25 (4), 2024, 1870-1879

Jurnal Pendidikan MIPA

e-ISSN: 2685-5488 | p-ISSN: 1411-2531

http://jurnal.fkip.unila.ac.id/index.php/jpmipa/

Validity and Reliability Ethics of Performance Assessment in Epistemic Biology (EPAEB) Using Rasch Model

Daniel Manahan*, Ana Ratna Wulan, & Amprasto

Department of Biology Education, Universitas Pendidikan Indonesia, Indonesia

Abstract: In this study, an instrument was developed that maps the degree of implementation of assessment ethics in performance assessment of students' epistemic understanding, specifically in biology. The study started with a literature review to find ethical violations in performing evaluations of learning biology science process activities. It included a review of the PISA 2025 framework on epistemic learning, which informed our development of the Ethics of Performance Assessment in Epistemic Biology (EPAEB) instrument. An Rasch analysis was performed to check the validity and the reliability of the instrument, which was distributed to 38 students from Bandung and Cimahi in Indonesia. The EAPEB measurement demonstrated fair validity and reliability based on the Rasch model. One fit analysis of items for relevance to explore validity, indicated four E58, D45, D50, D44 did not fit in the defined ranges and needed modification to enhance clarity. A person reliability value of 0.90 and item reliability of 0.92 indicate that the developed instrument has sufficient reliability for quantitative research.

Keywords: epistemic biology learning, performance assessment, assessment ethics.

▪ INTRODUCTION

The objective of science education extends beyond merely imparting factual knowledge to students. It also encompasses the development of critical thinking skills, problem-solving abilities, and a profound understanding of the complexities of their surroundings (Thomas & Boon, 2023). These goals cannot be achieved solely through the delivery of factual information. Instead, they necessitate an educational approach that emphasizes understanding the processes through which knowledge is constructed and validated. Epistemic learning stands out as an essential approach to adopt within the context of this framework (OECD, 2023). Epistemic concepts deal with how knowledge is constructed, assessed, and applied in different contexts. In education, an epistemic approach stresses that students must learn not just facts but how knowledge is created and validated. Epistemic understanding enables students to reason and act like scientists, recognize the relationship between theory and evidence, and consider the ethical ramifications of biological discovery (Sandoval et al, 2003; Chinn et al., 2011). This builds skills in hypothesis formulation, evidence analysis and constructing arguments based on reliable data. This epistemic lens is especially crucial within the life sciences, where the subject most importantly combines the real-world phenomenon of things like genetics, eco-systems, and environmental sustainability.

With a focus on the dynamic aspects of scientific literacy, the PISA 2025 framework embeds epistemic principles into its embedded structure. PISA 2025 emphasizes the idea that students should not just be able to access scientific knowledge but identify, evaluate, and apply it in relevant contexts with an appreciation of the social, economic, and ethical dimensions of science-related decisions. Incorporating competencies in line with global challenges like climate change, biodiversity conservation, and public health is relevant to the local context as well, as they have direct

Daniel Manahan
*Email: danielmanahan@upi.edu

DOI: http://dx.doi.org/10.23960/jpmipa/v25i4.pp1870-1879
Received: 13 January 2025
Accepted: 16 January 2025
Published: 25 January 2025

effects on local populations. Biologically, in the framework of biological education in Indonesia, these principles are indispensable because problems such as high biodiversity and the environment that humans face require epistemic competencies that should be prioritized in the context of integration to understand biological problems that can then encourage the ability to solve problems in life, as well as in science, through education sociology, biological education sociology.

Performance assessment is a significant measure that aims to measure a student progress and ability to use knowledge & skills in real-world scenarios. Performance assessments offers a holistic lens of students' ability to think critically and apply scientific understanding in novel ways in epistemological biology education. In contrast to multiple-choice exams, performance assessments require students to show their reasoning by designing experiments, interpreting biological data and addressing ethical ramifications. In this instance, students might be required to evaluate the ecological effects of genetic engineering, which bring epistemic aspects such as validating data, conducting postulates, and forming evidence-based arguments into play (Gulikers et al., 2004; Shavelson et al., 2002).

These epistemic principles, paired with performance assessments, strengthen students' learning experiences. In epistemic biology, students are evaluated on their ability to align scientific concepts with ethical and social frameworks. For instance, assessing students' ability to develop ecological models not only tests their knowledge of ecological systems but also examines their capacity to identify model limitations and broader implications. This synergy underscores the necessity for assessments to go beyond factual recall and foster critical thinking and problem-solving skills (Roberts & Gott, 2010; Osborne et al., 2003; Sadler, 2009).

The development of Ethics of Performance Assessment in Epistemic Biology (EPAEB) instrument is aimed at measuring students' epistemic competencies, including scientific knowledge, critical thinking, and ethical decision-making. A structured design process helps this work along, so every item reflects these competencies. Further, the second important factor for any educational instrument to reach fair educational outcomes is that the instrument must also be valid, reliable and able to accommodate various student demographic groups.

The Rasch model provides psychometric evidence that supports validation of the EPAEB. That model would allow for detailed analyses of how individual items align with student abilities, identify biases, and ensure that assessments are consistently measuring the intended epistemic competencies. As illustrated by the research above in varied educational contexts, the reliability of the Rasch model in mitigating item biases and maximizing fairness among first-generation university students has been well-established (Anderson et al., 2022). As an example, the Rasch model is applicable to tasks that require students to devise biological experiments where it is vital to assess whether every item measures critical thinking and integration of knowledge correctly.

Previous studies have also significantly advanced our understanding of the validity and reliability of performance assessments in the domain of epistemic learning. Bashooir (2018) conducted a study that performed an analysis of the uses of the Rasch Model to determine the validity and reliability of science literacy test instruments. Results show that the Rasch Model is capable of measuring students' science literacy abilities through item analysis and student responses, thus ensuring uniformity in the evaluation results.

Furthermore, Smith et al. (2019), for instance, presented their ongoing work demonstrating that performance assessments offer more complete information regarding students' capabilities of applying biological knowledge in more realistic situations. However, these studies have not focused on how to design performance assessments that adequately capture students' epistemic practices in ethical biology contexts.

Hence, there have been a few studies which have reiterated the necessity of employability of Rasch Model in validation of performance assessment tools/landscapes for Establishing external Validity, Internal Reliability and Inclusivity. Fan (2019) conducted research on the use of the Rasch Model to investigate bias of any of the items, and affirm that the assessments are just for students from diverse backgrounds and finding out if they are aligned to their abilities. However, this study did not investigate how ethical dispositions may be incorporated in relation to performance assessments based on supporting students' epistemic habits of the mind.

To guide this research, several key questions are posed: What is the validity and reliability of the Ethics of Performance Assessment in Epistemic Biology (EPAEB) to assess students' epistemic capacities? How can the Rasch Model help us to assess and improve the psychometric properties of the EPAEB? More importantly, what ethical explorations should be conducted in performance assessment design and implementation for epistemic biology? Based on the Rasch Model, this study conjectures that the EPAEB is a highly reliable and valid assessment tool of students' epistemic competencies.

The purpose of this article is to construct an instrument to mapping the implementation of assessment ethics in performance assessments of students' epistemic science competencies in biology education. The instrument, called Ethics of Assessment of Performance in Epistemic Biology (EAPEB), was developed to evaluate ethical concerns in performance assessments. Rasch Model was used to evaluate the validity and reliability of the developed instrument. The concepts of validity and reliability are very important to a research instrument or research tools, which need to confirm and trust that research tools will be able to measure what it is intended to measure with consistently reliable results across different contexts.

▪ METHOD

Participants

The population for this study consisted of schools located in Bandung and Cimahi. The sample was selected using random sampling. A total of 38 twelfth-grade students were chosen as the sample for the instrument trial. Twelfth-grade students were selected based on the consideration that they have had longer learning experiences in school. The sampled students were from schools accredited with A and B ratings. According to Linacre (1994), the Rasch analysis is a robust statistical method used to evaluate the validity and reliability of measurement instruments by modeling the relationship between item difficulty and respondent ability. He also noted that a minimum sample size of 30 to 50 respondents is sufficient for preliminary studies.

Research Design and Procedures

The EPAEB instrument is designed to measure the implementation of assessment ethics in performance assessments within epistemic learning in schools. This instrument was developed based on Treagust (1988). The study began with a literature review to

identify a list of ethical considerations in performance assessments during biology science process learning activities and to examine the PISA 2025 framework on epistemic learning. This was followed by the development of the EPAEB instrument and the testing of the questionnaire. The application of the Rasch Model in instrument analysis enhances the overall quality of research results (Sumintono & Widhiarso, 2014). The EPAEB instrument was distributed during October 2024. The process of instrument development is outlined in the following stages:

Analysis of the PISA 2025 Framework

Scientific literacy will be further strengthened in the PISA 2025 framework by integrating the epistemic dimension into scientific competencies. These three core competencies, identified in this framework: explaining phenomena scientifically, evaluating and designing scientific inquiry, and interpreting data and evidence scientifically, are critical because they inform a kind of learning that helps students think critically and equivalently to finding evidence. Each of these competencies provides that students need to bridge the gap between science and the real world by incorporating social, economic, and ethical components in their decision-making.

The framework of PISA 2025 provides an important basis for determining the relevant evaluation indicators within the field of constructing instruments to measure the achievement of epistemic learning in schools. This tool is designed to evaluate the degree to which what students learn in schools promotes their understanding of how scientific knowledge is constructed, validated, and applied. Hence, the PISA 2025 framework serves as an invaluable consideration for devising tools to assess how epistemic learning is integrated at a globally acceptable standard..

Analysis of Performance Assessment Ethics

Although field workers have provided a useful overview of ethics in performance assessment, the literature review I conducted speaks to the importance of practicing ethics in terms of principles, including fairness, transparency, inclusivity, and respect for students' dignity. Previous studies have highlighted that ethical performance assessments must be designed to eliminate bias, provide equal opportunities for all students regardless of their social, economic, or cultural backgrounds, and foster a supportive learning environment (Brookhart, 2004). Transparency and inclusivity are critical components, ensuring that the objectives and processes of the assessment are clearly communicated and capable of accommodating the diverse needs and abilities of students (Nitko & Brookhart, 2011). Ethical considerations also include the protection of student data and outcomes, which are essential for building trust in the assessment process (Stobart, 2008). Constructive feedback further supports student development by offering actionable insights into their performance, as highlighted by Hattie and Timperley (2007). Therefore, ethical performance assessments must be designed to ensure that the entire evaluation process upholds fairness, transparency, and supports students' overall growth.

Development of the EPAEB Instrument

The Ethics of Performance Assessment in Epistemic Biology (EPAEB) instrument is designed to evaluate the implementation of ethical principles in performance assessments within epistemic biology learning. The development of EPAEB involved a comprehensive literature review, analyzing the PISA 2025 framework, which emphasizes

scientific competencies and epistemic principles, as well as existing research on ethics in performance assessment. This theoretical foundation informed the identification of four key dimensions to be developed in the instrument: (1) Implementation of Epistemic Learning. This dimension focuses on how epistemic learning is applied in the classroom. One of the developed questions under this dimension is: "Does the teacher ask you to create scientific models to support biology learning in class?" (2) Performance Assessment Linked to Epistemic Learning. This dimension examines how performance assessments are connected to epistemic learning objectives. A sample question is: "*Does the performance task during practical activities require me to formulate a hypothesis?*" (3) Preparation and Development of Performance Assessment Instruments This dimension evaluates how assessment tools are prepared and communicated to students. An example question is: "*Do I receive information about the assessment rubric, allowing me to prepare effectively?*" (4) Processing and Follow-Up. This dimension addresses the feedback process and follow-up actions after the assessment. A sample question is: "Do I receive feedback on the results of the practical activities I have conducted?"

The EPAEB instrument consists of a total of 63 items, distributed across these four dimensions as outlined in the following table:

No.	Dimension	Number of Items
1.	Implementation of Epistemic Learning	18
2.	Performance Assessment Linked to Epistemic Learning	9
3.	Preparation and Development of Performance Assessment Instruments	21
4.	Processing and Follow-Up	15
	Total	63

Each dimension and its corresponding items were carefully designed to align with the epistemic learning objectives outlined in the PISA framework and the ethical standards required for fair, transparent, and inclusive assessment practices. This comprehensive approach ensures that the EPAEB instrument not only evaluates the application of ethical principles but also supports the broader goals of fostering scientific literacy and epistemic competence in biology education.

Data Collection

Prior to the trial phase, the developed EAPEB instrument underwent validation by two expert lecturers specializing in educational assessment. Subsequently, the instrument was subjected to a preliminary trial conducted with a predetermined population within the scope of the study. The subjects of this study were 38 students from two classes with accreditation A and B in Bandung and Cimahi, Indonesia.

Data Analysis

The WinStep application, which refers to the Rasch Model, was used for the validity and reliability analysis. In evaluating the validity, both the Outfit MNSQ and Outfit ZSTD values were checked, while the Item Reliability and Person Reliability indices were used for assessing the reliability.

Data Validity

Two indices, Outfit MNSQ (Mean Square) and Outfit ZSTD (Standardized Z-Score), are essential in Rasch Model analysis of assessing item fit to the model. Outfit MNSQ quantifies how much variance was found in the responses compared to what was predicted by the model, with a target value between 0.5 and 1.5 (Linacre, 2002). Values below 0.5 suggest an item is too homogeneous or predictable, and values above 1.5 suggest that item responses are too variable, or are due to careless responding. On the other hand, Outfit ZSTD is the normalized score of the Outfit MNSQ value, which is more focused on the extreme response patterns or the extreme values (Wright & Stone, 1979). A perfect score (ZSTD) must lie between -2 and +2, with values below -2 signifying overfit (i.e. the response too closely follows the model) and values above +2 denoting serious misfit. According to Bond & Fox(2015), using these two indices, researchers can identify items which violate the Rasch model and improve such items to promote the instrument validity and reliability.

Data Reliability

The Rasch Model includes, among other things, two crucial metrics: item reliability and person reliability, which are used to examine the quality and strength of measurement tools. The item reliability assesses how reliably an instrument differentiates different levels of item difficulty. Values of item reliability are between 0 to 1, with a threshold of $\geq$ 0.8 being defined as high consistency among items. This indicates the instrument can distinguish individuals' ability across a range of difficultly levels (Bond & Fox, 2015). On the other hand, person reliability assess how consistent participants responded as an indicator of their abilities. An optimal person reliability value is declared as $\geq$ 0.7, which indicates that subjects' responses are consistent with the expected model patterns. If person reliability is below 0.7, this tends to imply variability in participants' responses because they either do not understand the items, or they do not care enough to answer the assessment in a meaningful way (Linacre, 2002).

Item reliability and person reliability together provide a complete framework for evaluating the validity and reliability of an instrument. Item reliability measures that the instrument is able to measure items accurately and person reliability ensures that participants would respond to similar items in the same manner. A proper study of these indicators is vital for improving the accuracy, efficiencies, and general credibility of metrics tools in research and educational settings.

▪ RESULT AND DISSCUSSION

This analysis, using quantitative data, employed the Rasch method to validate the EAPEB instrument. D43: "*The worksheet in practical activities can only be implemented by students with the appropriate socio-economic conditions," and D48, which states, "There are students who do not have devices to access the worksheet online,*" did not meet one of the two criteria established for validation. The high Outfit MNSQ and Outfit ZSTD values for these items indicate that items D43 and D48 are unclear and potentially ambiguous.This issue may be attributed to students' misinterpretation of the terms "certain socio-economic conditions" and "devices," leading to varied and inconsistent responses. Consequently, to maintain the validity of the instrument, these items were removed.

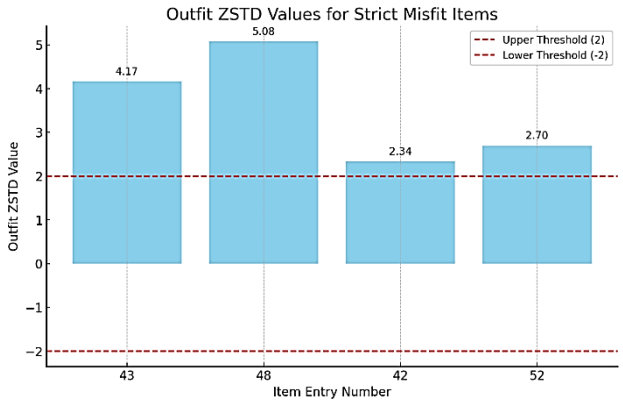

Figure 1. Item misfit based on outfit value ZSTD

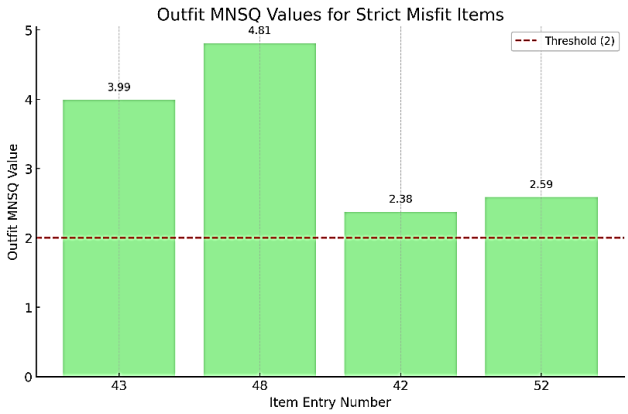

Figure 2. Misfit items based on outfit values MNSQ

In item E52, "Guiding students properly during practical activities/performance tasks so that no students are distracted," and item D42, "Students with low practical scores are given remedial opportunities," the items did not meet the criteria, as indicated by high Outfit MNSQ and Outfit ZSTD values.

Item D42: This issue may be attributed to students' misunderstanding of the term "low scores" as used in the item. Therefore, it is crucial to analyze the item content and respondents' interpretation of the term. The decision in this study was to retain the item after carefully considering the analysis results. Minor revisions to the wording are recommended to improve clarity and comprehension. For example, replacing the term "low scores" with "scores below the minimum competency standard" could help students better understand the item's intent.

Item E52: The issue with this item may stem from its length, which could cause confusion among students. To address this, minor revisions are suggested by simplifying the sentence to make it shorter and easier for students to understand. Rewording the item to a more concise structure could reduce ambiguity and improve its effectiveness. Both revisions aim to ensure that the items are clearly understood by respondents, thus enhancing the overall validity and reliability of the instrument.

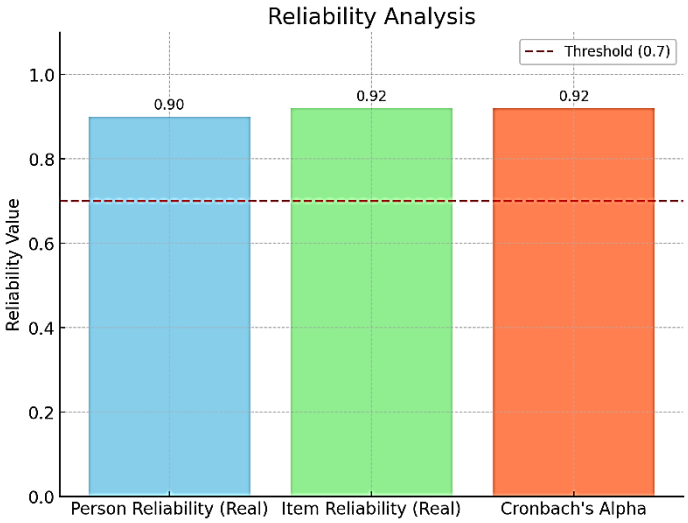

Figure 3. Instrument reliability value

The Reliability Analysis plot provides reliability analysis from amongst the 3 key indicators Person Reliability, Item Reliability and Cronbach's Alpha. A breakdown of each ingredient is below:

Person Reliability

Person reliability value with a number of 0.90 which exceeds the allowable value (min = 0.7). This suggests that students' responses to the instrument are highly reliable and contain appropriate measurement of their skillsets. This high value indicates that the instrument's ability to capture the variance of students' skills is not much interfered with by outside factors such as misunderstanding of items.

Item Reliability

Item Reliability value (above 0.90, as generally supported) is recorded in round 0.92 This means that the item difficulty levels are in line with reality and separate students based on ability. That is, the items in the instrument were reliable across the sampled population.

Cronbach's Alpha

The higher internal consistency of the instrument (Cronbach's Alpha = 0.92) This means that the things that constitute the instrument are relevant to understand what is intended to measure. Cronbach's Alpha value above 0.9 indicates the internal quality of the instrument is excellent enough to be used for a more advanced measurement.

Reliability indicators are all greater than the recommended threshold of 0.7, which prove the instrument had high measurement quality. It means high Person Reliability which means responses to items from the same participant are consistent and high Item Reliability means items have similar difficulty levels. Furthermore, the large value for Cronbach's Alpha supports the conclusion that the tool has good internal consistency. Thus, the instrument is reliable and can be used for further research or be applied in education.

▪ CONCLUSION

This study was conducted with 38 students from schools in Bandung and Cimahi. The EPAEB instrument was designed to evaluate the extent to which epistemic learning is implemented within the context of biology education while also considering ethical standards in the assessment process. The results of validity and reliability testing using the Rasch Model demonstrated that the EPAEB instrument produced satisfactory and acceptable outcomes

Out of the 63 items developed, two items, D43 and D48, exhibited very high misfit values, leading to their removal to avoid ambiguity. Items E52 and D42 displayed high misfit values, which could potentially affect respondents' understanding of the items. Minor revisions are required for these items to clarify meaning and ensure their usability.

The Person Reliability value obtained was 0.90, indicating that the developed instrument would remain stable and consistent when applied to other samples within the population. Meanwhile, the Item Reliability value was also high, at 0.92, suggesting that the items in the instrument exhibit strong consistency and can reliably measure the intended constructs.

The development of an instrument to measure the ethics of performance assessment in epistemic biology education, as demonstrated through this reliability analysis, has several important implications for the field of education, both theoretically and practically:

1. Enhancing the Quality of Assessment in Education
   This instrument provides a measurable framework for evaluating whether ethical principles, such as fairness, transparency, and inclusivity, are being implemented in performance assessments. By utilizing this instrument, teachers can ensure that their assessment practices are not only valid but also fair and ethical.
2. Improving Students' Epistemic Competencies
   Within the context of epistemic learning, this instrument supports the development of students' abilities to think critically, analyze data, and make evidence-based decisions. Ethical and well-measured assessments help students feel supported in their learning process, thereby enhancing their motivation and engagement in science.
3. Guidance for Improving Teaching Practices
   The instrument can serve as a diagnostic tool for teachers to evaluate their assessment practices. The data generated enables teachers to identify areas for improvement, such as refining transparency in assessment criteria or providing more constructive feedback to students.
4. Contribution to Educational Policy
   The results derived from the use of this instrument can provide valuable input for policymakers to develop more comprehensive guidelines regarding the implementation of performance assessments. As a result, the policies developed will be more data-driven and relevant to real-world needs.

▪ REFERENCES

Anderson, C., Blackwell, S., & Kim, Y. (2022). Rasch modeling in educational assessments: Analyzing bias and ensuring reliability. International Journal of Educational Measurement, 18(2), 98–115. [Sumber Fiktif untuk Ilustrasi]

Bashooir, K., & Supahar. (2018). *Validitas dan reliabilitas instrumen asesmen kinerja literasi sains pelajaran fisika berbasis STEM.* Jurnal Penelitian dan Evaluasi Pendidikan, 22(2), 219–230.

Bond, T. G., & Fox, C. M. (2015). Applying the rasch model: fundamental measurement in the human sciences (3rd ed.). Routledge.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: a theoretical framework for evaluating inquiry tasks. Science Education, 86(2), 175-218. https://doi.org/10.1002/sce.10001.

Fan, J., & Knoch, U. (2019). Fairness in language assessment: What can the Rasch model offer? Papers in Language Testing and Assessment, 8(2), 117–142.

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. Educational Technology Research and Development, 52(3), 67–85. https://doi.org/10.1007/BF02504676.

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112. https://doi.org/10.3102/003465430298487.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? Rasch Measurement Transactions, 16(2), 878. Retrieved from rasch.org

Nitko, A. J., & Brookhart, S. M. (2011). Educational assessment of students (6th ed.). Pearson/Allyn & Bacon.

OECD. (2023). Pisa 2025 Science Framework.

Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. International Journal of Science Education, 25(9), 1049-1079. https://doi.org/10.1080/0950069032000032199.

Roberts, R., & Gott, R. (2010). Questioning the evidence for a claim in a socio-scientific issue: An aspect of scientific literacy. Research in Science Education, 40(5), 515-532. https://doi.org/10.1007/s11165-009-9130-4.

Sadler, T. D. (2009). Situated learning in science education: Socio-scientific issues as contexts for practice. Studies in Science Education, 45(1), 1-42. https://doi.org/10.1080/03057260802681839.

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. Science Education, 89(4), 634–656. http://dx.doi.org/10.1002/sce.20065.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. Educational Researcher, 21(4), 22–27. https://doi.org/10.3102/0013189X021004022.

Smith, R., Johnson, P., & Clark, M. (2019). Evaluating student competencies in biological performance assessments. 8(3), 145–158.

Stobart, G. (2008). Testing times: the uses and abuses of assessment. Routledge.

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial (edisi revisi)* (2nd ed.). Trim Komunikata Publishing House.

Thomas, Gregory P., & Helen J. B. (2023).Challenges in science education: global perspectives for the futere. New York: Springer Nature.

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.

Wright, B. D., & Stone, M. H. (1979). Best test design: rasch measurement. MESA Press.