Diagnostic Assessment in Physics: Developing and Validating a Two-Tier Instrument on Heat Concepts Using the Rasch Model

Asesmen Diagnostik dalam Fisika: Pengembangan dan Validasi Instrumen Dua-Tingkat pada Konsep Kalor Menggunakan Model Rasch

Andi Mustari[1], Intan Wulan Sari[2], Abu Bakar[3], Septri Rahayu[4], Fatma Wati[5]
Akademi Komunitas Negeri Aceh Barat[1,2], Politeknik Aceh[3], Balai Diklat Keagamaan Ambon[4], Politeknik Negeri Padang[5]
andimustari@aknacehbarat.ac.id[1], intanwulansari@aknacehbarat.ac.id[2], abu@politeknikaceh.ac.id[3], septrirahayu@gmail.com[4], Fatma_wati@fmipa.unp.ac.id[5]
DOI: https://doi.org/10.52048/inovasi.v19i2.668

ABSTRACT

This study aims to develop and analyze a two-tier diagnostic test instrument designed to identify students' misconceptions regarding heat. The instrument includes 20 multiple-choice items (10 for content and 10 for justification) aimed at measuring students' understanding and reasoning. The data were collected from 57 preservice teachers and analyzed using the Rasch model. The results showed an item reliability value of 0.92 and an item separation index of 3.47, indicating a high level of internal consistency and the instrument's capacity to differentiate item difficulty levels. The person-item map showed a moderate match between student ability and item difficulty, but gaps were identified at the lower and upper ends of the ability distribution. Infit and Outfit statistics revealed three items-S8, A4, and S4-that did not fit the Rasch model and were considered ineffective in identifying students' misconceptions. These findings suggest that the Rasch model is a robust analytical tool for evaluating diagnostic test quality and support the need for well-calibrated items in order to accurately reveal students' conceptual understanding and misconceptions in physics learning.
Keywords: Diagnostic Test, Heat, Misconceptions, Rasch Model

ABSTRAK

Penelitian ini bertujuan untuk mengembangkan dan menganalisis instrumen tes diagnostik dua tingkat yang dirancang untuk mengidentifikasi miskonsepsi mahasiswa pada topik kalor. Instrumen ini terdiri atas 20 butir soal pilihan ganda (10 butir untuk konten dan 10 butir untuk alasan) yang mengukur pemahaman dan penalaran mahasiswa. Data diperoleh dari 57 mahasiswa calon guru dan dianalisis menggunakan model Rasch. Hasil penelitian menunjukkan reliabilitas butir sebesar 0,92 dan indeks pemisahan butir sebesar 3,47, yang menandakan konsistensi internal yang tinggi serta kemampuan instrumen dalam membedakan tingkat kesulitan butir. Peta *person–item* menunjukkan tingkat kesesuaian sedang antara kemampuan mahasiswa dan tingkat kesulitan butir, namun ditemukan kesenjangan pada bagian bawah dan atas distribusi kemampuan. Statistik *Infit* dan *Outfit* mengindikasikan adanya tiga butir yakni S8, A4, dan S4 yang tidak sesuai dengan model

Rasch dan dianggap kurang efektif dalam mengidentifikasi miskonsepsi mahasiswa. Temuan ini menegaskan bahwa model Rasch merupakan alat analisis yang andal untuk mengevaluasi kualitas tes diagnostik serta menekankan pentingnya butir yang terkalibrasi dengan baik agar dapat mengungkap pemahaman konseptual dan miskonsepsi mahasiswa dalam pembelajaran fisika secara akurat.
Kata Kunci: Miskonsepsi, Model Rasch, Kalor, Tes Diagnostik

INTRODUCTION

The concept of heat is a fundamental topic in Physics, both at the secondary and higher education levels (Cahyaningtyas et al., 2023; Ugwuanyi et al., 2023; Yuliana et al., 2023). A correct understanding of heat is essential, as it underpins various real-life phenomena, including phase changes, heat transfer, and other thermodynamic processes.

However, numerous studies have consistently revealed that misconceptions related to heat are highly prevalent (Cahyaningtyas et al., 2023; Yuliana et al., 2023), not only among students and university learners, but also among preservice teachers who are expected to teach this topic in the future (Aydeniz et al., 2017).

Misconceptions are defined as understandings or mental models that deviate from scientifically accepted concepts and are often resistant to change even after instruction (Mukhlisa, 2021). In physics education, misconceptions are particularly common because students attempt to explain physical phenomena using everyday reasoning rather than scientific principles. These misconceptions can originate from inappropriate analogies, intuitive experiences that conflict with formal theory, or instructional methods that fail to challenge prior conceptions (Boateng, 2024; Oladejo et al., 2023). For example, students may believe that heat is a material substance that flows from hot to cold objects, that temperature and heat represent the same physical quantity, or that a metal feels colder because it "contains less heat." Such alternative conceptions demonstrate how students' intuitive reasoning can obscure their understanding of energy transfer and thermal equilibrium (Hoppe et al., 2020; Irmak et al., 2023).

Case study results indicate that misconceptions are still frequently found among preservice teachers, a concern that becomes increasingly critical given their future role as educators. As future educators, preservice teachers play a pivotal role in conveying accurate scientific concepts to their students (Buchari, 2018). If these misconceptions are not addressed early, there is a risk that they will propagate incorrect understandings in the

classroom, ultimately affecting the quality of science education in the long term (Wangdi et al., 2017). Therefore, there is a pressing need for effective diagnostic tools to identify and address misconceptions among preservice teachers (Aydeniz et al., 2017). The two-tier diagnostic test is considered one of the most effective methods for detecting misconceptions. This type of test consists of two components: the first tier is a multiple-choice question that assesses students' factual understanding, while the second tier requires them to provide a reason or explanation for their chosen answer. This format allows for a deeper analysis of conceptual understanding, which cannot be captured through conventional multiple-choice tests alone (Duit & Treagust, 2003; Treagust, 1988).

In recent years, an increasing number of studies have emphasized the importance of using two-tier tests in science education. These tests not only help educators identify students' misconceptions but also provide valuable insights into how students or preservice teachers reason about specific scientific concepts (Liu et al., 2024). For instance, research by Hoppe et al (2020) and Irmak et al (2023) revealed that preservice teachers continue to hold significant misconceptions about the concept of heat, despite having studied the topic on multiple occasions.

The use of valid and reliable diagnostic instruments is essential in education, particularly for preservice teachers (Istiyono et al., 2023; Jumadi et al., 2023). Developing a trustworthy instrument enables teachers and lecturers to diagnose and address misconceptions before preservice teachers begin their professional careers. Validating and testing the reliability of such instruments are also crucial to ensure their effectiveness across various educational contexts (Laeli, 2020; Tsui & Treagust, 2010). One analytical approach that can be used to evaluate the quality of an instrument is the Rasch Model (Jumadi et al., 2023; Suparman et al., 2024). As part of Item Response Theory (IRT), the Rasch Model allows for in-depth analysis of item characteristics, including item difficulty, discrimination power, and the overall fit between the model and empirical data (Suparman et al., 2024). By applying the Rasch Model, the developed two-tier instrument can be quantitatively validated, resulting in a more accurate and dependable assessment tool.

Given the critical role of preservice teachers in shaping students' scientific understanding, the availability of a valid and reliable diagnostic instrument is essential. Before misconceptions spread in the classroom, a well-designed tool can help identify and

address them. Therefore, this study aims to develop and validate a two-tier diagnostic test focused on the concept of heat, specifically targeting misconceptions. The resulting instrument is expected to support the improvement of conceptual understanding in physics education and to minimize the risk of misconception propagation in future teaching practices.

RESEARCH METHOD

This study employed a Research and Development (R&D) approach, which involved the stages of instrument development, trial implementation, and data analysis. The validity and reliability of the instrument were analyzed using the Rasch Model.

1. Research Design

The study adopted a quantitative descriptive approach with the aim of identifying and analyzing misconceptions held by preservice teachers, specifically related to the concept of heat. The research was conducted through several stages, including the development of the diagnostic instrument, validation and reliability testing using the Rasch Model, and In the final phase of the study, trial data were evaluated to determine the instrument's effectiveness in identifying students' misconceptions.

2. Research Procedures

2.1. Development of the 2-Tier Diagnostic Test

The two-tier diagnostic test instrument consists of two parts, both in the form of multiple-choice questions. The first tier assesses students' factual knowledge, while the second tier examines the reasoning behind their chosen answers. The development of the instrument followed several stages as shown in figure 1:

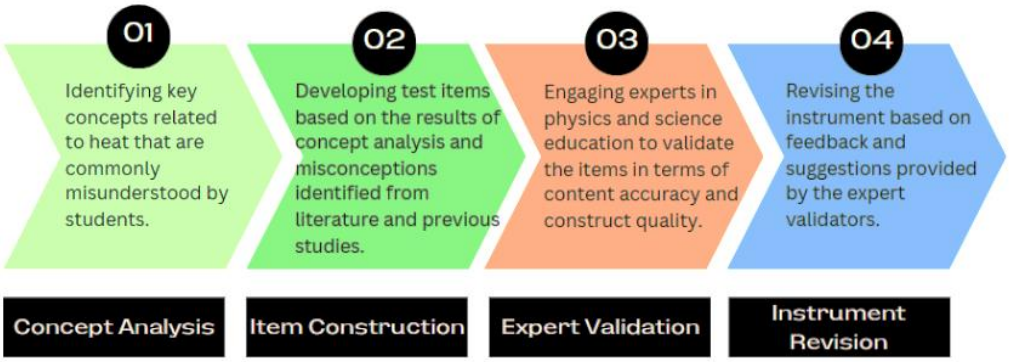

Figure 1. Instrument Development Stage

2.2. A Preliminary Test of the Instrument

Following the development phase, a preliminary test of the instrument was conducted with students from the Science Education Department who are preservice science teachers. A total of 57 participants from Universitas Negeri Padang were involved in this stage. The participants were selected using purposive sampling based on specific criteria: they were undergraduate students in their fourth and fifth semesters who had completed fundamental physics courses, including *Thermodynamics* and *Fundamental of Physics*, which cover the concept of heat and temperature. These criteria ensured that participants possessed sufficient conceptual background to respond meaningfully to the diagnostic test. The collected data consisted of participants' responses to the multiple-choice items from both tiers of the diagnostic instrument.

2.3. Data Analysis

The data obtained from the preliminary test were analyzed using the Rasch model to evaluate the validity and reliability of the diagnostic test. Identification of preservice teachers' misconceptions was conducted by examining their responses to the two-tier items, specifically by analyzing the consistency between their answers in Tier 1 (content) and Tier 2 (justification). The analysis was performed using Winsteps software, which supports Rasch model procedures. The analysis included item fit statistics (Infit and Outfit), item reliability, item separation index, and the construction of a person-item map to assess the alignment between item difficulty and participant ability.

a. Instrument Validity

Item fit in the Rasch Model is used to assess whether each test item aligns with the theoretical expectations of the model. The primary statistics used are Infit Mean Square (MNSQ) and Outfit Mean Square (MNSQ), with ideal values ranging between 0.5 and 1.5. Items falling outside of this acceptable range are subject to further review and may be revised or removed. Winsteps software generates Infit and Outfit statistics for each item, serving as indicators of how well individual items conform to the Rasch model. Items with excessively high Infit or Outfit values suggest that the question may be too difficult, misunderstood by respondents, or not functioning as intended within the Rasch framework.

b. Person-Item Map

The Rasch model can generate a map that displays both participant ability (person) and item difficulty on the same scale. This provides valuable information about whether the items have a range of difficulties that match the participants' abilities. The Rasch model produces a Wright map, also known as a person-item map, which presents the distribution of participant abilities on one side and item difficulties on the other. This map is useful for examining how participant abilities are distributed and whether the items adequately cover the full range of those abilities.

The interpretation of the person-item map can be guided by the relative positions of item difficulty and student ability along the logit scale. If most of the items are located above the distribution of student abilities, it indicates that the test items are generally too difficult for the respondents. Conversely, if the items are mostly positioned below the ability distribution, it suggests that the items are too easy and may not effectively discriminate among students with varying levels of understanding. Ideally, the items should be evenly distributed across the range of student abilities. Such a pattern would indicate that the instrument provides good diagnostic coverage and is capable of accurately measuring students across a wide spectrum of abilities.

c. Instrument Reliability

Reliability analysis was conducted using two Rasch-based indicators, as described below. First, Item Reliability measures the consistency of test items in distinguishing participants with varying levels of ability. A high reliability value (above 0.80) indicates that the items effectively differentiate between individuals with different ability levels. Second, the Item Separation Index reflects how well the instrument can classify participants based on their abilities. The separation index is also considered an indicator of instrument reliability, as a high-quality instrument should yield consistent and accurate measurements. A higher separation index suggests that the observed variation in item responses is meaningful and not due to random chance. For instance, a high separation value indicates that respondents with different ability levels tend to respond differently to the items.

d. Effectiveness in Identifying Misconceptions

Fit statistics (Infit and Outfit) are used to determine which items are particularly effective in identifying students with misconceptions. By analyzing individual item reports (Item Fit Report) and response patterns, it is possible to trace which items reveal conceptual

misunderstandings. To assess the items most effective in identifying misconceptions, the analysis focuses on Infit and Outfit statistics at the item level within the Rasch model framework. These statistics provide insights into how well each item functions according to Rasch expectations, specifically whether the item can distinguish between respondents who hold misconceptions and those who do not.

Infit (Information-weighted Fit Statistic) reflects how well responses to an item align with the expected response pattern, particularly for individuals whose ability levels are close to the item's difficulty. Infit is sensitive to unexpected responses from students to items that should be relatively easy or difficult for them. When a student has a strong misconception about a particular concept, an item with good Infit can effectively identify that misunderstanding.

Outfit (Outlier-sensitive Fit Statistic), on the other hand, is more sensitive to unusual or extreme responses, especially on items that are very easy or very difficult relative to the respondent's ability. High Outfit values may indicate response anomalies, such as when a student answers in a way that deviates significantly from model expectations, potentially due to guessing or deep-rooted misconceptions.

RESULTS AND DISCUSSION

1. Results

1. 1. Instrument Test 2-Tier

Table 1 presents the main physics concepts assessed in the diagnostic test on heat. Each concept is paired with a specific indicator that shows what students are expected to understand. This table helps identify students' misconceptions.

Table 1. The indicators of questions on Heat

Concepts	Assessment Indicators
Difference between temperature and heat	Students can explain that temperature does not directly indicate the amount of heat in an object.
Heat as energy, not a substance	Students understand that what transfers is energy (heat), not a substance or particles.
Temperature vs. quantity of heat	Students can conclude that an object with a higher temperature does not necessarily contain more heat.
Definition of heat and temperature	Students understand that temperature is a measure of how hot something is, while heat is transferred energy.

Heat exists in all objects	Students realize that all objects above 0 Kelvin contain thermal energy (heat).
Cold objects still contain heat	Students can state that cold objects still contain heat, although in smaller amounts.
Direction of heat transfer	Students recognize that heat does not always "rise", but moves from hotter to cooler areas.
Specific heat capacity	Students can explain that the rate of temperature change depends on the material's specific heat capacity.
Thermal conductivity	Students understand that metal feels colder because it transfers heat from the hand more quickly.
Heat transfer through container material	Students understand that metal conducts heat faster, causing tea to cool down more quickly.

The next stage of this research is to develop a set of two-tier diagnostic questions, as exemplified in the previously designed items. This instrument consists of total 20 two-tier items, comprising 10 items for the first tier and 10 items for the second tier, with both tiers presented in multiple-choice format. The two-tier format is particularly effective in detecting common misconceptions and differentiating between correct answers based on sound understanding and those based on guessing or flawed reasoning.

Concept: Difference between temperature and heat

1. **Tier 1:** Two objects of the same size have different temperatures: the first object has a temperature of 80°C and the second object has a temperature of 20°C. What can you conclude about the heat in these two objects?
   A. The object with the higher temperature has more heat.
   B. The object with the lower temperature has less heat.
   C. Temperature does not directly indicate the amount of heat in an object.
   D. Both objects have the same amount of heat because they are the same size.

   Tier 2: Why did you choose that answer?
   A. Because heat is always directly related to the temperature of an object.
   B. Because the higher the temperature of an object, the more heat it has.
   C. Because heat depends on both temperature and size, but cannot be concluded from temperature alone.
   D. Because the size of an object determines the amount of heat it can store.

Figure 2. The Example of Two-tier Diagnostic Question

1. 2. Validity

During the analysis using Winsteps software, the items in tier 1 were labeled with the code S, while the items in tier 2 were labeled with A. The results showed that three items-S8, A4, and S4-had Outfit Mean Square values outside the ideal range of 0.5 to 1.5, as suggested by the Rasch model. Therefore, these three items need to be revised.

INPUT: 57 Person 20 Item REPORTED: 57 Person 20 Item 2 CATS MINISTEP 5.8.1.0

Person: REAL SEP.: 1.24 REL.: .61 ... Item: REAL SEP.: 3.41 REL.: .92

Item STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
9	14	57	1.72	.33	1.02	.18	.89	-.38	.38	.36	71.9	77.7	S5
17	17	57	1.41	.31	1.40	2.49	1.46	1.97	-.02	.37	61.4	73.8	S9
10	18	57	1.32	.31	.79	-1.59	.68	-1.76	.60	.37	73.7	72.8	A5
13	19	57	1.22	.30	.95	-.36	1.04	.30	.41	.38	75.4	71.7	S7
14	21	57	1.04	.30	1.18	1.50	1.19	1.14	.20	.38	61.4	69.6	A7
11	22	57	.95	.29	.75	-2.38	.69	-2.10	.63	.38	82.5	68.6	S6
12	24	57	.78	.29	.84	-1.64	.78	-1.58	.55	.38	75.4	67.2	A6
5	27	57	.53	.29	.92	-.85	.88	-.84	.46	.37	70.2	65.9	S3
18	27	57	.53	.29	1.13	1.34	1.18	1.23	.24	.37	56.1	65.9	A9
4	28	57	.45	.29	.99	-.12	1.02	.16	.38	.37	68.4	65.6	A2
6	28	57	.45	.29	.94	-.65	.92	-.56	.44	.37	68.4	65.6	A3
1	30	57	.29	.29	1.04	.46	.99	-.04	.34	.37	57.9	65.1	S1
2	30	57	.29	.29	1.04	.46	.99	-.04	.34	.37	57.9	65.1	A1
16	32	57	.12	.29	1.07	.71	1.04	.31	.30	.37	61.4	65.3	A8
19	37	57	-.30	.30	1.04	.35	1.05	.33	.31	.35	66.7	69.1	S10
20	43	57	-.88	.33	.97	-.13	.92	-.21	.36	.31	77.2	76.9	A10
3	46	57	-1.23	.35	1.01	.09	.89	-.21	.31	.29	78.9	81.1	S2
15	48	57	-1.50	.38	1.13	.61	1.99	1.99	.03	.27	84.2	84.3	S8
8	55	57	-3.24	.73	.89	.05	.31	-.70	.33	.14	96.5	96.5	A4
7	56	57	-3.96	1.02	.93	.25	.25	-.43	.26	.10	98.2	98.3	S4
MEAN	31.1	57.0	.00	.36	1.00	.04	.96	-.07			72.2	73.3	
P.SD	12.2	.0	1.46	.18	.14	1.09	.36	1.07			11.6	9.7	

Figure 3. Presents the Infit and Outfit statistics for each item in the instrument.

1.3. Reliability

The summary table (Figure 4) below presents the reliability and separation values for the 20 items measured in this study.

SUMMARY OF 20 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	31.1	57.0	.00	.36	1.00	.04	.96	-.07
SEM	2.8	.0	.34	.04	.03	.25	.08	.25
P.SD	12.2	.0	1.46	.18	.14	1.09	.36	1.07
S.SD	12.5	.0	1.50	.18	.14	1.12	.37	1.10
MAX.	56.0	57.0	1.72	1.02	1.40	2.49	1.99	1.99
MIN.	14.0	57.0	-3.96	.29	.75	-2.38	.25	-2.10

REAL RMSE	.41	TRUE SD	1.40	SEPARATION	3.41	Item RELIABILITY	.92	
MODEL RMSE	.40	TRUE SD	1.40	SEPARATION	3.47	Item RELIABILITY	.92	
S.E. OF Item MEAN = .34								

Item RAW SCORE-TO-MEASURE CORRELATION = -.97

Global statistics: please see Table 44.

UMEAN=.0000 USCALE=1.0000

Figure 4. Item Reliability and Separation Indeks

1.4. Person-Item Map

The figure 5. illustrates the Person-Item Map, which displays the distribution of participants' abilities (upper part, in blue) and item difficulties (lower part, in red) on a common logit scale.

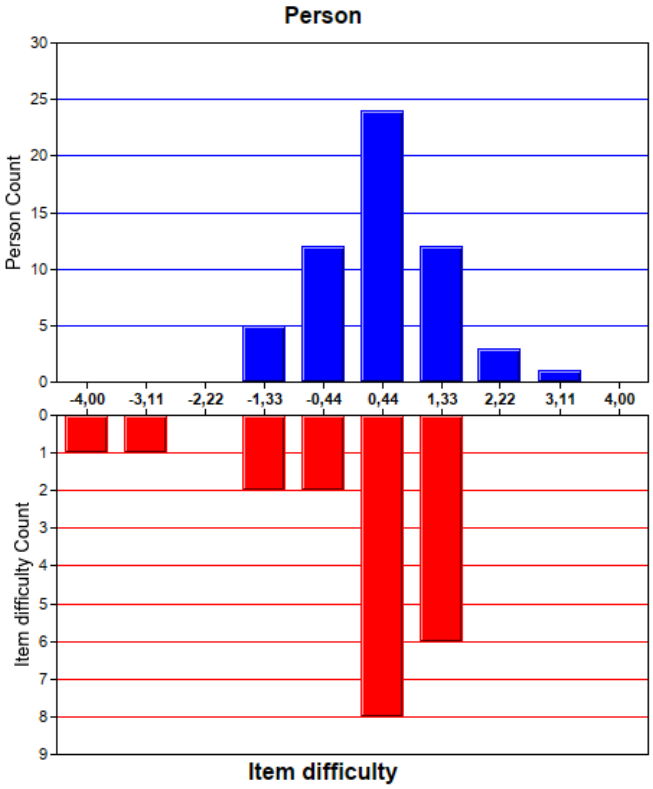

Figure 5. Person-Item Map

1.5. Effectiveness in Identifying Misconceptions

Infit Mean Square (MNSQ) and Outfit Mean Square (MNSQ) values are interpreted according to the following criteria:

- Ideal value: 1.0 (represents perfect fit according to the Rasch model).
- Acceptable range: 0.5 to 1.5 for diagnostic tests; values within this range indicate that the item functions optimally in measuring misconceptions.
- Too high (>1.5): Indicates misfit or inconsistency with the model. Items with Infit or Outfit values above 1.5 may not function effectively in identifying misconceptions.
- Too low (<0.5): Suggests that the item may be too predictable or too easy, and may provide limited information in distinguishing between students who hold misconceptions and those who do not.

- Items with misfitting Infit or Outfit values will be reviewed or removed, as such values imply that the item may not provide consistent or meaningful information regarding students' conceptual understanding.

Based on the item fit output (see Figure 3), one item was identified as misfitting: S8 (Outfit MNSQ = 1.99). In addition, two items were found to be potentially too easy: A4 (Outfit MNSQ = 0.31) and S4 (Outfit MNSQ = 0.25).

2. Discussion

The Outfit value for item S8 reached 1.99, which is considered high and indicates unexpected or inconsistent responses that do not align with the general response pattern of examinees. In Rasch theory, a high Outfit value suggests that the item may not be functioning properly in measuring the intended construct-possibly due to being too easily guessed, misunderstood by respondents, or containing ineffective distractors. Conversely, an Outfit value that is too low (<0.5) may indicate redundancy or that the item is overly predictable, thus reducing its discriminative power . Therefore, revision of items S8, A4, and S4 is necessary to improve the validity and psychometric quality of the instrument in accordance with Rasch measurement principles.(Jumadi et al., 2023)

The Person-Item Map (Figure 5) provides a visual representation of the alignment between student ability and item difficulty along a common logit scale. The upper part of the map displays the distribution of students, most of whom are clustered around the ability range of –1.33 to 2.22 logits. The lower part shows the distribution of item difficulties, which are generally spread from -0.44 to 1.33 logits. This distribution suggests that the items are moderately well targeted to the population. However, the presence of gaps-particularly at the lower and upper ends of the scale-indicates that the instrument may be less effective for students with very low or very high abilities.

According to (Jumadi et al., 2023), a well-constructed diagnostic instrument should contain items that span the full range of participant abilities to ensure measurement precision. The absence of very easy or very difficult items suggests a need to develop additional items to fill these extremes. By improving item coverage across the ability spectrum, the diagnostic tool can better differentiate students who hold strong conceptual understanding from those who exhibit misconceptions. As (Xiao et al., 2018) emphasized,

effective targeting of item difficulty to respondent ability is fundamental to achieving accurate and meaningful measurement outcomes in Rasch analysis.

The participants are densely clustered around the center, particularly at 0.44 logits, with the highest count exceeding 20 individuals. In contrast, the items are skewed more toward the moderate to slightly difficult range, with relatively fewer items falling below -1.33 logits. This indicates a potential mismatch for lower-ability students, who may not encounter enough items suited to their level, thus limiting the measurement precision for this group.

There are noticeable gaps in item distribution below –2.22 logits and above 2.22 logits, indicating a lack of very easy or very difficult items. As a result, students with extremely low or high ability levels may not be adequately assessed. This highlights the need to develop additional items that target both ends of the ability spectrum. Overall, the instrument sufficiently covers the central range of participant ability. However, to enhance validity and measurement precision, future test development should consider adding items with lower difficulty (logits < –1.33) and adding items with higher difficulty (logits > 2.22).

Based on Figure 4, the instrument shows an item reliability of 0.92 and an item separation index of 3.47. This high reliability indicates a strong consistency in the instrument's ability to distinguish between items of varying difficulty levels. In Rasch analysis, item reliability values above 0.80 are considered excellent, and a value of 0.92 suggests that the item difficulty hierarchy is stable and can be reliably interpreted across similar populations (Kania et al., 2020).

Moreover, the item separation value of 3.47 indicates that the items can be grouped into more than three distinct difficulty strata. A separation index above 2.0 demonstrates the instrument's effectiveness in distinguishing between easy, moderate, and difficult items (Salele et al., 2025). Therefore, the instrument has a good spread of item difficulties and provides sufficient information for valid interpretation of student performance levels.

Items S8 exceeded the upper threshold of acceptable fit, indicating a substantial degree of misfit with the Rasch model. High Outfit values suggest unexpected or inconsistent responses from students, often resulting from items that may be poorly worded, ambiguous, or not aligned with the intended construct (Salele et al., 2025). Although

S4 still falls within the acceptable range, its relatively high value suggests it may be approaching a level of reduced measurement precision and should be monitored or revised.

These results indicate that S8, A4 and S4 should be critically reviewed and potentially revised in subsequent test versions to enhance the instrument's ability to detect conceptual misconceptions. The misfit of these items may have been caused by unclear wording, overlapping concepts with other items, or distractors that failed to reflect students' actual reasoning patterns. Such weaknesses highlight the importance of aligning each item with the underlying construct of the concept being measured and ensuring that distractors are based on empirically identified misconceptions. To avoid similar issues in future test versions, item validation should include expert review and pilot testing with varied ability levels to confirm item clarity and conceptual focus. Evaluating item fit is therefore essential, not only for confirming the structural validity of the instrument, but also for ensuring that each item contributes meaningfully to the identification of student learning difficulties within the topic of heat. This approach also reflects one of the key principles in diagnostic test development, namely that each item must directly represent a misconception category supported by theoretical and empirical evidence.

CONCLUSION

To achieve a more balanced alignment between item difficulty and student ability, and to ensure accurate measurement across the full range of the ability spectrum, we will develop additional items representing the lower and upper extremes of the logit scale. This will ensure that the instrument not only evaluates students with moderate abilities but also effectively captures the performance of those at both the lower and higher ends. In conclusion, the instrument demonstrates strong measurement properties and is highly suitable for diagnostic assessment in the conceptual domain of heat.

ACKNOWLEDGMENTS

The author would like to express sincere gratitude to the Direktorat Penelitian dan Pengabdian kepada Masyarakat, Direktorat Jenderal Riset dan Pengembangan, Kementerian Pendidikan Tinggi, Sains, dan Teknologi Republik Indonesia, for the financial support provided through the 2025 research funding scheme. This study would not have

been possible without their generous support and dedication to advancing scientific research and education.

REFERENCES

Aydeniz, M., Bilican, K., & Kirbulut, Z. D. (2017). Exploring Pre-Service Elementary Science Teachers' Conceptual Understanding of Particulate Nature of Matter through Three-Tier Diagnostic Test. *International Journal of Education in Mathematics, Science and Technology*, *5*(3), 221–234. https://doi.org/10.18404/ijemst.296036

Boateng, S. (2024). Assessing conceptual difficulties experienced by pre-service chemistry teachers in organic chemistry. *EURASIA Journal of Mathematics, Science and Technology Education*, *20*(2), em2398. https://doi.org/10.29333/ejmste/14156

Buchari, A. (2018). Peran guru dalam pengelolaan pembelajaran. *Jurnal Ilmiah Iqra'*, *12*(2), 106–124. http://journal.iain-manado.ac.id/index.php/JII

Cahyaningtyas, C. D., Fatma, E., Rianto, P. A. M., Nuha, U., Wahyuni, S., & Yusmar, F. (2023). Analisis miskonsepsi siswa smp pada materi konsep suhu dan kalor. *Jurnal Ilmiah Wahana Pendidikan*, *9*(15), 71–75. https://doi.org/10.5281/zenodo.8200897

Duit, R., & Treagust, D. F. (2003). Conceptual change: A powerful framework for improving science teaching and learning. *International Journal of Science Education*, *25*(6), 671–688. https://doi.org/10.1080/09500690305016

Hoppe, T., Renkl, A., Seidel, T., Rettig, S., & Riess, W. (2020). Exploring how teachers diagnose student conceptions about the cycle of matter. *Sustainability*, *12*(10), 4184. https://doi.org/10.3390/su12104184

Irmak, M., Inaltun, H., Ercan-Dursun, J., Yaniş-Kelleci, H., & Yürük, N. (2023). Development and application of a three-tier diagnostic test to assess pre-service science teachers' understanding on work-power and energy concepts. *International Journal of Science and Mathematics Education*, *21*(1), 159–185. https://doi.org/10.1007/s10763-021-10242-6

Istiyono, E., Dwandaru, W. S. B., Fenditasari, K., Ayub, M. R. S. S. N., & Saepuzaman, D. (2023). The Development of a Four-Tier Diagnostic Test Based on Modern Test Theory in Physics Education. *European Journal of Educational Research*, *12*(1).

Jumadi, J., Sukarelawan, M. I., & Kuswanto, H. (2023). An investigation of item bias in the four-tier diagnostic test using Rasch model. *Int J Eval & Res Educ ISSN*, *2252*(8822), 8822.

Kania, V. I., Samsudin, A., Purwanto, A. H. A., Rasmitadila, R. R., Jermsittiparsert, K., & Nurtanto, M. (2020). Multitier of greenhouse effect (Moge) instrument development to identify middle school students' mental model in Thailand with rasch analysis,". *Int. J. Adv. Sci. Technol*, *29*(7), 3223–3237.

Laeli, C. M. H. (2020). Misconception of science learning in primary school students. *3rd International Conference on Learning Innovation and Quality Education (ICLIQE 2019)*, 657–671. https://doi.org/10.2991/assehr.k.200129.083

Liu, L., Cisterna, D., Kinsey, D., Qi, Y., & Steimel, K. (2024). AI-Based Diagnosis of Student Reasoning. *Uses of Artificial Intelligence in STEM Education*, 162. https://books.google.co.id/books?hl=en&lr=&id=JYQoEQAAQBAJ&oi=fnd&pg=PA162&dq=AI-Based+Diagnosis+of+Student+Reasoning&ots=v3JNRovhfC&sig=wyivSUIoKeA6bk9CGRL

fxziS5hw&redir_esc=y#v=onepage&q=AI-Based%20Diagnosis%20of%20Student%20Reasoning&f=false

Mukhlisa, N. (2021). Miskonsepsi pada peserta didik. *SPEED Journal: Journal of Special Education*, *4*(2), 123–133. https://doi.org/10.31537/speed.v4i2.403

Oladejo, A. I., Ademola, I. A., Ayanwale, M. A., & Tobih, D. (2023). Concept Difficulty in Secondary School Chemistry--An Intra-Play of Gender, School Location and School Type. *Journal of Technology and Science Education*, *13*(1), 255–275. https://doi.org/10.3926/jotse.1902

Salele, N., Khan, M. S. H., Hasan, M., & Ali, S. (2025). Advancing Four-Tier Diagnostic Assessments: A Novel Approach to Mapping Engineering Students' Conceptual Understanding in Microwave Engineering Course. *IEEE Access*.

Suparman, A. R., Rohaeti, E., & Wening, S. (2024). Development of Computer-Based Chemical Five-Tier Diagnostic Test Instruments: A Generalized Partial Credit Model. *Journal on Efficiency and Responsibility in Education and Science*, *17*(1), 92–106. https://doi.org/10.7160/eriesj.2024.170108

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

Tsui, C., & Treagust, D. (2010). Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. *International Journal of Science Education*, *32*(8), 1073–1098. https://doi.org/10.1080/09500690902951429

Ugwuanyi, C. S., Ezema, M. J., & Orji, E. I. (2023). Evaluating the instructional efficacies of conceptual change models on students' conceptual change achievement and self-efficacy in particulate nature matter in physics. *SAGE Open*, *13*(1), 21582440231153852. https://doi.org/10.1177/21582440231153851

Wangdi, D., Kanthang, P., & Precharattana, M. (2017). Development of a hands-on model embedded with guided inquiry laboratory to enhance students' understanding of law of mechanical energy conservation. *Asia-Pacific Forum on Science Learning and Teaching*, *18*(2), 1–26.

Xiao, Y., Han, J., Koenig, K., Xiong, J., & Bao, L. (2018). Multilevel Rasch modeling of two-tier multiple choice test: A case study using Lawson's classroom test of scientific reasoning. *Physical Review Physics Education Research*, *14*(2), 20104.

Yuliana, I., Artawan, P., & Heny, A. P. (2023). Profil miskonsepsi siswa pada materi suhu dan kalor. *NUSRA: Jurnal Penelitian Dan Ilmu Pendidikan*, *4*(4), 1161–1166. https://doi.org/10.55681/nusra.v4i4.1763