RASCH model analysis: validity and reliability of concept understanding test instrument on temperature and heat

Kafa Pramitha Anggraini Indhira Artanti[1]*, Sherly Eka Putri Damayanti[1], Shinta Nuriyah Mahbubiyah Royani[1], Bintan Nuril Irvaniyah[1], Natasya Mustika Elvina Rossa[2], Anggita Meilina Putri[1], Khoirun Nisa[3]

[1]Department of Physics, Universitas Negeri Malang, Cakrawala Street No.5, Malang, 65145, Indonesia,

[2]Junior High School Antartika, Banyu Urip Kidul Street No.37 Surabaya, 60254, Indonesia,

[3]Department of Education and Human Potentials Development, National Dong Hwa University, Hualien, Taiwan.

*Corresponding author, email: kafapramitha13@gmail.com

Article History
Received: 7 November 2025
Revised: 19 May 2026
Accepted: 25 June 2026

Keywords
Conceptual understanding
Instrument validation
RASCH model
Temperature and heat

Abstract
This study aims to analyze the quality of a concept understanding test instrument on temperature and heat using the Rasch Model approach, focusing on validity, reliability, and item difficulty level. This research employed a descriptive method with a quantitative approach. The sample was selected through simple random sampling and involved 75 students of class XI at Senior High School. The research instrument consisted of 15 multiple-choice items covering concepts related to temperature and heat. Data were analyzed using the Rasch Model assisted by Ministep software to examine item fit, person reliability, item reliability, and the distribution of item difficulty. The results showed that all 15 items were valid based on the INFIT and OUTFIT criteria of MNSQ and ZSTD values. The person reliability was categorized as poor, whereas the item reliability was categorized as very good. The item difficulty level was distributed into four categories, consisting of 4 easy items, 2 moderate items, 6 difficult items, and 3 very difficult items. These findings indicate that the Rasch Model is effective in identifying the validity, reliability, and difficulty distribution of test items. The study implies that the developed instrument can be used by physics teachers and researchers as an empirical basis for evaluating students' conceptual understanding of temperature and heat. However, the low person reliability suggests the need for further refinement by involving a more heterogeneous sample and developing additional items with a broader range of difficulty levels to improve the instrument's ability to distinguish students' conceptual understanding.

Artanti, K. P. A. I., Damayanti, S. E. P., Royani, S. N. M., Irvaniyah, B. N., Rossa, N. M. E., Putri, A. M., & Nisa', K. (2026). RASCH model analysis: validity and reliability of concept understanding test instrument on temperature and heat. *Momentum: Physics Education Journal*, *10*(1), 85-93. https://doi.org/10.21067/mpej.v10i1.13156

1. Introduction

In education, measuring students' understanding of physics concepts is an important aspect in evaluating their level of understanding of the subject matter (Lichtenberger et al, 2024; Putranta & Supahar, 2019). The success of physics learning is highly dependent on students' ability to understand the concepts (Royani et al, 2025). As mentioned by Dewi and Ibrahim (2019), understanding concepts is a key factor in achieving overall success in science learning. Concept understanding plays a crucial role in helping students solve problems related to the concepts they have learned (Ncube & Luneta, 202). Physics itself has a close relationship with concepts and their application in everyday life. Therefore, the ability to understand concepts is crucial before students can apply the theories they have previously learned (Royani et al, 2025; Shidik, 2020).

Concept understanding is one of the graduation requirements in physics learning at the senior high schools (SMA), Islamic senior high schools (MA), and vocational high schools (SMK) level. Students are expected to have factual, conceptual, procedural, and metacognitive knowledge related to science (Antonia and Prudente, 2022; Dicipulo and Bautista, 2022). In addition to the knowledge aspect, concept understanding includes cognitive dimensions that include remembering (C1), understanding (C2), applying (C3), analyzing (C4), evaluating (C5), and creating (C6) (Setyaningrum & Fadli, 2023) . Students concept understanding ability in Indonesia is still relatively low. Research conducted by Gourlay using a concept mapping experiment reveals that the majority of students

doi: 10.21067/mpej.v10i1.13156

ISSN: 2548-9135

experience misconceptions in learning physics, which are attributed to their inadequate understanding of key concepts (Gourlay, 2017).

Errors in concept interpretation often occur because students are still in the stage of understanding. Although the error may have become a pattern in students' understanding, it does not fully reflect the correct scientific concept. It is important for teachers to quickly identify errors in students' interpretation of science concepts (Muna, 2017). One strategy to help students overcome these errors and assess their concept understanding is to identify the problem, find the cause, and develop appropriate strategies. Teachers have the alternative to use isomorphic test instruments as a tool to evaluate students' concept understanding of a particular material. The process of preparing such test instruments involves creating a series of questions that have similar indicators.

By using effective test instruments, educators can more accurately and thoroughly assess students' levels of understanding. However, the quality of the test instrument determines the final result of learning in the form of students' understanding of the formulated learning objectives (Yusuf & Widyaningsih, 2018). Therefore, it is necessary to analyze the test instrument items to determine whether the instrument is of good quality or not. The process of analyzing test instrument items is a crucial step in evaluating and improving the quality of each question before it is used in the test. In this process, information from learners' answers is collected to assess the quality of test items (Surahman & Wang, 2022). Item analysis can be done qualitatively to evaluate its content and form, and quantitatively to assess its statistical features. Quantitative aspects include validity, reliability, difficulty, distinguishing power, and question difficulty level (Putri & Sayono, 2026).

One of the approaches used to analyze and improve the quality of test instruments in this study is Rasch Analysis. Rasch Analysis is a powerful statistical method for measuring the characteristics of test instruments, including validity, reliability, and item discrimination. In addition, Rasch modeling is an effective method to evaluate validity and reliability directly by utilizing the principle of probability (Hamzah et al., 2022). By using Rasch Analysis, we can gain a deeper understanding of the performance of test instrument items in measuring the understanding of physics concepts. The advantages of the Rasch model include its ability to predict missing data based on systematic response patterns, produce measurement standard errors to improve the accuracy of calculations, and calibrate simultaneously on the measurement scale, respondents, and items (Sumintono & Widhiarso, 2015). This advantage is particularly relevant in this study, as it enables the analysis of items from physics concept understanding test instruments for students. Therefore, this study aims to analyze the items of physics concept understanding test instruments using the Rasch Analysis approach.

Despite the growing number of studies on physics concept understanding assessment, most previous research has primarily focused on general cognitive achievement tests or misconception identification without specifically developing and validating instruments for temperature and heat concepts using the Rasch model approach. Several studies have examined concept understanding instruments in physics; however, many of them still rely on Classical Test Theory (CTT), which has limitations in estimating item characteristics and respondent ability simultaneously. In contrast, Rasch Analysis provides more comprehensive information regarding item validity, reliability, item difficulty, and person ability on the same measurement scale, allowing for more accurate instrument calibration and interpretation. Furthermore, temperature and heat remain among the most challenging topics in physics learning due to the abstract nature of the concepts and the high prevalence of misconceptions among students. Previous studies also indicate that students frequently experience difficulties distinguishing between heat and temperature, understanding thermal equilibrium, and interpreting heat transfer phenomena. Nevertheless, validated diagnostic instruments specifically designed to measure conceptual understanding on this topic at the senior high school level are still limited. Therefore, this study addresses an important gap in the literature by developing and analyzing a conceptual understanding test instrument for temperature and heat using the Rasch model framework. The novelty of this study lies in the integration of Rasch-based psychometric analysis with the development of an instrument specifically targeted at senior high school students' conceptual understanding of temperature and heat concepts. This study is expected to contribute theoretically to physics education assessment research and practically to provide educators with a valid and reliable instrument for evaluating students' conceptual understanding more accurately and comprehensively.

Based on this explanation, the problem to be investigated in this study is how to analyze the validity, reliability, and difficulty level of instrument test. The test was designed for temperature and heat material for physics subject in senior high school level. Futhermore, this study is expected to make a significant contribution to improving the quality of designed-instrument test to masure students' conceptual understanding of physics concepts.

2. Method

This study employed a descriptive quantitative research design. The population consisted of eleventh-grade science students (XI MIPA) at Senior High School in Bendosari. The sample involved 75 students selected through a simple random sampling technique. The instrument analysis in this study used the Rasch Model through the Ministep application. The Rasch Model was intentionally selected because it provides several methodological advantages compared to Classical Test Theory (CTT), particularly in educational measurement studies with relatively limited sample sizes. Unlike CTT, which is highly dependent on sample characteristics, the Rasch model is capable of estimating person ability and item difficulty independently on the same logit scale, thereby producing more stable and objective measurements (Sumintono & Widhiarso, 2015). In addition, Rasch analysis remains appropriate for moderate sample sizes because it focuses on probabilistic modeling of response patterns rather than solely relying on large-sample assumptions.

The use of the Rasch model is also particularly relevant considering the relatively homogeneous characteristics of students within the same academic level and school environment. In homogeneous datasets, conventional analysis methods often encounter limitations in distinguishing item functioning and respondent ability accurately. Rasch analysis addresses this issue by evaluating item fit, person fit, item difficulty distribution, and response consistency simultaneously, allowing subtle differences in conceptual understanding to be identified more precisely. Furthermore, the Rasch model enables the detection of misfitting items and unexpected response patterns that may not be observable through conventional validity and reliability analysis. Therefore, the Rasch approach was considered suitable for evaluating the quality of the conceptual understanding test instrument on temperature and heat topics. The criteria used to assess item fit (validity) in the Rasch Model referred to the standards proposed (Rahmayani et al, 2022). Table 1 shown a criteria for question validity.

Table 1. Criteria for Question Validity

MNSQ Outfit Infit Value	ZSTD Infit Outfit Value	Pt Value Mean Corr	Description
> 1.3	$> +2.0$	> 0.85	Misfit
$0.7 - 1.3$	$-2.0\ s.d. +2.0$	$0.4 - 0.85$	Fit
< 0.7	< -2.0	< 0.4	Fit

Cronbach's alpha formula is the formula used in determining the reliability of a question in the Rasch Model theory using (KR-20), which is based on the following view in Table 2 (Sumintono & Widhiarso, 2015):

Table 2. Criteria for the Reliability Coefficient

Value Criteria	Description
> 0.8	Very good
0.70 – 0.80	Good
0.60 – 0.70	Simply
0.50 – 0.60	Enough
< 0.50	Bad

The next stage in this study is to identify the difficulty index of a question item according to Lestari & Yudhanegara (2017) as follows Table 3:

Table 3. Criteria for the Reliability Coefficient

Value Criteria	Description
$Measure\ logit > SD\ Logit$	Very difficult item
$0 \leq Measure\ logit \leq SD\ Logit$	Difficult item
$SD\ Logit \leq Measure\ logit \leq 0$	Medium item
$Measure\ logit < -SD\ Logit$	Easy item

The research instrument utilized a test instrument to understand the concepts of temperature and heat, comprising multiple-choice questions with 14 indicators and 15 items.

3. Results and Discussion

This section provides a comprehensive explanation of three essential aspects of the instrument analysis, namely: (1) the validity of the test results, which discusses the extent to which the test items accurately measure students' conceptual understanding of temperature and heat; (2) the reliability of the test, which examines the consistency and stability of students' responses across items and measurement occasions; and (3) the item analysis, which includes the examination of problem difficulty levels to identify whether each test item appropriately discriminates between students with high and low levels of understanding. These three analyses together ensure that the developed instrument is both psychometrically sound and suitable for assessing conceptual comprehension in the topic of temperature and heat.

3.1. Validity Test

The results of the validity test for the instrument items, assessing understanding of the concepts of temperature and heat, have been analyzed using the Rasch model. This analysis aims to evaluate the extent to which the instrument used meets the validity criteria, which include measuring the accuracy and suitability of the items using the Rasch model (Alnahdi et al, 2025; Papini et al., 2021). Thus, this analysis not only assesses the reliability of measurements statistically, but also reviews the logical consistency between items based on modern measurement theory.

Previous studies have also shown that the application of the Rasch Model is effective in assessing the validity of instruments related to understanding science concept. For example, research by Akbari (2025) confirmed that the Rasch Model can provide more in-depth information related to item quality, such as item difficulty and person ability, compared to classical analysis. In addition, Lestari and Samsudin (2020) and Safitri et al. (2025) reported that Rasch analysis provides a more accurate picture of the structure of students' conceptual understanding in the topics of temperature and heat, as it is able to identify items that do not match the expected response patterns. With reference to these findings, the use of the Rasch Model in this study is relevant to ensure that the instruments used are capable of measuring the understanding of temperature and heat concept in a valid, objective, and consistent manner across the entire range of respondents' abilities. The table below shows the validity analysis results obtained through the Rasch Model approach. Table 4 shown a validity analysis results.

Table 4. Validity Analysis Results

No Item	Infit		Outfit		PT Measure Corr	Description
	MNSQ	ZSTD	MNSQ	ZSTD		
14	0.92	-0.14	0.99	0.14	0.54	Valid
9	1.11	0.51	1.19	0.62	0.39	Valid
11	1.05	0.27	1.24	0.73	0.40	Valid
15	1.00	0.08	1.02	0.17	0.45	Valid
6	1.03	0.20	0.98	0.02	0.43	Valid
2	1.12	0.80	1.13	0.64	0.32	Valid
3	0.79	-1.43	0.70	-1.64	0.58	Valid
8	1.12	0.99	1.21	1.29	0.28	Valid
13	0.99	-0.02	1.00	0.07	0.38	Valid
5	0.96	-0.56	0.95	-0.51	0.36	Valid
1	1.07	0,91	1.10	0.90	0.26	Valid
12	1.01	0.08	0.94	-0.32	0.28	Valid
7	0.91	-0.68	0.88	-0.60	0.33	Valid
4	0.89	-0.57	0.80	-0.64	0.31	Valid
10	1.01	0.16	0.75	-0.24	0.15	Valid

Based on the results of the analysis using the Rasch Model, all items measuring understanding of the concepts of temperature and heat were declared valid. The Infit MNSQ values ranged from 0.7 to 1.21, and the Outfit MNSQ values ranged from 0.7 to 1.3, which are still within the recommended tolerance limits of 0.5 ≤ MNSQ ≤ 1.5 (Bond & Fox, 2015). This indicates that each item has a good level of fit with the model and consistently represent the respondents' abilities. The ZSTD (Z-standardized) values for most items within the range of -2 to +2, indicating that no items exhibited extreme response patterns or significant deviations from the model predictions. In addition, the PT

Measure Correlation values ranging from 0.15 to 0.58, indicate a positive correlation between item scores and participant ability, suggesting that all items contribute to measuring the expected construct: understanding the concepts of temperature and heat.

These findings align with Muller's (2020) research, which demonstrates that the Rasch Model can validate item quality through a combination of fit statistics and point measure correlation analysis, assessing the logical consistency between items. Similar results were also found by Hermita et al (2021), who used the Rasch Model to analyze the transfer concepts of temperature and heat, and showed that items with fit values within the ideal range reflect a valid instrument suitable for measuring students' conceptual understanding. Thus, the results of this analysis indicate that all items on the instrument meet the validity criteria based on the Rasch Model. This instrument can be said to have a good ability to objectively, consistently, and proportionally measure the conceptual understanding of temperature and heat in relation to the respondents' ability levels.

3.2. Reliability Test

The results of the Rasch Model analysis showed a Cronbach's alpha value of 0.82, which falls into the very good category, indicating strong internal consistency between respondents and the test items. In addition, the item reliability value reached 0.96, categorized as excellent, while the person reliability value was 0.58, categorized as poor. These findings indicate contrasting measurement characteristics between the quality of the items and the distribution of students' abilities. The very high item reliability demonstrates that the developed items possess stable difficulty hierarchies and consistent measurement properties. In other words, the instrument was capable of consistently distinguishing item difficulty levels, ranging from easy to very difficult items, with relatively small measurement error. This condition suggests that the quality of the test items was well-constructed and sufficiently robust for measuring conceptual understanding of temperature and heat.

In contrast, the relatively low person reliability indicates that the instrument was less effective in differentiating students according to their ability levels. This phenomenon may occur because the participants tended to have relatively homogeneous conceptual understanding, causing their response patterns to appear similar across many items. In Rasch analysis, person reliability is highly influenced by the variability of respondent ability; when most students possess comparable achievement levels, the instrument will have difficulty separating respondents into distinct ability categories. Another contributing factor may be the limited spread of item difficulty relative to the students' ability distribution. Although the instrument included items categorized as easy, moderate, difficult, and very difficult, the concentration of participants within a similar ability range may have reduced the discriminatory power of the test.

This condition has also been reported by Yang et al. (2023) and Christensen et al. (2024), who found that person reliability values are frequently lower than item reliability when participant ability distributions are homogeneous or when the sample size is relatively limited. Furthermore, Tennant and Küçükdeveci (2023) emphasized that reliability in the Rasch model should be interpreted differently from Cronbach's alpha. While Cronbach's alpha primarily measures internal consistency, person reliability reflects the instrument's capability to distinguish respondents based on ability variation. Therefore, a high Cronbach's alpha value does not necessarily indicate that the instrument can effectively separate participants into multiple ability strata.

The coexistence of exemplary item reliability and deficient person reliability in this study suggests that the primary limitation lies not in the quality of the instrument items, but rather in the characteristics of the respondent sample and the alignment between item difficulty and participant ability. These findings imply that future instrument development should consider expanding the range of item difficulty levels and involving more heterogeneous participant groups to improve the instrument's discriminating power. As suggested by L'Boy (2023), broader variability in item difficulty and respondent characteristics can enhance person reliability and provide a more accurate representation of students' conceptual understanding levels. Therefore, the interpretation of reliability results in Rasch analysis should simultaneously consider Cronbach's alpha, person reliability, and item reliability in order to obtain a more comprehensive evaluation of instrument quality.

3.3. Problem Item Analysis (Problem Difficulty Level)

The level of item difficulty reflects the likelihood that respondents will answer an item correctly (Arikunto, 2017). An item is categorized as difficult if only a few students answer it correctly, and vice versa, an item is easy if many students answer it correctly. According to Subando et al. (2025), a balanced level of difficulty is crucial to ensure that the test can fairly distinguish between high- and low-ability students. The table below shows the results of the analysis of the level of difficulty through the Rasch Model approach.

Table 5. Results of Problem Difficulty Analysis

Item	Measure Value (logit)	Category
S1	-0.81	Medium
S2	0.81	Difficult
S3	0.72	Difficult
S4	-2.21	Easy
S5	-0.62	Medium
S6	1.29	Difficult
S7	-1.61	Easy
S8	0.41	Difficult
S9	1.66	Very Difficult
S10	-3.59	Easy
S11	1.66	Very Difficult
S12	-1.46	Easy
S13	0.07	Difficult
S14	2.15	Very Difficult
S15	1.53	Difficult
SD logit = 1.61		

The results obtained after analyzing the items using the Rasch Model approach show that most item difficulty levels can be observed from the measure value in logit units for each item. Based on the criteria presented in Table 3, the measure value (logit) of items is distributed in four groups of difficulty levels, namely easy (4 questions), moderate (2 questions), difficult (6 questions), and very difficult (3 questions), with an almost even distribution (Jumini et al., 2023). In the difficulty level group, with easy criteria consisted of questions number 4, 7, 10, and 12; the medium level criteria consisted of questions number 1 and 5; difficult level criteria consisted of numbers 2, 3, 6, 8, 13, and 15; finally for the very difficult level consisted of questions number 9, 11, and 14.

These results align with the research conducted by Sumintono and Widhiarso (2015), which demonstrates that the use of the Rasch model can effectively identify the distribution of item difficulty. Their research also revealed that the Rasch model provides more detailed insight into the distribution of difficulty for each item in the test, thus helping to prepare a more balanced and representative test. Metsämuuronen (2023) explains that estimating difficulty levels in logit units in the Rasch Model allows researchers to identify biases and adjust parameters for more accurate measurement results. Yulianto and Widodo (2023) also found that categorizing items into easy, moderate, difficult, and very difficult using logit ±SD thresholds effectively balances the range of item difficulty. In addition, Isnawati et al. (2023) emphasized that item logit mapping is important to ensure that the proportion of questions covers the entire range of participants' abilities, while Çebi et al. (2024), in a study on digital literacy scales, highlighted that an even logit distribution indicates construct validity and the instrument's ability to measure competence at various levels. Thus, the distribution of item difficulty in this study can be considered relatively balanced and indicates that the instrument has been able to measure students' abilities across different levels of difficulty.

The findings of this study have important implications for both instructional design and assessment development in physics education, particularly in the topic of temperature and heat. The Rasch analysis revealed that the instrument possesses strong item validity and high item reliability, indicating that the developed items are capable of consistently measuring students' conceptual understanding across different levels of ability. More importantly, the variation in item difficulty levels provides meaningful insights into students' reasoning patterns and conceptual challenges related to thermal physics concepts.

The existence of several difficult and very difficult items suggests that students still encounter substantial difficulties in understanding abstract concepts such as heat transfer mechanisms, thermal equilibrium, and the distinction between heat and temperature. This finding supports previous studies reporting that misconceptions in thermal physics remain prevalent among secondary school students. Therefore, the results imply that future instructional practices should not focus solely on procedural problem-solving, but should emphasize conceptual reasoning and scientific interpretation. Learning approaches such as inquiry-based learning, conceptual change strategies, interactive simulations, and contextual laboratory activities may help students develop deeper conceptual understanding and reduce misconceptions related to temperature and heat.

Furthermore, the distribution of item difficulty identified through the Rasch model provides valuable information for the development of more adaptive and balanced assessment frameworks. Easy items may serve as indicators of students' foundational conceptual understanding, while difficult and very difficult items can be used to assess higher-order reasoning skills and students' ability to apply concepts in unfamiliar situations. Consequently, the developed instrument can function not only as an evaluative assessment tool, but also as a diagnostic instrument capable of identifying specific conceptual weaknesses and learning difficulties among students.

From an assessment perspective, the Rasch-based analysis offers richer psychometric information compared to conventional methods. The ability of the Rasch model to simultaneously map person ability and item difficulty on the same logit scale enables educators to determine whether the assessment instrument appropriately corresponds to students' ability distributions. Such information is beneficial for designing differentiated instruction and adaptive assessments that are aligned with students' cognitive readiness and learning progression.

In addition, the relatively low person reliability obtained in this study indicates that students' conceptual understanding levels were relatively homogeneous. This finding suggests the need for future instructional designs that encourage greater variation in conceptual engagement, critical thinking, and problem-solving experiences. Future studies may also improve the instrument by incorporating a wider range of item difficulties, involving larger and more heterogeneous samples, and integrating reasoning-based diagnostic questions to further explore students' scientific thinking processes in physics learning.

Overall, this study contributes not only to the development of a valid and reliable instrument for measuring conceptual understanding of temperature and heat, but also to the broader advancement of evidence-based assessment practices in physics education. The findings are expected to support educators in designing more effective instructional strategies and more comprehensive conceptual assessment frameworks for physics learning.

4. Conclusion

This study aims to analyze the validity and reliability of concept understanding test instruments for temperature and heat materials using the Rasch Model, with the analysis conducted through the Winsteps application. The results showed that all 15 items were declared valid based on INFIT and OUTFIT criteria using the MNSQ and ZSTD. For reliability, Person Reliability was categorized as poor, while Item Reliability was very good. The level of difficulty of the questions was evenly distributed, with 4 easy questions, 2 medium questions, 6 difficult questions, and 3 very difficult questions, indicating that the Rasch Model was effective in identifying the distribution of question difficulty.

Author Contributions

Kafa Pramitha Anggraini Indhira Artanti: Conceptualization, Methodology, Writing original draft preparation. Sherly Eka Putri Damayanti: Data curation, Formal analysis, Visualization. Shinta Nuriyah Mahbubiyah Royani: Conceptualization, Writing review & editing. Bintan Nuril Irvaniyah: Investigation, Project administration, Resources. Natasya Mustika Elvina Rossa: Software, Data curation, Visualization. Anggita Meilina Putri: Methodology, Formal analysis, Writing original draft preparation. Khoirun Nisa: Conceptualization, Writing review & editing.

Funding

No funding support was received.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Akbari, A. (2025). The Rasch analysis of item response theory: An untouched area in evaluating student academic translations. *SKASE Journal of Translation and Interpretation, 18*(1), 50–77.

Alnahdi, A. H., Alsubiheen, A. M., & Aldaihan, M. M. (2025). Rasch measurement model supports the unidimensionality and internal structure of the Arabic Oswestry Disability Index. *Journal of Clinical Medicine, 14*(4), 1259. https://doi.org/10.3390/jcm14041259

Antonio, R. P., & Prudente, M. S. (2022). Effectiveness of metacognitive instruction on students' science learning achievement: A meta-analysis. *International Journal on Studies in Education, 4*(1), 1–20.

Arikunto, S. (2017). *Pengembangan instrumen penelitian dan penilaian program*. Pustaka Pelajar.

Avinç, E., & Doğan, F. (2024). Digital literacy scale: Validity and reliability study with the Rasch model. *Education and Information Technologies, 29*(17), 22895–22941. https://doi.org/10.1007/s10639-024-12657-2

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Christensen, K. S., & Ammentorp, J. (2024). Rasch analysis of the self-efficacy (SE-12) questionnaire measuring clinical communication skills. *PEC Innovation, 4*, 100296. https://doi.org/10.1016/j.pecinn.2024.100296

Dewi, S. Z., & Ibrahim, T. (2019). Pentingnya pemahaman konsep untuk mengatasi miskonsepsi dalam materi belajar IPA di sekolah dasar. *Jurnal Pendidikan Universitas Garut, 13*(1), 130–136.

Discipulo, L. G., & Bautista, R. G. (2022). Students' cognitive and metacognitive learning strategies towards hands-on science. *International Journal of Evaluation and Research in Education, 11*(2), 882–891. https://doi.org/10.11591/ijere.v11i2.21919

Hamzah, H. (2017). Learning about A-level physics students' understanding of particle physics using concept mapping. *Physics Education, 52*(5), Article 054002. https://doi.org/10.1088/1361-6552/aa7f3c

Hamzah, F. M., Rashid, M. N. A., Rahman, M. N. A., & Rasul, M. S. (2022). Evaluating the validity and reliability of authentic learning instruments using Rasch model. *International Journal of Global Optimization and Its Application, 1*(3), 182–189. https://doi.org/10.56225/ijgoia.v1i3.24

Handayani, S. (2022). Analisis UAS biologi kelas X dengan teori tes klasik dan item response theory (Rasch model). *Bio-Pedagogi, 11*(2), 76–84. https://doi.org/10.20961/bio-pedagogi.v11i2.63177

Hermita, N., Sakinah, S., Wijaya, T. T., Vebrianto, R., Alim, J. A., Putra, Z. H., Yulianti, E., & Jihe, C. (2021). Item analysis of heat transfer concept using Rasch model in elementary school. *Journal of Physics: Conference Series, 2049*(1), Article 012058. https://doi.org/10.1088/1742-6596/2049/1/012058

Isnawati, I., Sriyati, S., Agustin, R. R., Supriyadi, S., Kasi, Y. F., & Ismail, I. (2024). Analysis of question difficulty levels based on science process skills indicators using the Rasch model. *Tadris: Jurnal Keguruan dan Ilmu Tarbiyah, 9*(1), 31–41. https://doi.org/10.24042/tadris.v9i1.19374

Kementerian Pendidikan dan Kebudayaan Republik Indonesia. (2016). *Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 20 Tahun 2016 tentang Standar Kompetensi Lulusan Pendidikan Dasar dan Menengah*. Kemendikbud.

L'Boy, D., & Khan, R. N. (2023). A Rasch-model-based hierarchical framework for statistical literacy and learning. *International Journal of Mathematical Education in Science and Technology, 54*(9), 1874–1887. https://doi.org/10.1080/0020739X.2022.2060166

Lestari, A. S., & Samsudin, A. (2020). Using Rasch model analysis to analyze students' scientific literacy on heat and temperature. In *Proceedings of the 7th Mathematics, Science, and Computer Science Education International Seminar* (pp. xx–xx).

Lestari, K. E., & Yudhanegara, M. R. (2017). Analisis kemampuan representasi matematis mahasiswa pada mata kuliah geometri transformasi berdasarkan latar belakang pendidikan menengah. *Jurnal Matematika Integratif, 13*(1), 29–39. https://doi.org/10.24198/jmi.v13i1.11410

Lichtenberger, A., Hofer, S. I., Stern, E., & Vaterlaus, A. (2025). Enhanced conceptual understanding through formative assessment: Results of a randomized controlled intervention study in physics classes. *Educational Assessment, Evaluation and Accountability, 37*(1), 5–33. https://doi.org/10.1007/s11092-024-09458-5

Metsämuuronen, J. (2023). Seeking the real item difficulty: Bias-corrected item difficulty and some consequences in Rasch and IRT modeling. *Behaviormetrika, 50*(1), 121–154. https://doi.org/10.1007/s41237-022-00192-7

Müller, M. (2020). Item fit statistics for Rasch analysis: Can we trust them? *Journal of Statistical Distributions and Applications, 7*(1), Article 5. https://doi.org/10.1186/s40488-020-00108-7

Muna, I. A. (2017). Model pembelajaran POE (*predict–observe–explain*) dalam meningkatkan pemahaman konsep dan keterampilan proses IPA. *El-Wasathiya: Jurnal Studi Agama, 5*(1), 73–92.

Hamzah, F. M., Rashid, M. N. A., Rahman, M. N. A., & Rasul, M. S. (2022). Evaluating the validity and reliability of authentic learning instruments using Rasch model. *International Journal of Global Optimization and Its Application, 1*(3), 182–189. https://doi.org/10.56225/ijgoia.v1i3.24

Papini, N., Kang, M., Ryu, S., Griese, E., Wingert, T., & Herrmann, S. (2021). Rasch calibration of the 25-item Connor–Davidson Resilience Scale. *Journal of Health Psychology, 26*(11), 1976–1987. https://doi.org/10.1177/1359105320909876

Putranta, H., & Supahar, S. (2019). Development of physics-tier tests (PysTT) to measure students' conceptual understanding and creative thinking skills: A qualitative synthesis. *Journal for the Education of Gifted Young Scientists, 7*(3), 747–775. https://doi.org/10.17478/jegys.587203

Putri, A. P., & Sayono, J. (2026). Evaluation of item quality: Analysis of difficulty level and distinction power with quantitative methods. *Journal of Educational Sciences, 10*(1), 317–330.

Rahmayani, A., Tiurlina, T., & Alfarisa, F. (2022). Analisis kualitas butir soal ulangan harian matematika di kelas IV MI Al-Islamiyah menggunakan Rasch model. *Jurnal Perseda, 5*(3), 170–177.

Royani, S. N. M., Sutopo, S., Hidayat, A., & Parno, P. (2025). Research trends in physics learning strategies: A systematic literature review addressing students' conceptual understanding difficulties in kinematics. *Jurnal Penelitian Pendidikan IPA, 11*(3), 1–10.

Royani, S., Artanti, K., Putri, R., & Parno, P. (2025). Analysis of item difficulty and student misconceptions on temperature and heat using a two-tier diagnostic test. *Jurnal Ilmiah Pendidikan Fisika, 9*(1), 144–154. https://doi.org/10.20527/jipf.v9i1.

Safitri, F., Rusdiana, D., Samsudin, A., & Widiyatmoko, A. (2025). Development of understanding test instruments for grade 7 junior high school students on temperature, heat and expansion topics. *Jurnal Penelitian Pendidikan IPA, 11*(1), 395–404. https://doi.org/10.29303/jppipa.v11i1.

Shidik, M. A. (2020). Hubungan antara motivasi belajar dengan pemahaman konsep fisika peserta didik MAN Baraka. *Jurnal Kumparan Fisika, 3*(2), 91–98. https://doi.org/10.33369/jkf.3.2.91-98

Setyaningrum, N., & Fadli, F. (2023). Cognitive development analysis of C4, C5, and C6 in early childhood education students and their implications for teacher efforts at Raudhatul Athfal. *Proceeding International Conference on Religion, Science, and Education, 2*, 487–496.

Subando, J., Astoko, D. B., Amin, L. H., Budiharjo, & Embong, R. (2025). A multidimensional item response theory approach in the item analysis of Arabic language tests in Madrasah Aliyah. *Jurnal Penelitian dan Evaluasi Pendidikan, 29*(2), 271–284.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan*. Trim Komunikata.

Suparno, P. (2005). *Miskonsepsi dan perubahan konsep dalam pendidikan fisika*. Grasindo.

Rismawati, T. A., & Pujiastuti, H. (2020). Pengaruh model Search Solve Create and Share (SSCS) terhadap kemampuan pemahaman konsep matematis. *Jurnal Kajian Pendidikan Matematika, 5*(2), 183–190. https://doi.org/10.30998/jkpm.v5i2.6332

Tennant, A., McKenna, S. P., & Hagell, P. (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health, 7*(Suppl. 1), S22–S26. https://doi.org/10.1111/j.1524-4733.2004.7s106.x

Yang, J., Reheman, Z., Liu, Y., & Zhao, S. (2023). Applying Rasch analysis in refinement and validation of interpersonal skills measure for gifted children. *Frontiers in Psychology, 14*, Article 1236640. https://doi.org/10.3389/fpsyg.2023.1236640

Yulianto, A., & Widodo, A. (2020). Disclosure of difficulty distribution of HOTS-based test questions through Rasch modeling. *Indonesian Journal of Primary Education, 4*(2), 197–203. https://doi.org/10.17509/ijpe.v4i2.26691

Yusuf, I., & Widyaningsih, S. W. (2018). Profil kemampuan mahasiswa dalam menyelesaikan soal HOTS di Jurusan Pendidikan Fisika Universitas Papua. *Jurnal Komunikasi Pendidikan, 2*(1), 42–49. https://doi.org/10.32585/jkp.v2i1.63