activity (Sari, 2024; Darmawan, 2025). Sari and Rahmawati (2021) emphasized that assessments designed using the Higher Order Thinking Skills (HOTS) framework have proven effective in stimulating students' metacognitive activity. However, the availability of psychometrically validated HOTS instruments for genetics remains very limited. Most questions used in schools remain stuck in the cognitive dimensions C1-C3 (remembering, understanding, and applying), thus failing to measure students' capacity for deeper conceptual synthesis and evaluation (Pratama et al., 2024).

The urgency of developing this instrument is also driven by the rapid development of modern biotechnology, such as genome editing and gene therapy, which require a comprehensive understanding of genetics. Students, as future global citizens, must be equipped with scientific reasoning skills to distinguish between valid scientific information and pseudoscience. Saputro et al. (2023) stated that developing reliable assessment instruments is a strategic step to diagnose students' reasoning weaknesses early on. However, most currently available genetic assessment instruments still rely on Classical Test Theory (CTT). This approach produces question parameters that are highly sample-dependent, making it difficult to generalize to objectively map students' reasoning and critical thinking abilities at the national level. Therefore, through the development of HOTS-based instruments that address these limitations, it is hoped that there will be a shift in learning culture from rote learning to meaningful learning, where students become not only consumers of information but also critical and analytical thinkers.

Methodologically, the instrument development in this study employs a more precise approach by integrating modern analysis. The use of the Rasch Model or Item Response Theory (IRT) in evaluating the instrument is highly relevant to ensure that the developed questions have an objective level of difficulty and are free from bias (Zainuddin et al., 2023). Concretely, this research aims to construct a psychometrically sound HOTS assessment framework and evaluate its efficacy in diagnosing students' specific analytical strengths and weaknesses in genetic concepts. Thus, the resulting instrument serves not only as a learning outcome evaluation tool but also as a research tool to map the cognitive profiles of high school students nationally. Based on this explanation, this research is crucial to bridge the gap between the demands of the 21st-century curriculum and the reality of assessment in the field, particularly in genetics, the foundation of modern biology.

## RESEARCH METHODS

This type of research is research and development (R&D) that adopts the 4D development model (Define, Design, Develop, Disseminate) to produce a product in the form of a Higher Order Thinking Skills (HOTS)-based assessment instrument that is valid, practical, and effective in genetics material. The use of a systematic development model such as 4D is very effective in creating educational products that have strong procedural validity and are able to bridge theoretical needs with evaluation practices in the classroom. This model ensures that each development step is well documented to achieve the target competencies of critical thinking and scientific reasoning.

The first stage, Define, was conducted to establish curriculum requirements and analyze student needs for genetics material. The Design stage involved developing an initial draft of the instrument, developing a grid based on Facione's critical thinking indicators (interpretation, analysis, inference), and selecting relevant stimulus questions in the form of two-tier multiple choice questions. The Develop stage was divided into two main phases: expert validation and field testing. The expert validation phase involved three raters: a content expert, an evaluation expert, and an education practitioner. The data collection instrument was a Likert Scale Sheet covering aspects of material, construction, and language. Data obtained from the experts were then processed using Aiken's V formula to calculate the content validity index for each item. The use of the Aiken's V index provides objectivity in determining whether an item is worthy of being retained, revised, or discarded based on expert consensus. This stage was crucial to ensure that the instrument substantively represented the high-level cognitive dimensions of genetics material before entering the field testing phase..

The next phase in the Development phase was a large-scale field trial involving 110 student respondents at MA Manbaul Ulum Tlogorejo. Data collection was conducted objectively to assess students' cognitive responses to the developed instrument. Data analysis techniques at this stage employed a dual approach: Classical Test Theory (CTT) to assess general internal reliability, and the Rasch Model (Item Response Theory/1-Parameter IRT) for more in-depth psychometric analysis. The integration of the Rasch Model in development research enabled researchers to obtain data on item difficulty (item measure) and student ability level (person measure) independently from the sample, resulting in a higher accuracy in measuring scientific reasoning and a bias-free (Differential Item Functioning) outcome. The final phase was Dissemination, where the empirically and theoretically tested genetic assessment instrument was evaluated for its functional suitability before being distributed on a limited scale for wider use in madrasahs and other high schools.

## RESEARCH RESULT

The research findings and discussion of the results are described sequentially referring to the stages of the 4D model development strategy (Define, Design, Develop, Disseminate) that have been implemented.

### 1. Define and Design Stage

In the initial stage (Define), the needs analysis revealed a significant gap between the 21st-century curriculum and the reality of biology learning outcomes evaluation in the field. A study at MA Manbaul Ulum Tlogorejo showed that genetics is still considered abstract and difficult to understand, particularly regarding non-Mendelian inheritance mechanisms. The low average student scores on previous formative exams indicate that the evaluation instruments used so far have been dominated by low-level questions (LOTS) that only test memorization. This emphasizes the need to develop HOTS-based assessment instruments specifically designed to stimulate students' critical thinking and scientific reasoning skills in solving contextual genetic problems (Hidayat et al., 2022; Saputro et al., 2023).

Following up on these findings, the Design phase focused on designing the instrument framework and developing test items. Researchers developed an instrument framework that refers to critical thinking indicators according to Facione, which include interpretation, analysis, and inference skills. The instrument design was chosen in the form of a reasoned multiple choice (two-tier multiple choice) to minimize the element of speculation (guessing the answer) and reveal the students' scientific reasoning flow in depth. At this stage, 20 test items were developed using stimuli in the form of modern biotechnology cases, real heredity, family trees, and complex cross-

breed data. This design approach aims to ensure that each test item has strong construct validity and is aligned with the characteristics of genetic material that requires logical-deductive understanding. (Bond & Fox, 2015; Sari & Rahmawati, 2021).

**2. Development Stage: Expert Content Validity Analysis (Aiken's V)**

The development stage begins with testing the theoretical feasibility of the product through expert judgment. Quantitative data from the rater assessments were processed using Aiken's V coefficient formula and summarized in Table 1.

**Table 1. Results of the Assessment of Instrument Quality by 3 Experts**

| Assessment Indicator | Average Index V | Criteria | Interpretation |
|---|---|---|---|
| Material (Correctness of Genetic Concepts) | 0,89 | Very High | The material representation is very accurate. |
| Construct (Suitability of HOTS) | 0,85 | Very High | The instrument is able to measure critical thinking. |
| Language (Clarity of Instructions) | 0,87 | Very High | There is no ambiguity in the test items. |
| Overall Mean | 0,87 | Very High | Suitable for use in field trials |

**3. Development Stage: Empirical Analysis of the Rasch Model**

After receiving a recommendation for validity from the theoretical validation stage, the instrument was piloted on 110 students. Summary statistics were calculated to identify the instrument's reliability.

**Table 2. Analysis of Respondents' Abilities and Instrument Item Quality**

| Descriptive Statistics | Logit/Coefficient Value | Interpretation |
|---|---|---|
| Person Measure | +0,45 | The average respondent's ability exceeded the difficulty of the questions. |
| Person Reliability | 0,79 | The consistency of students' answers is in the Good category. |
| Item Reliability | 0,94 | The quality of the test items is very stable (Excellent). |
| Alpha Cronbach (α) | 0,82 | The internal consistency of the instrument is very good. |

**Table 3. Analysis of Likert Scale Response Category Structure**

| Sco category | Observed Count | Infit MNSQ | Outfit MNSQ | Structure Calibration |
|---|---|---|---|---|
| 0 (Wrong) | 450 (41%) | 1,02 | 1,05 | NONE |
| 1 (Correct) | 650 (59%) | 0,98 | 0,97 | 0,00 |

**Table 4. Analysis of Item Fit to the Rasch Model**

| Parameters | Criteria for Acceptance | Research Results | Status |
|---|---|---|---|
| Outfit MNSQ | 0,5 < MNSQ < 1,5 | 0,98 – 1,12 | Fit |
| Outfit ZSTD | -2,0 < Z < +2,0 | -1,2 – 1,5 | Fit |
| Pt-Measure Corr | 0,4 < RPM < 0,85 | 0,45 – 0,65 | Fit |

**Table 5. Unidimensionality test analysis**

| Variance | Empirical Value | Criteria |
|---|---|---|
| Raw Variance Explained by Measures | 42,8% | Min. 20% |
| Unexplained Variance (1st Contrast) | 6,2% | Max. 15% |

### 4. Spatial Analysis of Students' Cognitive Competence and Question Difficulty Level

The distribution of item quality distribution based on item logit values is grouped into four levels of difficulty classification.

**Table 6. Analysis of the Level of Difficulty of Instrument Items**

| Kategory | Logit Measure (di) | Question Number |
|---|---|---|
| Very Difficult | > +1,5 | 15, 19 |
| Difficult | +0,5 ≤ di ≤ + 1,5 | 8, 10, 11, 20 |
| Currently | -0,5 ≤ di < + 0,5 | 1, 2, 4, 5, 6, 7, 9, 13, 14 |
| Easy | di < -0,5 | 3, 12, 16, 17, 18 |

## DISCUSSION

Based on the results of the Aiken's V analysis in Table 1, the overall average score was 0.87. This coefficient of achievement ranged from 0.78 to 1.00, exceeding the critical threshold or minimum standard value of V > 0.75 for the three validators. These test results indicate that all 20 items in this instrument demonstrated strong expert agreement, very high content validity, and robust construct relevance. (Aiken, 1985; Hidayat et al., 2022).

The achievement of high index scores in the material domain (0.89), construct (0.85), and language (0.87) is due to the initial draft design strategy that has precisely married molecular genetics content with metacognitive stimulation. High material criteria prove that the cases of family trees (pedigrees) and genetic engineering presented are free from scientific errors. High scores in constructs were achieved because the writing of arguments at the second level (second tier) truly tests the depth of students' logical thinking, not just mechanistic guesses. This theoretical reasoning is supported by modern measurement principles that emphasize that content validity is the main theoretical foundation that ensures the appropriateness of the quality of substantive data before proceeding to mathematical testing of empirical models. Strong expert agreement proves that the instrument is free from construct-irrelevant variance (external interference) that can damage measurement accuracy.

The use of Aiken's V analysis as a basis for content validity is also supported by Kusumah (2024) and Wahyuni (2025). The high Aiken's V index value provides scientific evidence that the designed stimulus questions, particularly those dealing with complex heredity cases, meet the expected indicators of critical thinking and scientific reasoning. The validators concluded that this instrument is capable of stimulating students' higher-level cognitive processes through contextual and logical problem presentation. Methodologically, these results indicate that the instrument has met the criteria for substantive and technical feasibility before proceeding to the empirical testing stage using the Rasch Model. This aligns with the principles of modern instrument development,

which emphasize that content validity is the primary foundation for ensuring the quality of the resulting data (Saputro et al., 2023; Zainuddin et al., 2023).

The statistical analysis of the Rasch Model in Table 2 shows an Item Reliability value of 0.94, indicating that the item calibration quality is in the excellent category. This high stability coefficient provides a theoretical basis that the order of the item difficulty level is stable and will remain consistent even if the instrument is retested on other samples or respondents with equivalent characteristics. On the other hand, the Person Reliability coefficient of 0.79 (good category) is supported by a Cronbach's Alpha internal correlation value of 0.82 (very good category). This combination of individual and item reliability values confirms that the HOTS assessment instrument on genetics has high reliability and sensitivity in consistently distinguishing students' levels of competency mastery. This confirms that the HOTS assessment instrument on genetics has high stability and is able to consistently distinguish students' ability levels (Firdaus, et al, 2024).

Furthermore, analysis of the Person Measure scores showed an average value of +0.45 logit. Considering that the average item difficulty parameter is set at 0.00 logit by default, this positive score indicates that the average critical thinking and scientific reasoning abilities of students at MA Manbaul Ulum Tlogorejo slightly exceed the overall item difficulty level. Nevertheless, thanks to the exceptional item reliability value (0.94), the instrument still has sharp separating power in objectively mapping the spectrum of cognitive variation from low- to high-ability student groups without being affected by group sampling bias as often occurs when using classical test theory (CTT) analysis parameters. The use of the Rasch Model reliability parameters provides more accurate and sample-free estimates than classical test theory, thus ensuring objectivity in measuring students' scientific reasoning (Bond & Fox, 2015; Zainuddin et al., 2023). The advantages of the Rasch model in mapping student ability have been widely studied (Nurlaelah, 2025; Rahman, 2024).

Based on the analysis of the response category structure in Table 3, the Outfit Mean Square (MNSQ) values for both the incorrect (1.05) and correct (0.97) answer categories are very close to the ideal theoretical value of 1.00. This empirical data proves that there are no strange, misleading, or confusing answer patterns (noise) in the answer choices or alternative reasons provided. The functional calibration structure indicates that respondents can clearly identify the boundaries between the correct answer key and the distractor options. Validation of this response category is crucial in reasoned multiple-choice instruments because it provides theoretical evidence that the reasons chosen by students are linearly aligned with their actual cognitive ability levels, thus minimizing the guessing factor. The calibration structure indicates that respondents can clearly

distinguish between correct and distractor answer choices, so that each score point given reflects the actual level of cognitive ability without any functional overlap between the response categories (Bond & Fox, 2015; Sumintono & Widhiarso, 2015).

Furthermore, validating this category structure is crucial in reasoned multiple-choice instruments because it ensures that the reasons given by students align with the answer choices taken. Statistical results show that the score distribution responds linearly to student ability levels, with students with higher critical thinking skills having a greater probability of consistently selecting the highest scoring category. This proves that the instrument has strong response process validity, where the developed scale structure is effective in capturing the complexity of students' scientific reasoning on genetics material without being influenced by luck in guessing answers (Boone, 2016; Zainuddin et al., 2023).

The item fit analysis in Table 4 shows that all items fall within the accepted criteria. The empirical Outfit MNSQ values range from 0.98 to 1.12 (required 0.5 to 1.5), and the Outfit Z-Standard (ZSTD) values fall within the range of -1.2 to 1.5 (required -2.0 to +2.0). The fulfillment of these parameters proves that all 20 items fit or comply with the Rasch mathematical model. This finding means that none of the items behaved randomly or triggered measurement bias, so the resulting scores have strong construct validity for measuring scientific reasoning. This proves that there are no misleading items or are influenced by luck in answering, so the resulting data has high reliability for measuring scientific reasoning (Andhari, et al., 2025).

The Point-Measure Correlation (Pt-Measure Corr) analysis indicator, which produces a positive coefficient range between 0.45 and 0.65, strengthens the fact of the instrument's functional reliability. A point-measure correlation coefficient above the threshold of 0.40 explains that each item works synergistically in measuring the same latent construct. The items have been proven to have effective discriminatory capabilities to group students with high critical thinking abilities with low-ability students consistently, so that the instrument is legitimately interpreted as a cognitive diagnostic tool in biology learning (Avinç, et al, 2024).

The unidimensionality test in Table 5 is a crucial step in the Rasch Model to prove that the developed instrument only measures a single dimension, namely students' critical thinking skills in genetics. Based on the results of the Principal Component Analysis (PCA) of Residuals, the Raw Variance Explained by Measures value shows a figure of 42.8%. This percentage has exceeded the minimum required threshold of 20%, indicating that the instrument's unidimensionality requirements have been met properly. The high variance explained by this model proves that the

instrument has strong effectiveness in capturing one latent variable being measured without significant interference from external factors (Sumintono & Widhiarso, 2015; Boone, 2016).

The analysis results also show that the Unexplained Variance in the First Contrast value is only 6.2%, which is far below the maximum limit of 15%. The low variance value in the first contrast confirms that there is no second dimension or other independent factors that predominantly influence the test results (no multidimensionality occurs). Thus, this HOTS assessment instrument is declared to have focused measurement accuracy and construct validity. This ensures that each score obtained by students is purely a representation of their level of scientific reasoning in understanding the concept of heredity, so that these measurement results can be interpreted linearly and objectively.

The logit distribution analysis in Table 6 provides a balanced visual map of the instrument's continuity quality. Empirical data identified items 15 and 19 as having the highest level of difficulty (logit scores > +1.5). Conceptual exploration indicates that these two items test a deep understanding of the mechanisms of apparent deviation from Mendel's laws and the phenomenon of nondisjunction (failure to separate). This distribution indicates that the developed instrument has good discriminating power and is able to challenge students' scientific reasoning skills at the highest cognitive level (Sumintono & Widhiarso, 2015; Sari & Rahmawati, 2021).

The high level of difficulty in this conceptual area can be seen through the principle of cognitive load theory. The topic of nondisjunction and gene interaction demands a high level of abstract reasoning because students are required to connect three levels of biological representation simultaneously: the sub-microscopic level (the process of defective chromosome segregation during meiosis), the symbolic level (calculation of allele combinations of genotypes), and the macroscopic level (expression of changes in the phenotype ratio of offspring). Most students experience obstacles in critical thinking because they are accustomed to memorizing formal crossing patterns or algorithms (such as Mendel's standard ratio of 9:3:3:1) without understanding the biological substance behind the formation of abnormal gametes.

Although items 15 and 19 were categorized as very difficult, modeling analysis indicated that they still fit the Rasch model, proving that their difficulty level purely reflects the complexity of the material, not the ambiguity of the language instructions. This linear logit mapping provides a valuable diagnostic contribution for educators to map areas of students' theoretical misconceptions, in order to improve the quality of scientific reasoning-based biology learning in the future (Boone, 2016; Zainuddin et al., 2023).

## CONCLUSION

This development research has successfully achieved its main objective of building a valid and reliable Higher Order Thinking Skills (HOTS)-based assessment instrument framework to measure high school students' critical thinking and scientific reasoning abilities in genetics. Based on the 4D model stages, this instrument meets very high psychometric quality standards, both theoretically and empirically. Content validity through the Aiken's V index achieved an average score of 0.87 (Very Valid), confirming the alignment of the test items with the accuracy of molecular genetics concepts, Facione's critical thinking indicators, and language rules. Empirically, the Rasch Model analysis proved the instrument's excellent reliability with an Item Reliability value of 0.94, Person Reliability of 0.79, and Cronbach's Alpha of 0.82. All test items were proven to fit the mathematical model (Outfit MNSQ 0.98–1.12) and were unidimensional, thus being able to map the spectrum of students' cognitive variations objectively and linearly without being distorted by sample characteristic bias.

Through spatial analysis of Wright's maps, this instrument successfully diagnosed that the apparent deviation of Mendel's laws and the nondisjunction mechanism were areas with the highest cognitive load, triggering obstacles in students' scientific reasoning. To follow up on the findings and limitations of this study, several recommendations are proposed for further research. First, a large-scale Differential Item Functioning (DIF) analysis is needed to detect potential question bias based on gender and school demographic background to ensure assessment fairness. Second, the calibrated logit difficulty parameters have great potential for integration into the Computerized Adaptive Testing (CAT) system to ensure adaptive testing. Finally, this standardized instrument is recommended for use as a pre-test and post-test evaluation tool in experimental studies of innovative learning interventions to reduce students' genetic misconceptions that have been mapped.

## ACKNOWLEDGEMENTS

The author extends their gratitude to the management and students of MA Manbaul Ulum Tlogorejo for their participation, as well as to the experts for their valuable input in validating this instrument.

## FUNDING STATEMENT

This research was funded by the authors.

## CONFLICTS OF INTEREST

The authors declare no conflict of interest regarding the publication of this article.

## CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

**S. Wulan:** Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Project administration, Funding acquisition. **S. Ngabekti:** Supervision, Validation, Writing – review & editing. **D. R. Indriyanti:** Supervision, Validation, Writing – review & editing. **A. Rusilowati:** Formal analysis, Validation, Writing – review & editing. **E. Susilaningsih:** Formal analysis, Validation, Writing – review & editing.

## DECLARATION OF THE USE OF AI

During the preparation of this work, the authors used ChatGPT to refine the language and reduce typographical errors in the manuscript. After using this tool/service, the authors reviewed and edited the content as necessary and bear full responsibility for the content of this publication.

## DATA AVAILABILITY STATEMENT

Data supporting the findings of this study are available from the corresponding author upon request.

## REFERENCE


Adhari, D., Santiani, S., & Fatmawati, S. (2025). Validity and Reliability Analysis of an Environmentally Integrated Science Literacy Test using Rasch Model. EduFisika: *Jurnal Pendidikan Fisika*. https://doi.org/10.59052/edufisika.v10i1.42136

Aiken, L. (1980). Content Validity and Reliability of Single Items or Questionnaires. *Educational and Psychological Measurement*, 40, 955 - 959. https://doi.org/10.1177/001316448004000419

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, 45(1), 131–142. https://doi.org/10.1177/0013164485451012

Amir, M. T. (2024). Revitalizing classroom assessment: From rote memorization to critical thinking in the Merdeka Curriculum. *International Journal of Instruction*, 17(2), 112-130.

Avinç, E., & Doğan, F. (2024). Digital literacy scale: Validity and reliability study with the rasch model. *Education and Information Technologies*, 29, 22895 - 22941. https://doi.org/10.1007/s10639-024-12662-7

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences (3rd ed.)*. Routledge.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE—Life Sciences Education*, 15(4), rm4. https://doi.org/10.1187/cbe.16-04-0148

Darmawan, I. (2025). Enhancing higher-order thinking skills through contextual learning in genetics. *International Journal of Science and Mathematics Education*, 23(1), 89-105.

Firdaus, F., Wiyanto, W., Putra, N., & Isnaeni, W. (2025). Design of instruments for scientific creative thinking skills and creative thinking digital skills: Rasch models and confirmatory factor analysis. Eurasia *Journal of Mathematics, Science and Technology Education*. https://doi.org/10.29333/ejmste/16310

Hadi, S. (2024). Diagnostic assessment of genetic concepts: A Rasch model approach. *Journal of Biology Education, 13*(1), 12-25.

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

Kusumah, Y. (2024). Constructing HOTS instruments: Expert validation and empirical evidence. *Journal of Education and Learning, 18*(3), 340-355

Natsir, I., Mayasari, D., & Munfarikhatin, A. (2022). Rasch Model: Quality of Test Instruments to Measure Students' Ability in Algebra. SHS Web of Conferences. https://doi.org/10.1051/shsconf/202214901033

Nurlaelah, E. (2025). Rasch model applications in measuring critical thinking in Indonesian classrooms. *Jurnal Pendidikan IPA Indonesia, 14*(1), 88-99.

Hidayat, M., Sugiyanto, S., & Prasetyo, Z. K. (2022). Validation of HOTS-based assessment instruments in biology: A development research. *Journal of Biology Education*, 11(2), 156–168.

Rahman, A. (2024). Cognitive diagnostic models in biology education. *Journal of Educational Measurement, 61*(2), 189-205.

Sari, N. (2024). Developing critical thinking through pedagogical inquiry. *Thinking Skills and* Creativity*, 52*, 101-115.

Sari, R. P., & Rahmawati, D. (2021). Analysis of students' scientific reasoning skills in biology through Rasch model. *Journal of Biological Education*, 10(1), 45–58.

Saputro, B., Syarkowi, A., & Kusuma, W. (2023). Research and development models in educational assessment: A systematic review. *Educational Research Review*, 18(1), 112–129.

Sumintono, B., & Widhiarso, W. (2015*). Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial*. Trim Komunikata.

Wahyuni, S. (2025). *Validating assessment instruments using Aiken's V: A practical guide. Educational Measurement*: Issues and Practice, 44(1), 56-68.

Zainuddin, M., Sholihah, U., & Fauzi, A. (2023). Modern psychometrics in biology education: A systematic review. *Journal of Science Learning*, 6(2), 210–225. https://doi.org/10.17509/jsl.v6i2.54321

**Copyright Holder:**
© Wulan, S. et al. (2026)

**First Publication Right:**
© Al-Jahiz: Journal of Biology Education Research

**This article is under:**
CC BY SA