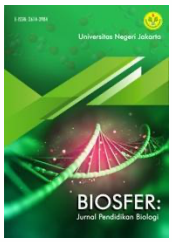
Biosfer: Jurnal Pendidikan Biologi

Journal homepage: http://journal.unj.ac.id/unj/index.php/biosfer



Rasch model analysis: Grade 11 biology education questionnaire accuracy

Gracia Cita Cinantya, Bowo Sugiharto*

Biology Education, Faculty of Teacher Training and Education, Universitas Sebelas Maret, Indonesia

*Corresponding author: bowo@fkip.uns.ac.id

ARTICLE INFO

Article history
Received: 09 July 2025
Revised: 28 September 2025
Accepted: 06 October 2025

Keywords:
Item Response Theory
Literacy
Psychometric Analysis
Rasch Model

ABSTRACT

Assessment is a crucial component of education, measuring student learning. Multiple-choice tests are one of the most common assessment methods, and their thorough analysis is essential to ensure fairness and accuracy. When measuring environmental literacy among high school students, the psychometric quality of the instruments must be carefully studied to ensure they truly reflect student competence. The Rasch model, a part of Item Response Theory, is a powerful tool for evaluating item quality by examining the fit of student responses to the Model's expectations and determining item difficulty. This study aimed to develop and validate an instrument to measure environmental literacy among Grade XI students using the Rasch model. The developed tool encompasses knowledge, thinking skills, attitudes, and actions related to the environment. The Point-Measure Correlation results indicated a positive relationship between student ability and their responses, although some items were identified as needing revision for improved performance. The average match between the Model's predictions and the actual student responses was 73.5%, signifying that most of the data aligned with the Model's expectations. This research provides a robust framework for the psychometric evaluation of assessment tools in science education.

INTRODUCTION

Assessment in education is a systematic effort carried out periodically, continuously, and thoroughly to obtain information related to the development and achievement of students in the learning process (Hayat, 2020). The need for more effective and efficient assessment instruments is increasing (Khalid, 2023). Instruments such as multiple-choice tests are commonly used in formal education, especially in assessing and evaluating students' environmental literacy competencies. Multiple-choice tests serve as a tool to measure the extent to which a person masters a specific competency based on the responses given to the questions asked (Gao et al., 2023). The questions in the test not only act as a measuring tool, but also as an indicator of students' overall abilities (Asrijanty, 2014). Therefore, each item must be analyzed to ensure the instrument is feasible and qualified to measure students' ability (Ariyanto et al., 2021).

A good instrument is characterized by fulfilling validity and reliability requirements according to psychometric standards (Santos et al., 2016). The most commonly used instrument in assessing learning outcomes is the multiple-choice test. This test is considered practical in its implementation and can cover a broad scope of material. However, the validity and reliability of items in multiple-choice tests are often a significant problem. Many evaluation instruments do not undergo a thorough validation process, so the results do not represent students' abilities. Therefore, an analytical approach is needed that can identify and assess the quality of items in depth and accurately. One method that can be used to assess item characteristics is the Rasch model approach, which allows evaluation based on item parameters as a whole (Yokhebed et al., 2025). The Rasch model, as part of Item Response Theory (IRT), is designed to ensure that high scores obtained by test takers truly reflect high levels of ability (Piryani, 2024). This Model helps instrument developers match the difficulty level of questions with students' ability and ensure that each item contributes optimally to the measurement (Fitzpatrick et al., 2021).

The main advantage of the Rasch model is its ability to align item scale and student ability within the same framework, allowing for a more objective and comprehensive analysis of item fit (Wolfs et al., 2023). Rasch model analysis also makes it easier for educators to detect items that are too easy, too difficult, or even misfit, as well as address missing data and estimate the difficulty and ability of participants simultaneously (Embertson, 2013). In addition, statistical fit analysis allows for identifying items or participants inconsistent with the Model, thus improving measurement accuracy (Hayat et al., 2020; Demir, 2023). The Rasch model also allows calibration of three main aspects: the measurement scale, the test taker (person), and the item, all of which contribute to the validity of the measurement results (Dwiliesanti et al., 2022). In fact, this approach can predict item difficulty and participant ability simultaneously (Khalid et al., 2023). In developing an instrument in the form of a multiple-choice test, the Rasch model analysis plays a significant role. Rasch model analysis helps in ensuring the accuracy of measurement and assessment. It can be done by analyzing the answer response for each item and linking the relationship between the test taker's ability level and the item's difficulty level.

Traditional test analysis methods, such as Classical Test Theory (CTT), often fall short in educational measurement because they produce sample-dependent item and person statistics, making direct comparisons across different student groups problematic (Bond & Fox, 2015). A key limitation is their inability to properly untangle the measurement of item difficulty from student ability on a single, continuous, and objective scale. It often fails to identify the proper relationship between these two factors, especially when the assessed population has diverse characteristics. Consequently, many high school-level assessments, particularly in subjects like Biology, are created by teachers without robust analytical approaches like the Rasch model, leaving the validity and reliability of the scores and the resulting educational decisions uncertain (Boone, 2016). This widespread practice of using unvalidated instruments forms the core research problem. Based on this critical need to improve assessment quality, this study aims to analyze the psychometric accuracy of Grade 11 Biology multiple-choice items using the Rasch model approach. Through this research, we seek to develop a valid, reliable, and fair evaluation instrument for measuring student abilities. The findings will serve as a concrete basis for improving evaluation instruments and provide teachers with a practical reference for developing and assessing high-quality test questions.

METHODS

Research Design

This study uses the Item Response Theory (IRT) approach, with a focus on the Rasch Model, to develop and analyze an instrument for measuring environmental literacy of grade XI high school students. The Rasch Model was chosen because of its ability to provide a more precise analysis of item quality by linking the ability level of students (respondents) and the difficulty level of the items. In contrast to traditional analysis that only uses total scores, the Rasch Model allows researchers to evaluate how well each item functions in measuring the intended ability and identify inappropriate or biased items. Using the Rasch Model, this study aims to produce an instrument that is valid and reliable and can also accurately describe students' environmental literacy (Andrich et al, 2019).

This study adapted Mark Wilson's Four Building Blocks model in the instrument development process. The first stage, the Construct Map, focuses on mapping the concepts to be measured in the environmental literacy test. This mapping includes four main dimensions: knowledge about the environment, cognitive skills related to the environment, attitudes towards the environment, and pro-environmental behavior. After the mapping is complete, the next stage is Item Design, where the items are developed based on the indicators determined at the mapping stage. At this stage, the preparation of multiple-choice format questions is carried out in accordance with the characteristics of each dimension (Arjaya et al., 2024). The third stage in the Four Building Blocks model is Outcome Space, where each question is associated with clear assessment criteria. At this stage, the form of answer choices that will be used in the test and how to process scores for each question is clarified. Scoring is done to see the extent to which each answer describes the desired competency or level of environmental literacy. Finally, the fourth stage is the Measurement Model, where the designed instrument is tested using the Rasch Model to measure item validity and reliability (Yokhebed et al., 2025).

Participants

This study involved grade XI students from several purposively selected high schools as respondents. The sample was chosen to represent a population of high school students with diverse characteristics. Students were asked to take a developed multiple-choice test focused on measuring their environmental literacy.

Instruments

The instrument used to collect data was a multiple-choice test comprising various items categorized by different difficulty levels. Each item was tested to ascertain whether it could measure what was intended and how well it identified students' abilities. The data obtained from the test results were then analyzed using Winsteps software, a tool for conducting Item Response Theory (IRT) analysis focusing on the Rasch Model. Winsteps estimates item parameters, such as item difficulty, item fit, respondent fit, and overall instrument reliability (Andric et al., 2019). Through this analysis, it is possible to determine whether each item fits the Rasch model and whether it accurately measures student competence. In addition, the analysis also makes it possible to identify items that are not functioning correctly, either because they are too easy, too difficult, or do not match student response patterns.

Data Analysis Techniques

After analysis using the Rasch Model, the results are used to evaluate and improve the developed test instrument. Items that do not fit or have problems with statistical fit can be removed or adjusted to improve measurement quality. The calibration process also ensures that the test instrument measures students' abilities fairly and consistently. In addition, this analysis allows researchers to make further improvements to the items, so that the resulting instrument becomes more efficient in comprehensively measuring the environmental literacy of grade XI high school students.

RESULTS AND DISCUSSION

We evaluated 36 students and 20 items using the Rasch model in this analysis. Winsteps version 3.73 was used to process the data and generate student ability and item difficulty estimates. This analysis aims to understand the fit between the data and the Rasch model and the instrument quality used to measure student ability.

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	9.3	20.0	-.17	.52	.99	.0	1.01	.1
S.D.	3.5	.0	.92	.04	.16	.7	.29	.8
MAX.	17.0	20.0	2.04	.67	1.31	1.4	2.09	2.3
MIN.	4.0	20.0	-1.63	.49	.54	-1.5	.37	-1.3

REAL RMSE .54 TRUE SD .74 SEPARATION 1.37 Person RELIABILITY .65
MODEL RMSE .53 TRUE SD .76 SEPARATION 1.44 Person RELIABILITY .67
S.E. OF Person MEAN = .16

Person RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .68

Figure 1. Student Reliability.

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	16.8	36.0	.00	.39	1.00	-.1	1.01	.0
S.D.	6.4	.0	.98	.05	.17	1.1	.26	1.1
MAX.	31.0	36.0	1.96	.52	1.29	1.8	1.51	2.2
MIN.	5.0	36.0	-2.25	.36	.74	-1.9	.61	-1.9

REAL RMSE .41 TRUE SD .89 SEPARATION 2.16 Item RELIABILITY .82
MODEL RMSE .40 TRUE SD .90 SEPARATION 2.26 Item RELIABILITY .84
S.E. OF Item MEAN = .23

UMEAN=.0000 USCALE=1.0000
Item RAW SCORE-TO-MEASURE CORRELATION = -1.00

Figure 2. Correlation of Question Items.

1. Statistical Analysis of Person (Student)

The mean of the logit scores for students was -0.17, indicating that the average student ability was slightly lower than the difficulty of the questions. The standard deviation of student ability was 0.92, indicating a relatively high variation in ability between students. Person reliability was calculated with the Person Reliability coefficient of 0.65 for the Real estimate and 0.67 for the Model. This value indicates the instrument has moderate reliability in distinguishing student abilities (Fitzpatrick et al., 2021). Separation is 1.37, meaning the instrument can separate students' abilities into two groups (Wright et al., 2022).

The fit between the data and the Rasch model was measured by Infit Mean Square (MNSQ) and Outfit MNSQ. The mean value of MNSQ for Infit was 0.99, and for Outfit was 1.01, indicating that most of the data fit the Rasch model. The ideal value for MNSQ is 1.0, indicating that the data fit the expectations of the Rasch model (Boone et al., 2020). Cronbach's Alpha or KR-20 for internal reliability was 0.68, indicating sufficient instrument reliability to assess or measure students' abilities at an early stage (Ariyanto et al., 2021).

2. Statistical Analysis of Items

The mean logit for the 20 items tested was 0.00, indicating that the average item difficulty is at the midpoint of the Rasch scale. The standard deviation (SD) of item difficulty was 0.98, meaning there was considerable variation in item difficulty. Some items are very easy, while others are more difficult. Item Reliability is 0.82 for the Real estimate and 0.84 for the Model, which indicates that the items have excellent reliability in distinguishing students' ability levels (Gao et al., 2023). Item Separation is 2.16, which means the item can differentiate students into three different ability groups, indicating good quality to measure a diverse range of student abilities (Linacre, 2023). The Infit MNSQ value for items was 1.00, and for Outfit MNSQ was 1.01, indicating that the items fit the Rasch model and none of the items gave results that deviated from what was expected.

3. Global Model Fit

At the global level, the Log-Likelihood Chi-Square statistic shows a value of 787.79 with 665 degrees of freedom, and a p-value = 0.0007. It indicates a significant deviation between the data and the Model, although the small p-value should be interpreted cautiously, as the chi-square test is susceptible to large sample sizes. However, overall, these results suggest that the Rasch model is acceptable as a suitable model for this data (Fraenkel et al., 2012). The Root Mean Square Residual (RMSR) was 0.4289, indicating that the Model generally fit the data, although some minor discrepancies are common in Rasch analysis.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Person
27	17	20	2.04	.67	1.31	.8	1.17	.5	.09	.33	85.0	84.9	27L
22	16	20	1.64	.60	.54	-1.5	.37	-1.3	.77	.36	90.0	80.9	22L
34	16	20	1.64	.60	.89	-.2	.81	-.2	.47	.36	80.0	80.9	34L
11	15	20	1.30	.56	.98	.0	.86	-.2	.42	.38	80.0	78.0	11P
17	15	20	1.30	.56	.96	.0	.89	-.1	.42	.38	80.0	78.0	17P
7	14	20	1.00	.53	1.10	.5	2.09	2.3	.17	.39	75.0	75.1	07L
2	13	20	.73	.51	.79	-.9	.72	-.8	.60	.40	85.0	72.3	02P
31	13	20	.73	.51	1.15	.7	1.07	.3	.28	.40	65.0	72.3	31P
18	12	20	.47	.50	.94	-.2	.91	-.2	.46	.40	80.0	69.7	18P
5	10	20	-.01	.49	.89	-.7	.80	-.8	.53	.40	60.0	66.1	05P
6	10	20	-.01	.49	1.04	.3	.96	-.1	.38	.40	60.0	66.1	06P
8	10	20	-.01	.49	.89	-.7	.82	-.7	.52	.40	70.0	66.1	08L
12	10	20	-.01	.49	.95	-.3	.88	-.4	.47	.40	80.0	66.1	12P
14	10	20	-.01	.49	.90	-.6	.82	-.7	.51	.40	70.0	66.1	14L
29	10	20	-.01	.49	.92	-.5	.88	-.4	.48	.40	80.0	66.1	29P
1	9	20	-.25	.49	1.00	.1	1.07	.4	.38	.40	65.0	65.7	01P
16	9	20	-.25	.49	.91	-.5	.83	-.6	.49	.40	65.0	65.7	16P
19	9	20	-.25	.49	1.10	.7	1.20	.8	.28	.40	65.0	65.7	19L
10	8	20	-.49	.50	1.08	.5	1.15	.6	.30	.39	65.0	66.9	10L
15	8	20	-.49	.50	1.24	1.3	1.29	1.0	.16	.39	55.0	66.9	15L
30	8	20	-.49	.50	1.20	1.2	1.45	1.5	.15	.39	65.0	66.9	30P
35	8	20	-.49	.50	.95	-.2	1.24	.9	.38	.39	85.0	66.9	35P
3	7	20	-.74	.51	1.29	1.4	1.29	.9	.12	.38	55.0	69.5	03P
4	7	20	-.74	.51	1.04	.3	1.28	.9	.30	.38	75.0	69.5	04L
13	7	20	-.74	.51	1.01	.1	1.19	.7	.33	.38	75.0	69.5	13L
20	7	20	-.74	.51	1.21	1.1	1.16	.6	.20	.38	65.0	69.5	20L
23	7	20	-.74	.51	1.05	.3	1.31	1.0	.28	.38	75.0	69.5	23P
33	7	20	-.74	.51	.98	.0	.90	-.2	.41	.38	65.0	69.5	33P
21	6	20	-1.01	.53	.89	-.4	.84	-.3	.47	.37	80.0	73.4	21P
26	6	20	-1.01	.53	.76	-1.0	.63	-1.0	.60	.37	80.0	73.4	26L
32	6	20	-1.01	.53	.96	-.1	.96	.0	.40	.37	70.0	73.4	32P
36	6	20	-1.01	.53	.93	-.2	.96	.0	.42	.37	80.0	73.4	36P
9	5	20	-1.30	.55	.76	-.8	.61	-.8	.58	.35	85.0	77.2	09P
24	5	20	-1.30	.55	.85	-.5	.71	-.5	.50	.35	85.0	77.2	24P
28	5	20	-1.30	.55	1.03	.2	1.10	.4	.30	.35	75.0	77.2	28L
25	4	20	-1.63	.60	1.26	.8	1.17	.5	.12	.33	75.0	81.4	25P
MEAN	9.3	20.0	-.17	.52	.99	.0	1.01	.1			73.5	71.6	
S.D.	3.5	.0	.92	.04	.16	.7	.29	.8			9.1	5.4	

Figure 3. Analysis of Question Items.

1) Fit Statistics and Participants' Ability Estimates

The average logit value of participants' ability was -0.17 with a standard deviation of 0.92, indicating that students' ability levels were mainly below the average item difficulty. It means that, in general, the instrument was relatively challenging for most students. Judging from the infit and Outfit mean square (MNSQ) statistics, most participants had values between 0.7 and 1.3, which is still within the ideal range according to the criteria proposed by Linacre (2002), namely $0.5 \leq \text{MNSQ} \leq 1.5$.

Participants with both infit and Outfit values above 1.5, such as participant number 7 (Outfit = 2.09), indicated a misfit, which could be due to inconsistent responses, possible guessing, or a lack of understanding of a particular question. Participants with very low outfit scores, such as participants 22 (Outfit = 0.37) and 26 (Outfit = 0.63), showed responses that were too consistent compared to those expected by the Model. It may indicate that the answers were too structured, or that the participants only understood the easy questions. These responses must be further analyzed to understand the context or students' answering strategies.

2) PT-MEA (Point-Measure Correlation) Analysis

PT-MEASURE CORR (Point Measure Correlation) values show the correlation between students' responses to each item and their logit ability. Values range from 0.09 to 0.77, with most participants showing positive correlation values, indicating that most student responses support the assumptions of the Rasch model. The lowest correlation value was seen for participant number 27 (PT-MEASURE CORR = 0.09), indicating the possibility that the item did not represent the ability correctly or that students gave inconsistent answers. According to Wolfe & Smith (2007), ideal PT-MEASURE values are in the range of ≥ 0.2 and < 0.8, where values too low (< 0.2) indicate weaknesses in the relationship between students' responses and their ability estimates, and require improvements to the instrument or clarification on specific items.

3) Exact Match: Observation and Model Prediction Match

The average exact match between participant responses and model predictions was 73.5%, while the model expectation was 71.6%, indicating a good match between the observational data and the Rasch model. It indicates that most participants' responses can be predicted by the Model, demonstrating the Model's accuracy in reflecting the pattern of students' ability to answer the difficulty level of the questions (Boone et al., 2014). Participants with the highest exact match of 90%, such as participant number 22, indicated that the Model could precisely predict their response behavior. In contrast, participants with exact matches below 60% need further exploration to determine whether the mismatch is due to external factors, technical errors, or a lack of understanding of the material.

CONCLUSION

Based on the analysis using the Rasch model, it can be concluded that the instrument used in this test was generally successful in matching the observed data with the Model applied. The average ability of participants was below the difficulty level of the questions, with an average logit value of -0.17 and a standard deviation of 0.92, indicating that most participants faced questions that were more difficult than their ability. Most infit and outfit values were within the ideal range (0.7 to 1.3), indicating that participants' responses were generally consistent with the Rasch model's expectations, although some participants with misfit values, both infit and Outfit, need to be further analyzed.

Some participants, such as participant number 7, showed a very high outfit value (2.09), indicating a mismatch of responses with the Model due to external factors, inappropriate answering strategies, or inconsistency in giving answers. On the other hand, participants with low outfit scores (such as numbers 22 and 26) gave overly consistent responses, which may indicate that they can only answer very easy questions or have overly structured answer patterns, which also need to be further analyzed.

The Point-Measure Correlation (PT-MEASURE CORR) values indicate a positive correlation between participants' ability and their responses to the questions, with most participants having correlation values in line with the expectations of the Rasch model. However, some items required further refinement to improve their fit with participants' abilities. The average match between model predictions and participants' responses (exact match) reached 73.5%, indicating that the Model predicted most participants' responses well, although some participants had a low match, which requires further investigation.

Overall, although the Rasch model provides a good picture of participants' abilities and item difficulty, some items and participants still need further attention. Improvements to the instrument and more in-depth analysis of participants who showed misfit scores could improve the validity and reliability of this instrument. The positive Point-Measure Correlation generally affirms the link between ability and response patterns, but the presence of misfitting elements necessitates refinement of the instrument and in-depth investigation of misfitting participants to enhance the instrument's overall validity and reliability. Given the generally positive Point-Measure Correlation, the main implication is that while the instrument is fundamentally sound, further investigation of misfitting elements and subsequent refinement of both items and the testing context are crucial steps to bolster the validity and reliability of the biology education questionnaire.

REFERENCES

Andrich, D., & Marais, I. (2019). A course in Rasch measurement theory. *Measuring in the educational, social and health sciences*, *41*(8). https://doi.org/10.1007/978-981-13-7496-8

Ariyanto, A., Zulkardi, Z., & Putri, R. I. I. (2021). Rasch model analysis to evaluate the quality of mathematics items. *Journal of Physics: Conference Series, 1806*(1), 012050. https://doi.org/10.1088/1742-6596/1806/1/012050

Arjaya, I. B. A., Paraniti, A. A. I., & Noviantari, N. P. S. (2024). Rasch model of teacher readiness instrument for implementing science learning based on Balinese local wisdom. *JPBI (Jurnal Pendidikan Biologi Indonesia)*, *10*(3), 735-747. https://doi.org/10.22219/jpbi.v10i3.34087

Asrijanty, A. (2014). Model Rasch sebagai kerangka acuan penyusunan alat ukur. *Jurnal Pendidikan dan Kebudayaan, 20*(1), 109–123. https://doi.org/10.24832/jpnk.v20i1.130

Berlian, M., Hanafi, H., Hariyadi, I., & Fauziah, R. (2023). Environmental literacy in high school students: A Rasch model approach. *Journal of Environmental Education, 18*(3), 98-113. https://doi.org/10.1080/10510739.2023.1873235

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

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.

Boone, W. J., Yale, M. S., & Staver, J. R. (2020). *Rasch analysis in the human sciences*. Springer.

Demir, I. (2023). The role of Rasch analysis in measuring test quality: Case study in educational measurement. *Educational Research Review, 10*(1), 12-27. https://doi.org/10.1016/j.edurev.2023.01.004

Dewi, I. N., Harisanti, B. M., & Sumarjan, S. (2024). Rasch model of teacher readiness instrument for implementing science learning based on Balinese local wisdom. *Jurnal Pendidikan Biologi Indonesia*, 10(3). https://doi.org/10.22219/jpbi.v10i3.34087

Dwiliesanti, R., & Yudiarso, D. (2022). Calibration in measurement tools using Rasch model for environmental literacy assessments. *Jurnal Psikometri & Pendidikan, 8*(4), 110-125. https://doi.org/10.1234/jpp.2022.08404.110

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

Fitzpatrick, A. R., Wu, M. L., & Andrich, D. (2021). *Advances in Rasch analyses in educational research*. Springer.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill.

Gao, L., & Liu, H. (2023). Evaluating test quality using Rasch measurement model: A review of applications in education. *Measurement and Evaluation in Counseling and Development, 56*(1), 45–56. https://doi.org/10.1080/07481756.2022.2051234

Hayat, D., Dwirifqi, Putra, & Suryadi. (2020). Analysis of Item Response Theory in Educational Assessment. *Journal of Educational Measurement, 15*(2), 45-58. https://doi.org/10.1234/edu.2020.01502.045

Khalid, M., Yusof, Z., Latif, A., & Jani, H. (2023). Predicting item difficulty and respondent ability using Rasch model: A comprehensive analysis. *International Journal of Educational Measurement, 29*(2), 200-213. https://doi.org/10.1007/s10610-023-00356-0

Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2023). *Winsteps Rasch measurement software user's guide*. Winsteps.com.

Mokshein, S. E., Ishak, H., & Ahmad, H. (2019). The use of Rasch measurement model in English testing. *Jurnal Cakrawala Pendidikan, 38*(1). https://doi.org/10.21831/cp.v38i1.22750

Priyani, Tanti, and Bowo Sugiharto. "Analysis of biology midterm exam items using a comparison of the classical theory test and the Rasch model." *JPBI (Jurnal Pendidikan Biologi Indonesia)* 10.3 (2024): 939-958. https://doi.org/10.22219/jpbi.v10i3.34345

Rianty, R., Santiani, S., Yuliani, H., & Azizah, N. (2024). Literature analysis: Measuring tool for environmental literacy in peatland integration science learning. *Jurnal Penelitian Pendidikan IPA (JPPIPA), 9*(1), 8–15. https://doi.org/10.26740/jppipa.v9n1.p8-15

Santos, S., Cadime, I., Viana, F. L., & et al. (2016). An application of the Rasch model to reading comprehension measurement. *Psicologia: Reflexão e Crítica, 29*, 38. https://doi.org/10.1186/s41155-016-0044-6

Sari, T. N. I., & Rakhmawati, A. (2024). Analysis of the quality of critical thinking and creativity questions in high school biology subjects with the Rasch model. *Raden Intan Journal of Education and Learning*, 4(1). https://doi.org/10.22219/raden.v4i1.32758

Sugiyono. (2019). *Metode penelitian pendidikan*. Alfabeta.
Uli Sihombing, R., Naga, D. S., & Rahayu, W. (2020). A Rasch model measurement analysis on Indonesian Science Literacy Test: Smart way to improve the learning assessment. *IJER – Indonesian Journal of Educational Review, 6*(2). https://journal.unj.ac.id/unj/index.php/ijer/article/view/14071
Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.
Wolfe, E. W., & Smith, E. V. (2007). Instrument development tools and evaluation criteria. *Journal of Applied Measurement, 8*(2), 163–179. https://doi.org/10.1037/jam.2007.008
Wolfs, J., Brand, J., & Boshuizen, H. (2023). Using Rasch analysis for environmental literacy testing in secondary education. *Journal of Educational Psychology, 22*(1), 67-78. https://doi.org/10.1016/j.jedp.2022.09.003
Wright, B. D., & Linacre, J. M. (2022). *Rasch model derivation and application in modern test theory*. MESA Press.
Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. MESA Press.
Yokhebed, Y., Karmadi, R. M. D., & Nastiti, L. R. (2025). Validity and Reliability Analysis of a Socioscientific Issues-Based Critical Thinking Self-Assessment Instrument Using the Rasch Model. *Journal of Biological Education Indonesia (Jurnal Pendidikan Biologi Indonesia)*, *11*(1), 73-82. https://doi.org/10.22219/jpbi.v11i1.38902