Jurnal Pendidikan MIPA
e-ISSN: 2685-5488 | p-ISSN: 1411-2531
URL: https://jpmipa.fkip.unila.ac.id/index.php/jpmipa
Jounal's DOI: http://doi.org/10.23960/jpmipa

Vol. 27, No. 2, 2026
Article ID: 1177
Pages: 867-886

Mapping Misconceptions in Heat and Temperature: Rasch Analysis of Four-Tier Diagnostic Responses

Roziqin[1], Achmad Samsudin[1,*], Duden Saepuzaman[1], Ida Kaniawati[1], Mimin Iryanti[1], & Nor Farahwahidah Abdul Rahman[2]

[1]Department of Physics Education, Universitas Pendidikan Indonesia, Indonesia
[2]Faculty of Education Science and Technology, Universiti Teknologi Malaysia, Malaysia

*Corresponding email: achmadsamsudin@upi.edu

Received: 31 December 2025 *Accepted: 26 April 2026* *Published: 13 May 2026*

Abstract: Understanding of heat and temperature has been a major obstacle in students' learning of thermodynamics, but many diagnostic studies still report descriptive data on the frequencies of students' responses, which lack evidence about the psychometric quality of the measurement, the ranking of items, and the validity of the response patterns. This paper aims to address this by combining a four-tier diagnostic test with the Rasch Partial Credit Model (PCM) to assess Grade XII students' conceptual knowledge of heat and temperature and to evaluate the diagnostic test's psychometric quality. We surveyed 60 senior high school students in Bandung, Indonesia, using a quantitative approach. They responded to a diagnostic test with four-tiered items on the concepts of temperature, thermal expansion, heating effect on temperature, states of matter, and heat transfer. We classified these responses into conceptual categories and applied Rasch-PCM to assess reliability, item and person fit, unidimensionality, category functioning, gender-based differential item functioning (DIF), Wright map, and the match between difficulty and ability. The findings show the instrument had satisfactory psychometric characteristics, including a reliable hierarchy of item difficulty, acceptable fit to the model, essential unidimensionality, and mostly well-functioning response categories. There was little evidence of measurement bias as items functioned similarly for both males and females. The diagnostic map indicated that students' concepts were characterized by partial understanding and misunderstanding, and by ongoing confusion among heat, temperature, and thermal energy. The most challenging concepts were density change upon heating, convection, thermal expansion on the particle level, thermal feelings, and mechanisms of heat transfer. The person-fit and scalogram analyses also showed irregular response patterns of a minority group of students, suggesting a potential for guessing, disjointed reasoning, or unstable understanding. These findings show that the use of four-tier diagnostic responses with Rasch-PCM offers a higher-standard approach to identifying misconception structures than descriptive diagnostic analysis.

Keywords: student misconceptions, four-tier diagnostic, rasch analysis, heat, temperature.

Article's DOI: https://doi.org/10.23960/jpmipa.v27i2.pp867-886

INTRODUCTION

It is necessary to understand the phenomena of thermodynamics and their numerous applications in science and engineering (Bandhu et al., 2023; Bezlutsky et al., 2024; Fauzan et al., 2021; Yang et al., 2020). Students are expected to understand the quantitative relationship between the two, such as recognizing temperature as an indicator of thermal state and heat as transferred energy. Nonetheless, the results of the study show that students and future educators often confuse temperature with the amount of heat or misunderstand temperature changes, which distort the basic notions (Busyairi et al., 2022; Fenditasari et al., 2020; Nabilah et al., 2019). As a result, they have limited ability to solve simple thermodynamic problems involving complex subjects such as heat transfer and phase

transitions. One study reveals that 58.73% of potential physics teachers hold misconceptions, whereas only 30.82% have an accurate understanding of the concept; most misconceptions concern thermal equilibrium. Moreover, there are many misconceptions in the distinction between the notions of temperature, heat, and thermal energy (Busyairi et al., 2022). This case highlights the fact that a sensitive and metric diagnostic tool should be used to ensure that the detection of misconceptions would present the basis of efficient learning interventions (Fenditasari et al., 2020; Samsudin et al., 2024). In this context, "diagnostic" implies more than reporting overall scores; it requires identifying which concepts become bottlenecks, how strongly they differentiate students, and which response patterns indicate stable misconceptions rather than momentary uncertainty (Caleon & Subramaniam, 2010; Diani et al., 2019).

To meet that requirement, the four-tier diagnostic test employs a more detailed method, using multiple-choice answers, reasonableness checks, and confidence levels to distinguish between ignorance and deep-rooted misconceptions (Diani et al., 2019; Kaltakci-Gurel et al., 2017; Salele et al., 2025). The technique detects response types that are undetected in two- or three-level tests (Caleon & Subramaniam, 2010; Hermita et al., 2017; Istiyono, 2022; Istiyono et al., 2023). The superiority of FTDT, with high Aiken V validity, INFIT MNSQ within a reasonable range, and reliability of 0.90, is evident in the study (Istiyono et al., 2023). The error structure in the conceptual model is more evident, as the eight categories of responses in the PCM model can be used to interconnect misconceptions across different mechanics subjects. Besides, web-based implementation enables large-scale response analysis and accelerates data collection (Istiyono et al., 2022). Nevertheless, four levels are usually used in descriptive statistics; therefore, trends of high confidence in inaccurate answers cannot be identified everywhere, so a more detailed analysis will be required. Therefore, an analysis framework that preserves the diagnostic nature of four-tier responses while providing interpretable measurement parameters is needed to connect response patterns to learning difficulties and misconception structure (Istiyono et al., 2023).

Rasch analysis in this case offers a metric framework that can transform an ordinal data to an interval scale that can be used to assess reliability, fit, and more informative person-item maps (Hermita et al., 2021; Ismail et al., 2021; Planinic et al., 2019). PCM and RSM are the polytomous models that are gaining increased relevance in analyzing categorical responses that involve confidence levels and reasons (Dhyaaldian et al., 2022; Xiao et al., 2023). Nevertheless, polytomous analysis should be studied carefully in terms of functions of categories and unidimensionality; otherwise, the misconception maps may be deceptive. Therefore, the use of the four-tier diagnostic test and the Rasch model needs to be psychometrically valid (Istiyono et al., 2024; Walfridsson et al., 2022). In a diagnostic context, Rasch parameters are meaningful for learning analysis: item "difficulty" can be interpreted as the relative conceptual demand of specific heat-temperature ideas on a common logit scale, while person "ability" represents students' overall mastery level within the same construct (Hermita et al., 2021; Planinic et al., 2019). Furthermore, item fit (INFIT/OUTFIT) helps indicate whether responses align with expected cognitive progression, where misfitting items may signal ambiguous reasoning options or heterogeneous misconception pathways; person fit can reveal aberrant response patterns such as guessing or fragmented understanding that are pedagogically informative rather than merely noise (Planinic et al., 2019). A scatter plot operationalizes this

connection by directly showing whether item difficulty appropriately targets the distribution of student ability, highlighting concept areas that are systematically too difficult or too easy for the sampled population (Hermita et al., 2021; Ismail et al., 2021).

Practically, it is also beneficial for integrating person maps with other analytics, not limited to analytics. Person item maps may expose those biased aspects, dysfunctional sets of beliefs, and respondent groups prone to some forms of misconceptions (Jumadi et al., 2023). Empirical research indicates that the application of Rasch enhances the validity and reliability of tools if the processes of item development are performed properly (Ali et al., 2017; Farida et al., 2024). In line with this, Busyairi et al. (2022) documented persistent misconceptions in temperature and heat, indicating that instructional innovation and targeted remediation remain urgent.

While other studies have used four-tier diagnostic tests (4T-DT) to detect students' misconceptions about heat and temperature, these tests have some limitations. For instance, Sukarelawan et al. (2021) diagnosed students' misconceptions using 4T-HTDT in Bantul, Yogyakarta, but the analysis primarily employed descriptive methods to report the percentage of students' conceptual comprehension. This approach provides an overview of the percentage of students' misconceptions. However, it does not provide details on item difficulty, response category functioning, model fit, or item bias. Furthermore, the misconception profile in one education system may not apply to other regions, as students' conceptual understanding may be affected by their learning experiences, school environments, and local teaching practices. The current study builds on previous research by investigating students' misconceptions about heat and temperature in Bandung, West Java, using a polytomous Rasch-PCM approach. This allows us to not only diagnose patterns of misconceptions but also assess the psychometric quality of the FTDT, identify a hierarchy of conceptual difficulty, and explore the functioning of response categories and the gender differences in DIF evidence. Therefore, the study not only helps diagnose misconceptions but also validates the four-tier diagnostic measurement across different regions.

The paper designed a Four-Tier Heat & Temperature Diagnostic Test and tested it using a polytomous Rasch model. This was aimed at comparing quantitative misconceptions in students with the tools that address the criteria of fit, reliability, and functionality (Hawkins et al., 2024; Hermita et al., 2021; Jumadi et al., 2023). The findings of the research include an analysis of misconceptions using a four-tier model that specifically concerns temperature and heat, an explanation of the reasons for using a polytomous Rasch model, and a misconception map that can be applied to enhance learning. Projected outcomes also furnish more noticeable evidence on how to plan to improve and allocate resources regarding certain misconceptions (Munggarani et al., 2021; Romine et al., 2015). Therefore, the study is guided by the following research questions:

1. What misconception patterns on heat and temperature are identified from FTDT response profiles?
2. Does the FTDT demonstrate adequate psychometric quality under the Rasch-PCM model in terms of reliability, fit, unidimensionality, and category functioning?
3. Do the items function equivalently across gender groups based on DIF evidence?

METHOD

Participants

This research was conducted on a group of 60 students (30 male and 30 female) from classes XII C and XII D of a government high

school in Bandung, aged 16-18 years. Purposive sampling was used in the paper because information on temperature and heat was formally taught at this stage. Consequently, misconceptions would be associated with real learning experiences. Participants were included if they (1) were enrolled in grade XII, (2) had completed instruction on temperature and heat, and (3) completed all FTDT responses via Google Forms. The data were collected in December 2025. Although Rasch analysis is frequently conducted with larger samples, a sample size of 60 was considered adequate for preliminary instrument validation using the Rasch Partial Credit Model (PCM), particularly for evaluating item fit, reliability, and category functioning. Nevertheless, the authors acknowledge that the sample was drawn from a single school and therefore the findings should be interpreted within this study context. This limitation is further addressed in the discussion section.

Research Design and Procedures

A quantitative survey was employed in this study, accompanied by psychometric analysis to map students' misconceptions about temperature and heat using the Four-Tier Diagnostic Test (FTDT). The reason a survey was selected is that the FTDT is better at capturing internal cognitive structures as a combination of answers, reasons, and confidence, because this approach avoids confining misconception data to bare self-reporting. The four-level tool generates polytomous responses; thus, the Rasch Partial Credit Model (PCM) was employed to obtain interval-scaled logit measures, test item and respondent reliability, and assess category functions. FTDT and Rasch are combined to enhance the identification of misconceptions, since, unlike classical analysis, this approach enables the identification of misfit items, the mapping of difficulty hierarchies, and the study of measurement structure consistency.

Instrument

The instrument used was a Four-Tier Diagnostic Test (FTDT), adopted from Sukarelawan et al. (2021). The FTDT format in this study followed the Four-Tier Heat and Temperature Diagnostic Test (4T-HTDT) structure proposed by Sukarelawan et al. (2021) without modification to the conceptual indicators and item intent. The FTDT consists of four tiers: Tier 1 (multiple-choice content answer), Tier 2 (confidence in Tier 1), Tier 3 (conceptual reasoning), and Tier 4 (confidence in Tier 3). The confidence scale followed a 4-point format (1 = only guessing, 2 = unsure, 3 = sure, 4 = very sure).

The instrument indicators were organized based on the Four-Tier Heat and Temperature Diagnostic Test (4T-HTDT) structure and covered five concept groups: (1) temperature, (2) thermal expansion, (3) the effect of heat on the temperature of an object, (4) the effect of heat on the state of matter, (5) heat and heat transfer. The adopted instrument consisted of 20 items distributed across the concept groups (6 items for temperature, 4 items for expansion, 2 items for the effect of heat on the temperature of an object, 2 items for the effect of heat on the state of matter, and 6 items for heat and heat transfer). An example item format includes a question about heat transfer phenomena in daily contexts (Tier 1), followed by a reasoning option explaining energy transfer mechanisms (Tier 3), with confidence ratings for both tiers.

Content validity was evaluated by experts for material relevance, construction, and linguistic clarity. Following the original development procedure, content validity was evaluated across three aspects (material, construction, and language/cultural clarity), and all items showed Aiken's V indices above the cut-off value of 0.75, indicating adequate representativeness and clarity. As the instrument was adopted without modification, the study focused on empirical

validation in the present sample using the Rasch Partial Credit Model (PCM) to confirm its measurement functioning in this context. Item fit, difficulty hierarchy, and unidimensionality were evaluated. All items fell within acceptable INFIT and OUTFIT MnSq ranges, except one item, which was retained because MnSq values between 1.50 and 2.00 are still considered productive for measurement. Reliability analysis showed high item separation and an item reliability coefficient of 0.97, indicating strong internal consistency and stable estimation of item difficulty.

Data Analysis

The method suggested by Boone et al. (2014) was assigned to evaluate the data, which presupposes the application of the Rasch Partial Credit Model (PCM). The PCM was selected because the FTDT produces polytomous response patterns across tiers, and Rasch modeling transforms raw scores into interval-level logit measures, facilitating more robust interpretation of item difficulty and student ability. Students have varying degrees of thinking, such as Solid Understanding (SU), Partial Understanding (PU), Partial Negative (PN), Misconception (MC), No Understanding (NU), and No Coding (NC). PU and PN differ in nature: PU is a response pattern of correct answering with inadequate reasoning and/or confidence, and PN is a response pattern of incorrect answering with reasoning and confidence patterns that indicate unstable, partially developed conceptual knowledge. While several PU or PN response patterns have identical scores, interpretation is based on the profile of tier co-occurrence rather than the scores. Thus, PU and PN response patterns were distinguished analytically by response-pattern diagnosis across tiers, rather than by scores. The results in Table 1 indicate the number of concepts students hold.

Table 1. Conception level category

Category	Tier 1	Tier 2	Tier 3	Tier 4	Score
SU	T	S	T	S	4
PU	T	S	T	U	3
	T	U	T	S	3
	T	U	T	U	3
	T	S	F	S	2
	T	S	F	U	2
	T	U	F	S	2
	T	U	F	U	2
PN	F	S	T	S	2
	F	S	T	U	2
	F	U	T	S	2
	F	U	T	U	2
MC	F	S	F	S	1
NU	F	S	F	U	0
	F	U	F	S	0
	F	U	F	U	0
NC	NCA				

Note: T=True; F=False; S=Sure; U=Unsure; NCA=No Clear Answer; NC=No Coding; T=S= score 1; F=U= score 0

The capacity of the instrument to differentiate the levels of the students' understanding and the distribution of the item difficulty was tested according to the standardized criteria, where reliability coefficients were excellent (e"0.80), good (0.71 to 0.79), fair (0.61 to 0.70), and very low (d"0.50). Indices of person and item separation were also examined to further assess the instrument's ability to differentiate among various levels of student understanding and item difficulty. The scale of construct validity was analyzed through the values of INFIT and OUTFIT MNSQ (range 0.5 to 1.5), Z-standardized scores (-2.0 to 2.0), and point measure correlation (0.40 to 0.85); items with at least one of the values were accepted, and those with all of the values were suggested to be changed. Principal Component Analysis of Residuals (PCAR) was used to assess unidimensionality based on the percentage of explained variance (>20%), the percentage of unexplained variance (<15%), and the first-eigenvalue contrast (<3%). The challenge for each item was then calculated, and the Wright Maps were applied to visualize the correspondence between student skills and the item hierarchy. Lastly, the misconception patterns were categorized using the FTDT to create a comprehensive conceptual mapping of the temperature and heat concepts. Item difficulty and person ability were expressed in logit units, and Wright Maps were used to interpret the item hierarchy and student distribution on the same measurement scale.

■ RESULT AND DISCUSSION

Quality of Instrument

The quality of the four-tier instrument for temperature and heat is assessed using elementary statistical analyses of person-item data within the Rasch model, on which the stability and reliability of the measurements are based. This is a rather significant step, since the subsequent Item fit, person fit, dimensionality, and ability mapping are highly reliant on model estimation accuracy (Hikmah et al., 2021; Jumadi et al., 2023; Saputra & Tania, 2024). It includes a four-level diagnostic tool, which is the most appropriate approach that helps researchers assess the quality of the instrument holistically, such as summary statistics of measurements, item and respondent goodness, unidimensionality measures, and the analysis of the possible intergroup bias, using DIF (Agustina et al., 2023; Istiyono et al., 2024; Salele et al., 2025). Each of the visualizations (Figures 1 to 8) is based on empirical data, with a Rasch-based interpretation and existing findings from diagnostic research, providing not only technical validation but also support for the instrument's diagnostic function.

Reliability and Measurement Quality

Figure 1 provides a statistical summary of the consistency of students' responses to each item. The person statistics indicate that students can be differentiated consistently into broad levels of conceptual understanding, so the instrument is suitable for diagnostic profiling rather than mere score ranking. The person statistics indicate that students can be differentiated consistently into broad levels of conceptual understanding, so the instrument is suitable for diagnostic profiling rather than mere score ranking. In practical terms, this means the FTDT can identify students with weak, emerging, and stronger conceptual structures, which is important for planning differentiated remediation in heat and temperature. Rather than indicating only statistical consistency, the person indices show that the ordering of students is stable enough to support meaningful instructional decisions. Accordingly, the instrument has value for grouping students into broad bands of understanding that can be linked to targeted remediation in heat and temperature.

The item statistics indicate that the test has a reliable hierarchy of conceptual difficulty (that

SUMMARY OF 60 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	21.2	20.0	-1.07	.28	.97	-.11	.99	-.07
SEM	1.2	.0	.09	.01	.05	.15	.05	.15
P.SD	9.5	.0	.68	.04	.39	1.16	.41	1.17
S.SD	9.6	.0	.69	.04	.39	1.17	.41	1.18
MAX.	49.0	20.0	.59	.42	2.39	3.44	2.47	3.36
MIN.	6.0	20.0	-2.53	.22	.28	-3.22	.30	-3.08

REAL RMSE	.30	TRUE SD	.61	SEPARATION 2.04	Person RELIABILITY	.81
MODEL RMSE	.28	TRUE SD	.62	SEPARATION 2.18	Person RELIABILITY	.83
S.E. OF Person MEAN = .09						

Person RAW SCORE-TO-MEASURE CORRELATION = .99

CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .84 SEM = 3.77

STANDARDIZED (50 ITEM) RELIABILITY = .92

SUMMARY OF 20 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	63.5	60.0	.00	.16	1.01	-.03	.99	-.14
SEM	5.0	.0	.12	.00	.05	.30	.05	.28
P.SD	21.9	.0	.52	.02	.24	1.29	.22	1.23
S.SD	22.4	.0	.53	.02	.25	1.33	.23	1.26
MAX.	114.0	60.0	.80	.19	1.54	2.42	1.31	1.48
MIN.	34.0	60.0	-1.08	.14	.61	-2.38	.61	-2.44

REAL RMSE	.17	TRUE SD	.49	SEPARATION 2.88	Item RELIABILITY	.89
MODEL RMSE	.16	TRUE SD	.49	SEPARATION 3.06	Item RELIABILITY	.90
S.E. OF Item MEAN = .12						

Item RAW SCORE-TO-MEASURE CORRELATION = -.99

Figure 1. Reliability and measurement quality

is, the arrangement of items from more to less difficult is interpretable and not due to measurement error). This is relevant for diagnostic purposes because difficult items can serve as a proxy for concepts that require understanding in terms of the coordination of principles. In contrast, easy items can serve as a proxy for concepts that are more accessible at this level of ability. Such stability is essential for diagnostic use because it supports consistent interpretation of which concepts are genuinely more challenging across administrations. Overall, the person and item evidence indicate that the instrument is adequately calibrated for diagnosis: it can map variation in student understanding and show which parts of the topic are systematically more demanding. Therefore, Figure 1 supports the use of the FTDT not only as a measurement tool, but also as a basis for identifying where conceptual support should be focused in instruction.

The finding that items can be organized into approximately four difficulty levels should be interpreted diagnostically, not only statistically. Based on the item measure ordering (Figure 2) and the Wright map targeting (Figure 7), the most difficult cluster is dominated by Q10, Q20, and Q9 (with Q19 also appearing in the upper-difficulty range). These items require students to coordinate multiple principles rather than rely on surface cues. For instance, Q10 probes density change under heating and requires integrating mass conservation with volume expansion and

the definition of density (m/V). Q20 targets convection and demands causal reasoning about temperature–density relations that drive fluid circulation. Q9 examines thermal expansion and challenges the common particle-number misconception by requiring a micro–macro mechanism (i.e., increased interparticle spacing). Q19 differentiates thermal sensation from temperature by invoking thermal equilibrium and conductivity. Therefore, these difficult items are diagnostically valuable because they tend to separate students who can coordinate models from those who rely on everyday intuition (e.g., "feels colder" means "lower temperature").

Cronbach's alpha complements the Rasch-based evidence by indicating that the items work together to measure a common construct, namely, students' conceptual understanding of heat and temperature. Within the Rasch framework, however, this coefficient is better interpreted as supporting evidence, while the separation, reliability, and error indices provide more direct information about measurement precision. Because the estimation error is relatively small, the resulting misconception map can be interpreted with reasonable confidence and used to support follow-up diagnostic decisions

Model Fit and Item Fit

In addition, the Model Fit review extends the item fit analysis shown in Figure 2, which provides estimates of the difficulty levels, standard errors, and Infit-Outfit MNSQ and ZSTD values. The point of this review is to ensure that the variation in student responses is attributable to conceptual ability rather than to item instability or measurement noise.

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
10	34	60	.80	.19	1.11	.59	1.12	.60	.33	.43	56.7	54.4	Q10
20	37	60	.70	.18	1.16	.83	1.13	.63	.39	.44	50.0	52.4	Q20
9	38	60	.66	.18	1.24	1.18	1.31	1.40	.26	.44	45.0	52.1	Q9
19	41	60	.57	.18	1.54	2.42	1.29	1.36	.46	.45	38.3	49.6	Q19
2	47	60	.38	.17	1.32	1.59	1.30	1.48	.41	.47	33.3	45.9	Q2
15	48	60	.35	.17	1.05	.34	1.11	.60	.34	.47	51.7	45.8	Q15
16	48	60	.35	.17	.85	-.76	.85	-.73	.62	.47	45.0	45.8	Q16
13	52	60	.24	.16	.61	-2.38	.62	-2.27	.58	.48	58.3	43.6	Q13
18	61	60	.01	.16	.69	-1.86	.61	-2.44	.39	.50	56.7	42.4	Q18
17	62	60	-.01	.16	.88	-.65	.92	-.38	.58	.50	38.3	42.4	Q17
5	63	60	-.04	.16	1.02	.17	.99	.03	.57	.50	41.7	42.5	Q5
1	64	60	-.06	.15	1.06	.36	.98	-.07	.51	.50	63.3	42.0	Q1
12	64	60	-.06	.15	.98	-.03	1.00	.05	.42	.50	48.3	42.0	Q12
8	67	60	-.13	.15	1.24	1.29	1.18	.99	.62	.50	16.7	41.0	Q8
14	72	60	-.25	.15	.71	-1.72	.72	-1.72	.68	.51	40.0	40.4	Q14
6	76	60	-.34	.15	.81	-1.06	.77	-1.34	.49	.52	48.3	40.2	Q6
7	82	60	-.46	.14	1.09	.54	1.05	.33	.48	.53	35.0	40.8	Q7
3	92	60	-.67	.14	.70	-1.80	.71	-1.80	.48	.54	48.3	39.3	Q3
4	108	60	-.97	.14	.83	-.99	.84	-.90	.64	.55	43.3	38.7	Q4
11	114	60	-1.08	.14	1.25	1.41	1.24	1.33	.54	.55	35.0	39.4	Q11
MEAN	63.5	60.0	.00	.16	1.01	-.03	.99	-.14			44.7	44.0	
P.SD	21.9	.0	.52	.02	.24	1.29	.22	1.23			10.4	4.6	

Figure 2. Model fit and item fit

The table depicted in Figure 2 presents the distribution of item fit and Infit-Outfit MNSQ and ZSTD. The average Infit MNSQ and Outfit MNSQ are 1.01 and 0.99, respectively, which are very similar to the optimum model values, as per the range of 0.5 to 1.5 commonly adopted in the studies of the validation of educational instruments (Hikmah et al., 2021). The value of ZSTD infit = -0.03 and ZSTD outfit = -0.14 implies that no item has a significant difference

between its expectations of the model, when the ZSTD guideline value is -2 to +2, implying that the items are functioning consistently in measuring the intended construct, rather than introducing systematic distortion (Avinç & Doðan, 2024).

The lowest point-measure correlations were found in Q9, Q10, Q15, Q18, and Q20, and these items appear to be conceptually demanding rather than merely defective. Q9 requires students to relate thermal expansion to increased inter-particle spacing at the microscopic level. Q10 requires coordination of mass conservation, volume expansion, and density change during heating. Q15 challenges the common misconception that "cold" can flow, rather than that heat is transferred from higher to lower temperatures. Q18 requires students to distinguish thermal sensation from temperature across different materials. Q20 demands a causal understanding of convection as a density-driven fluid process. Accordingly, the low correlations in these items are better interpreted as evidence of persistent conceptual difficulty and heterogeneous reasoning than as simple item-writing flaws. For details on the conception category (Solid Understanding, Partial Understanding, Partial Negative, Misconception, No Understanding), please refer to Appendix A.

These outcomes correspond to the results of the earlier four-tier tool applied to map misconceptions in fundamental physics topics (Çelikkanlý & Kýzýlcýk, 2022). When all three categories are met, it will be handled under the highly suitable category. When it has only two categories, then it comes under the category of Suitable. When one of the categories is met, it will be included under the Less Suitable category. When none of the three categories are met, it will be included under the Not Suitable category (Istiyono et al., 2024). From the perspective of the diagnosis, this classification suggests selective refinement of items, particularly to improve discrimination in areas related to persistent misconceptions, without compromising the instrument's overall construct validity.

Construct Validity and Unidimensionality

Figure 3 of the dimensionality analysis of raw variance explained by measures, with a standard of 20%, showed a value of 35.7%. It

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = Item information units

	Eigenvalue	Observed		Expected
Total raw variance in observations =	31.1014	100.0%		100.0%
Raw variance explained by measures =	11.1014	35.7%		35.3%
Raw variance explained by persons =	6.9074	22.2%		21.9%
Raw Variance explained by items =	4.1940	13.5%		13.3%
Raw unexplained variance (total) =	20.0000	64.3%	100.0%	64.7%
Unexplned variance in 1st contrast =	2.4680	7.9%	12.3%	
Unexplned variance in 2nd contrast =	2.1621	7.0%	10.8%	
Unexplned variance in 3rd contrast =	1.9021	6.1%	9.5%	
Unexplned variance in 4th contrast =	1.7301	5.6%	8.7%	
Unexplned variance in 5th contrast =	1.6477	5.3%	8.2%	

Essential Unidimensionality (Rasch/Common variance) = 52.8%

Figure 3. Construct validity and unidimensionality

means there is a single prevailing construct measured consistently by the instrument. The percentage of variance accounted for by the Rasch measures indicates that the primary dimension is sufficiently dominant in the measurement structure to represent the intended construct. The values of unexplained variance in the first contrast (7.9) and second contrast (7.0) are rather insignificant, which means that there are no significant secondary dimensions (Li et al., 2024). In Rasch measurement, contrasts of this magnitude are usually treated as noise rather than as evidence of extra latent traits.

Essential Unidimensionality of 52.8% also contributes to the fact that the given instrument corresponds to the assumption of

unidimensionality in practice and can be used to compare misconception levels among students using a logit scale (Sukarelawan et al., 2021). This finding is especially significant for four-tier diagnostic instruments, as it appears that the addition of confidence tiers does not contaminate the construct but rather operates within a single dominant dimension. These findings are sound, but the consistency of the measurement structure should still be checked through cross-sample analysis.

Item Characteristic Curves

The probability patterns in Figure 4 show distinct differences across ability levels, indicating that the response scale is generally effective at

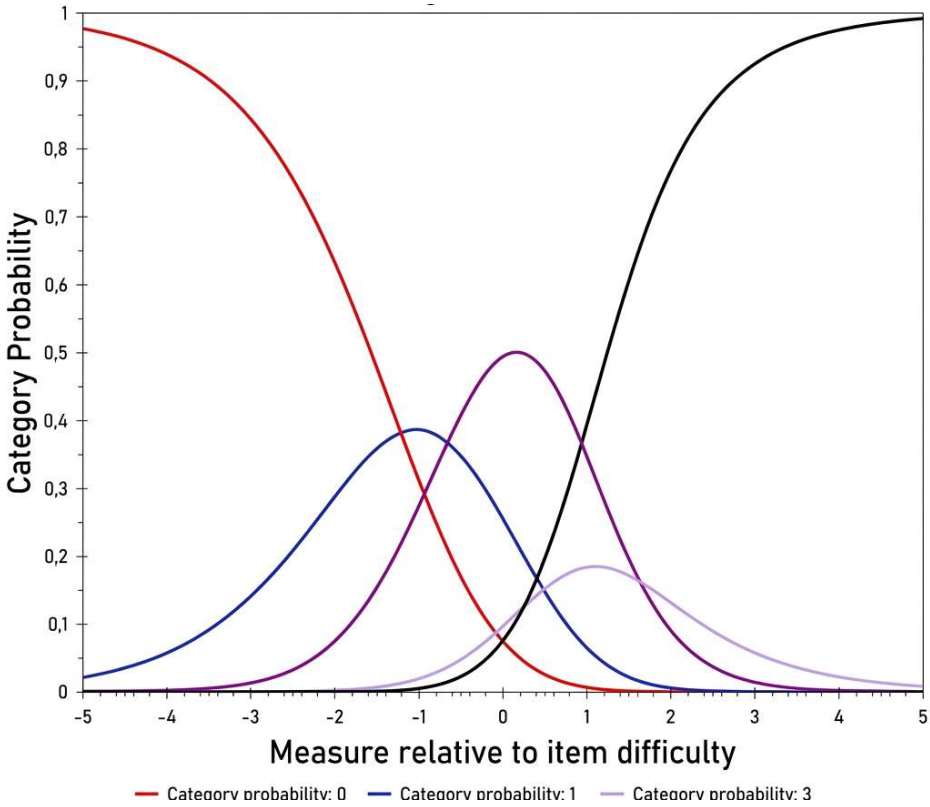


Figure 4. Item characteristic curves

distinguishing students' conceptual understanding and confidence. Category 0 relates to low ability, category 2 to medium ability, and category 4 to high ability. Categories 0, 1, 2, and 4 are clearly dominant across a range of abilities, while category 3 has a weaker, overlapping peak that is rapidly overrun by categories 2 and 4. This overlap indicates a lack of consistency for students in understanding the difference between intermediate confidence levels in holding partially correct conceptions. Nevertheless, the ordered Andrich thresholds and lack of irregularities suggest a logical, structured, and monotonic functioning scale along the latent trait continuum. Accordingly, the four-category scale is suitable for utilization within the Partial Credit Model, although category 3 does not work as efficiently and is maintained for more theoretical reasons, related to differentiating levels of student confidence (Oktaviani & Istiyono, 2023).

DIF Analysis

The findings of the gender DIF analysis in Figure 5 indicate that most items are near the zero

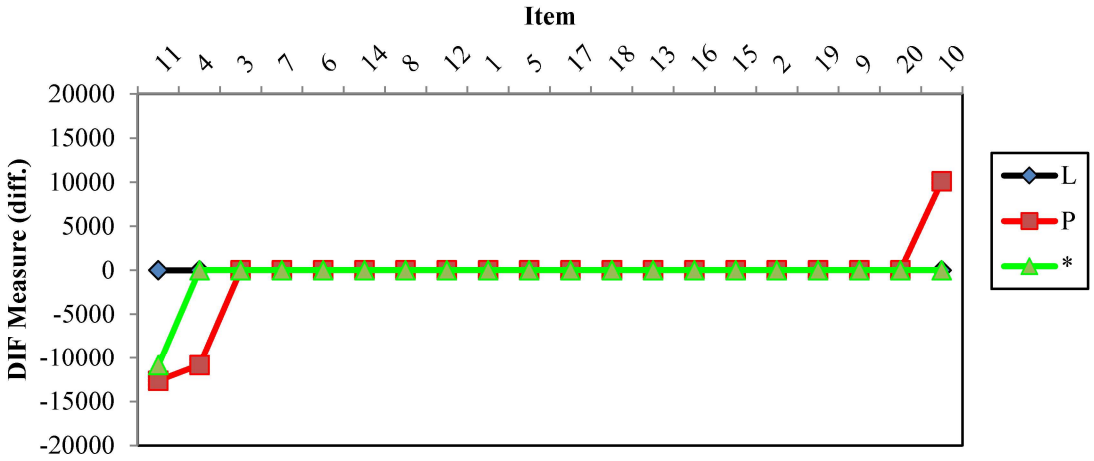
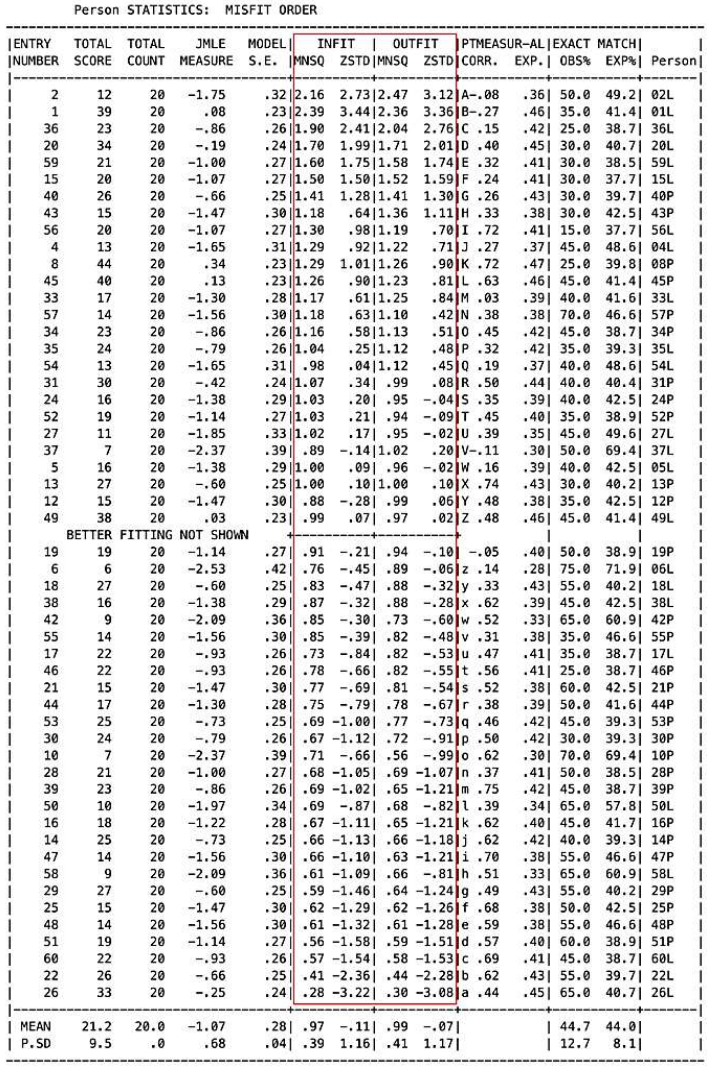


Figure 5. The result of DIF analysis based on gender (L: Male; P: Female)

line, suggesting no significant difference in difficulty between male and female students. The DIF values do not exceed the psychometrically acceptable threshold. This suggests that differences in student performance are more likely due to differences in conceptual understanding than gender related measurement bias. This is despite the fact that there are several other items with low DIF values, including Q11-Q3 and the change in Q10; the DIF values are within the psychometrically acceptable range (Farida et al., 2024). Such small patterns of DIF are often considered negligible and may reflect differences in context or instruction rather than systematic bias against a particular group. This is why the instrument can be regarded as comparatively unbiased with respect to gender; however, some amendments are suggested. However, these results provide useful information for item monitoring, as relatively large DIF values on items may warrant qualitative investigation in the future (Ozdemir & Alshamrani, 2020; Saputra et al., 2024; Steyn & de Bruin, 2020). Although DIF values remain within acceptable thresholds, small DIF tendencies may still be informative for diagnostic refinement. Items embedded in everyday contexts (e.g., kitchen, household materials, or tactile sensation) can be slightly sensitive to differences in prior experience or instructional emphasis, which may vary across student groups even when the underlying construct is the same. So, these items should be kept but monitored in future tests, and qualitative analysis can help determine whether their wording contributes to differences within small groups without indicating bias.

Person Fit Analysis

The person fit analysis in Figure 6 indicates that the pattern of responses from most respondents was as expected under the Rasch model, as evidenced by Infit and Outfit MNSQ values near 1.00 and ZSTD values near either -2 or +2 on average. This pattern suggests that most students responded to the items in a manner consistent with their estimated ability levels, supporting the overall validity of the instrument's person measures. Nevertheless, a few respondents exhibit misfit, as indicated by Outfit MNSQ > 1.5 or ZSTD > |2|. Typically, such misfit responses are those in which students get more difficult items right and easier ones wrong, and show irregular response behavior rather than random measurement error. For further clarification, please refer to Appendix B, which shows students' understanding of each test item. Appendix B presents the frequency distribution of students' conceptual understanding categories



Person STATISTICS: MISFIT 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%	Person
2	12	20	-1.75	.32	2.16	2.73	2.47	3.12	A -.08	.36	50.0	49.2	02L
1	39	20	.08	.23	2.39	3.44	2.36	3.36	B -.27	.46	35.0	41.4	01L
36	23	20	-.86	.26	1.90	2.41	2.04	2.76	C .15	.42	25.0	38.7	36L
20	34	20	-.19	.24	1.70	1.99	1.71	2.01	D .40	.45	30.0	40.7	20L
59	21	20	-1.00	.27	1.60	1.75	1.58	1.74	E .32	.41	30.0	38.5	59L
15	20	20	-1.07	.27	1.50	1.50	1.52	1.59	F .24	.41	30.0	37.7	15L
40	26	20	-.66	.25	1.41	1.28	1.41	1.30	G .26	.43	30.0	39.7	40P
43	15	20	-1.47	.30	1.18	.64	1.36	1.11	H .33	.38	30.0	42.5	43P
56	20	20	-1.07	.27	1.30	.98	1.19	.70	I .72	.41	15.0	37.7	56L
4	13	20	-1.65	.31	1.29	.92	1.22	.71	J .27	.37	45.0	48.6	04L
8	44	20	.34	.23	1.29	1.01	1.26	.90	K .72	.47	25.0	39.8	08P
45	40	20	.13	.23	1.26	.90	1.23	.81	L .63	.46	45.0	41.4	45P
33	17	20	-1.30	.28	1.17	.61	1.25	.84	M .03	.39	40.0	41.6	33L
57	14	20	-1.56	.30	1.18	.63	1.10	.42	N .38	.38	70.0	46.6	57P
34	23	20	-.86	.26	1.16	.58	1.13	.51	O .45	.42	45.0	38.7	34P
35	24	20	-.79	.26	1.04	.25	1.12	.48	P .32	.42	35.0	39.3	35L
54	13	20	-1.65	.31	.98	.04	1.12	.45	Q .19	.37	40.0	48.6	54L
31	30	20	-.42	.24	1.07	.34	.99	.08	R .50	.44	40.0	40.4	31P
24	16	20	-1.38	.29	1.03	.20	.95	-.04	S .35	.39	40.0	42.5	24P
52	19	20	-1.14	.27	1.03	.21	.94	-.09	T .45	.40	35.0	38.9	52P
27	11	20	-1.85	.33	1.02	.17	.95	-.02	U .39	.35	45.0	49.6	27L
37	7	20	-2.37	.39	.89	-.14	1.02	.20	V -.11	.30	50.0	69.4	37L
5	16	20	-1.38	.29	1.00	.09	.96	-.02	W .16	.39	40.0	42.5	05L
13	27	20	-.60	.25	1.00	.10	1.00	.10	X .74	.43	30.0	40.2	13P
12	15	20	-1.47	.30	.88	-.28	.99	.06	Y .48	.38	35.0	42.5	12P
49	38	20	.03	.23	.99	.07	.97	.02	Z .48	.46	45.0	41.4	49L
BETTER FITTING NOT SHOWN													
19	19	20	-1.14	.27	.91	-.21	.94	-.10	a -.05	.40	50.0	38.9	19P
6	6	20	-2.53	.42	.76	-.45	.89	-.06	z .14	.28	75.0	71.9	06L
18	27	20	-.60	.25	.83	-.47	.88	-.32	y .33	.43	55.0	40.2	18L
38	16	20	-1.38	.29	.87	-.32	.88	-.28	x .62	.39	45.0	42.5	38L
42	9	20	-2.09	.36	.85	-.30	.73	-.60	w .52	.33	65.0	60.9	42P
55	14	20	-1.56	.30	.85	-.39	.82	-.48	v .31	.38	35.0	46.6	55P
17	22	20	-.93	.26	.73	-.84	.82	-.53	u .47	.41	35.0	38.7	17L
46	22	20	-.93	.26	.78	-.66	.82	-.55	t .56	.41	25.0	38.7	46P
21	15	20	-1.47	.30	.77	-.69	.81	-.54	s .52	.38	60.0	42.5	21P
44	17	20	-1.30	.28	.75	-.79	.78	-.67	r .38	.39	50.0	41.6	44P
53	25	20	-.73	.25	.69	-1.00	.77	-.73	q .46	.42	45.0	39.3	53P
30	24	20	-.79	.26	.67	-1.12	.72	-.91	p .50	.42	30.0	39.3	30P
10	7	20	-2.37	.39	.71	-.66	.56	-.99	o .62	.30	70.0	69.4	10P
28	21	20	-1.00	.27	.68	-1.05	.69	-1.07	n .37	.41	50.0	38.5	28P
39	23	20	-.86	.26	.69	-1.02	.65	-1.21	m .75	.42	45.0	38.7	39P
50	10	20	-1.97	.34	.69	-.87	.68	-.82	l .39	.34	65.0	57.8	50L
16	18	20	-1.22	.28	.67	-1.11	.65	-1.21	k .62	.40	45.0	41.7	16P
14	25	20	-.73	.25	.66	-1.13	.66	-1.18	j .62	.42	40.0	39.3	14P
47	14	20	-1.56	.30	.66	-1.10	.63	-1.21	i .70	.38	55.0	46.6	47P
58	9	20	-2.09	.36	.61	-1.09	.66	-.81	h .51	.33	65.0	60.9	58L
29	27	20	-.60	.25	.59	-1.46	.64	-1.24	g .49	.43	55.0	40.2	29P
25	15	20	-1.47	.30	.62	-1.29	.62	-1.26	f .68	.38	50.0	42.5	25P
48	14	20	-1.56	.30	.61	-1.32	.61	-1.28	e .59	.38	55.0	46.6	48P
51	19	20	-1.14	.27	.56	-1.58	.59	-1.51	d .57	.40	60.0	38.9	51P
60	22	20	-.93	.26	.57	-1.54	.58	-1.53	c .69	.41	45.0	38.7	60L
22	26	20	-.66	.25	.41	-2.36	.44	-2.28	b .62	.43	55.0	39.7	22L
26	33	20	-.25	.24	.28	-3.22	.30	-3.08	a .44	.45	65.0	40.7	26L
MEAN	21.2	20.0	-1.07	.28	.97	-.11	.99	-.07			44.7	44.0	
P.SD	9.5	.0	.68	.04	.39	1.16	.41	1.17			12.7	8.1	

Figure 6. The result of person fit analysis

across the 20 exam questions. The results show that answers containing misconceptions were rarely found across all questions. The dominant categories among the 60 students were Partial Understanding (32 students) and No Understanding (24 students), followed by 2 students in the category of either No Understanding or Partial Understanding, 1 student in the category of either Misconception or Partial Understanding, 1 student in the Misconception category, 0 students in the Partial Negative category, and 0 students in the Solid Understanding category. The trend also follows when the students get challenging questions right and easy questions wrong, which is usually attributed to carelessness, guesswork, or certain misapprehensions (Rohima & Hasbiah, 2023). From a diagnostic standpoint, these misfit cases are especially informative because they may indicate non-linear reasoning and/or unstable conceptual understanding rather than inattentive responding. The importance of such findings lies in the fact that misfit persons can undermine the validity of the results interpretation and should establish the foundation on which a choice regarding the clustering or particular intervention types can be made. Instead of compromising measurement quality, these patterns of misfit

provide a basis for a specific pedagogical intervention or clustering strategy to address individual learning needs and misconceptions (Agustina et al., 2023).

To avoid redundancy, the person-fit interpretation is presented once, followed by a diagnostic illustration using response-pattern evidence from the scalogram in Figure 6. The response patterns of most participants were in line with the expectations of the Rasch model, as indicated by Infit and Outfit MNSQ values of approximately 1.00 and ZSTD values usually within the range of -2 to +2. Nonetheless, a few respondents had an Outfit MNSQ misfit of more than 1.5 or a ZSTD of more than 2. Students gave correct responses to challenging questions and gave incorrect answers to simple questions, which was frequently noticed when the latter were careless, guessing, or had a particular misconception (Rohima & Hasbiah, 2023). Results like these are highly significant since incorrect analysis of findings can be nullified by unqualified people. Moreover, such findings must be utilized to make decisions regarding a particular pedagogical intervention or clustering (Agustina et al., 2023).

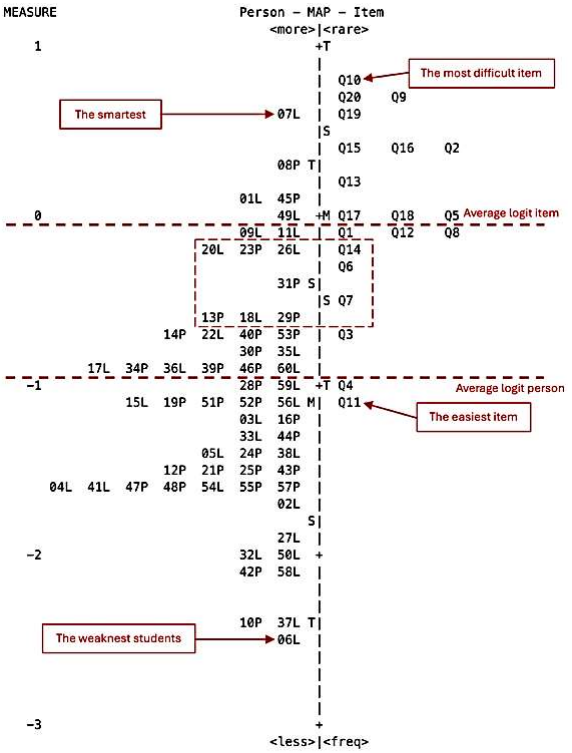


Figure 7. The wright map of students' abilities

Wright Map Targeting

Figure 7 shows that students' abilities range from +1 to -3 logits, with most students between -1 and 0 logits. This means that the items are mostly harder than average student ability and will, however, be usable as diagnostic tools. Such targeting is suitable for diagnostic purposes, as it is more about increasing sensitivity to partial understanding and emerging misconceptions than about picking out high-performing students. On the other hand, the item difficulty hierarchy may be considered stable and reliable, since items are distributed across the least challenging (Q11 and Q4) and the most challenging (Q10, Q20, and Q9). This ordered distribution suggests that the instrument spans a meaningful progression of conceptual difficulty in the domain of temperature and heat. The trend indicates the gaps at both ends of ability, which implies the necessity to include very easy and very difficult items, because that way the instrument will be able to record the variation in ability more equally, which is also supported by the Rasch instrument development literature (Hikmah et al., 2021; Sukarelawan et al., 2021). Such refinement would lead to better targeting precision, especially for students at the lower and upper ends of ability, thereby enhancing the instrument's diagnostic resolution.

Scatter Plot of Students Performance

Figure 8 shows a scatter plot illustrating the relationship between students' ability levels and item difficulty levels. The x-axis represents the item difficulty index, while the y-axis represents students' ability scores. Based on this visualization, the item difficulty index ranges from 3.13 to 4.27, while student ability scores range

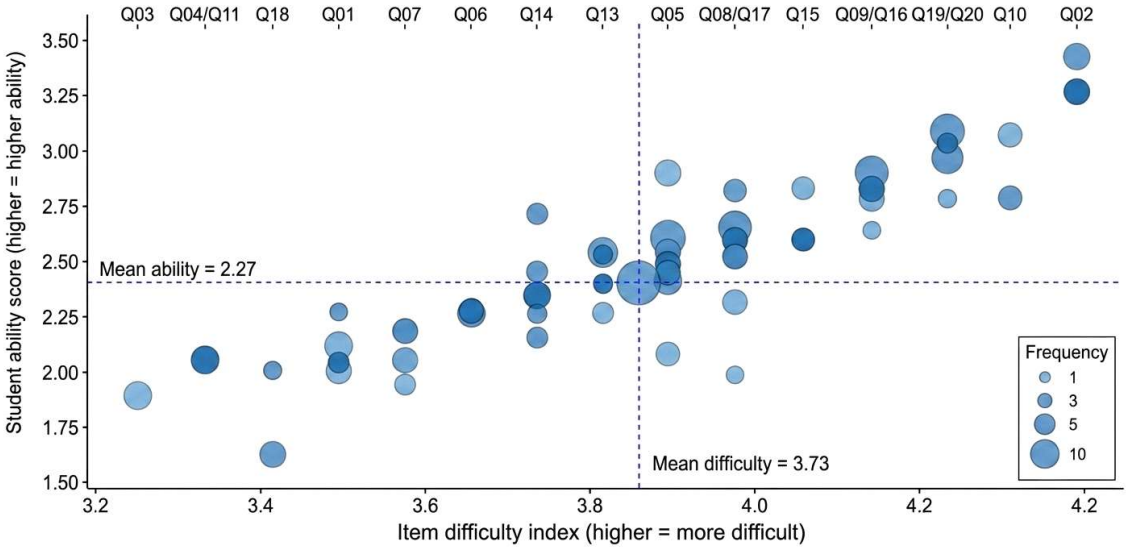
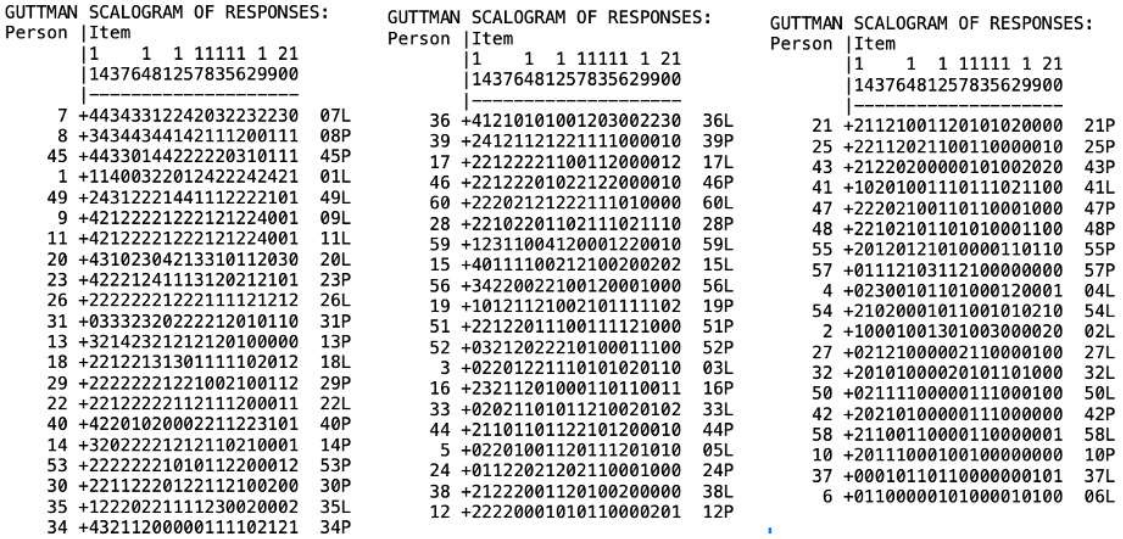


Figure 8. Scatter plot of students' ability levels and item difficulty levels

from 1.50 to 3.40. The average item difficulty of 3.73 is higher than the average student ability of 2.27. This indicates that most items are relatively difficult for the average student. Items Q02, Q09, and Q10 on temperature and expansion, and items Q15, Q16, Q19, and Q20 on heat transfer are the most difficult. In contrast, items with low difficulty levels on the temperature and expansion topic include Q01, Q03, Q04, and Q07. On the topic of the relationship between heat and an object's temperature and state, they include Q11 and Q14. In general, these results indicate that the topics of temperature and expansion are easier to understand for students at the low-to-moderate ability level, followed by the relationship between heat and the temperature and state of matter at the moderate level. In contrast, the topic of heat transfer demands higher conceptual ability. Thus, the scatter plot reveals a gap between students' abilities and the difficulty levels of some items. This can also be seen in Appendix A.

Response Patterns

In Figure 9, the Guttman Scalogram provides a full picture of student response patterns by item difficulty and shows that fit students exhibit monotonic response patterns, whereas misfit students exhibit non-monotonic patterns (Sideridis & Alghamdi, 2025; Wind, 2018). These irregularities can be explained by inconsistencies between students' expected and actual responses due to guessing, response laxity, or, more importantly, unstable and fragmented conceptual knowledge rather than random responses. This interpretation is consistent with person-fit statistics, in which misfit persons exhibit high Outfit MNSQ or ZSTD values outside the acceptable range (-2 to +2), suggesting violations of the model's item-difficulty assumptions. Methodologically, the Guttman Scalogram is a qualitative complement to the quantitative person fit indices, in which inappropriate person fit is associated systematically with patchy response structures (Wind, 2018; Yu et al., 2024). The overlap between scalogram patterns and person fit increases the validity of diagnostic inferences to support a strong triangulation basis for identifying responses that need reconsideration, items that might need revision, or student groups that might benefit from specific instructional remediation to provide a strong inference for informed decisions on item refinement and the



GUTTMAN SCALOGRAM OF RESPONSES:

Person	Item 1 1 1 11111 1 21	
	14376481257835629900	
7	+44343312242032232230	07L
8	+34344344142111200111	08P
45	+44330144222220310111	45P
1	+11400322012422242421	01L
49	+24312221441112222101	49L
9	+42122221222121224001	09L
11	+42122221222121224001	11L
20	+43102304213310112030	20L
23	+42221241113120212101	23P
26	+22222221222111121212	26L
31	+03332302222212010110	31P
13	+32142321212120100000	13P
18	+22122131301111102012	18L
29	+22222221221002100112	29P
22	+22122222112111200011	22L
40	+42201020002211223101	40P
14	+32022221212110210001	14P
53	+22222221010112200012	53P
30	+22112220122112100200	30P
35	+12220221111230020002	35L
34	+43211200000111102121	34P

GUTTMAN SCALOGRAM OF RESPONSES:

Person	Item 1 1 1 11111 1 21	
	14376481257835629900	
36	+41210101001203002230	36L
39	+24121121221111000010	39P
17	+22122221100112000012	17L
46	+22122201022122000010	46P
60	+22202121222111010000	60L
28	+22102201102111021110	28P
59	+12311004120001220010	59L
15	+40111100212100200202	15L
56	+34220022100120001000	56L
19	+10121112002101111102	19P
51	+22122011100111121000	51P
52	+03212022210100011100	52P
3	+02201221110101020110	03L
16	+23211201000110110011	16P
33	+02021101011210020102	33L
44	+21101101122101200010	44P
5	+02201001120111201010	05L
24	+01122021202110001000	24P
38	+21222001120100200000	38L
12	+22220001010110000201	12P

GUTTMAN SCALOGRAM OF RESPONSES:

Person	Item 1 1 1 11111 1 21	
	14376481257835629900	
21	+21121001120101020000	21P
25	+22112021100110000010	25P
43	+21220200000101002020	43P
41	+10201001110111021100	41L
47	+22202100110110001000	47P
48	+22102101101010001100	48P
55	+20120121010000110110	55P
57	+01112103112100000000	57P
4	+02300101101000120001	04L
54	+21020001011001010210	54L
2	+10001001301003000020	02L
27	+02121000002110000100	27L
32	+20101000020101101000	32L
50	+02111100000111000100	50L
42	+20210100000111000000	42P
58	+21100110000110000001	58L
10	+20111000100100000000	10P
37	+00010110110000000101	37L
6	+01100000101000010100	06L

Figure 9. Guttman scalogram of students' responses

design of group-specific intervention programs (Schooler, 1968; Yu et al., 2024).

With diagnostic follow-up, the erratic response patterns in the scalogram can be used to detect guessing and/or incomplete conceptual frameworks. For instance, a misfit response typically has several isolated high-category responses on difficult items and low-category responses on easy items. In a four-tier model, this can occur when the student successfully guesses Tier 1 but chooses an invalid reason (Tier 3) and/or is overconfident, producing a pattern that does not follow the expected monotonicity. These situations are interesting from a diagnostic perspective: they reveal which concepts elicit unstable reasoning and can inform a diagnostic session (interview, remediation tasks) to rule out "chance success".

■ CONCLUSION

This study shows that the Four-Tier Diagnostic Test (FTDT), analyzed using the Rasch Partial Credit Model, is effective at identifying students' misconceptions about heat and temperature. The most common conceptual understanding of temperature and heat among the 60 students was the "Partial Understanding" category, followed by the "No Understanding" category, then the "Misconception" category, the "Partial Negative" category, and finally the "Solid Understanding" category. The instrument proved highly reliable, clearly differentiated student proficiency levels, and showed no major gender bias; thus, it appears suitable for diagnostic purposes. The results are consistent with the research objectives and suggest alignment among the research context, methodology, and findings.

The calibrated FTDT can thus be used for teaching and remediation, as well as for curriculum development, to better address persistent conceptual misconceptions and develop effective teaching approaches. Larger sample sizes and the use of several thermodynamics topics and other qualitative methods (such as interviews or think-alouds) should be employed to reveal additional conceptual difficulties effective teaching strategies.

■ DECLARATION OF GENERATIVE AI USAGE IN THE WRITING PROCESS

The author used ChatGPT to help with editing this manuscript. The author has edited and reviewed the content and takes full responsibility for the content of the publication.

SUPPLEMENTARY MATERIAL

The dataset is openly available at: https://figshare.com/s/3572b61ecb783028f4c6

REFERENCES

Agustina, I., Astuti, D., Bhakti, Y. B., Prasetya, R., & Rahmawati, Y. (2023). Four tier-relativity diagnostic test (4T-RDT) to Identify Student Misconception. *JIPF (Jurnal Ilmu Pendidikan Fisika)*, *8*(1), 75–84. https://doi.org/10.26737/JIPF.V8I1.3668

Ali, K., Slade, A., Kay, E. J., Zahra, D., Chatterjee, A., & Tredwin, C. (2017). Application of Rasch analysis in the development and psychometric evaluation of the dental undergraduates' preparedness assessment scale. *European Journal of Dental Education*, *21*(4), e135–e141. https://doi.org/10.1111/eje.12236

Avinç, E., & Doðan, F. (2024). Digital literacy scale: Validity and reliability study with the rasch model. *Education and Information Technologies 2024 29:17*, *29*(17), 22895–22941. https://doi.org/10.1007/S10639-024-12662-7

Bandhu, D., Khadir, M., Kaushik, A., Sharma, S., Ali, H. A., & Jain, A. (2023). Innovative approaches to thermal management in next-generation electronics. *E3s Web of Conferences*, *430*. https://doi.org/10.1051/e3sconf/202343001139

Bezlutsky, V., Melnyk, O., Emelyanova, T., Lisnychiy, V., Lytvynenko, A., & Sova, S. (2024). Temperature loads in railway structures and general engineering practice. *Transport Means Proceedings of the International Conference*, *2024-Octob*, 581–586. https://doi.org/10.5755/e01.2351-7034.2024.P581-586

Boone, W. J., Yale, M. S., & Staver, J. R. (2014). Rasch analysis in the human sciences. In *Rasch Analysis in the Human Sciences*. Springer. https://doi.org/10.1007/978-94-007-6857-4

Busyairi, A., Munandar, R., Apsari, P. A. D., Wahyuni, A., Nurhasanah, Arni, K. J., Walihah, Z., & Diarta, M. H. (2022). Identification of prospective physics teacher's misconceptions of temperature and heat concept using the three tier test. *Amplitudo Journal of Science and Technology Innovation*, *1*(2), 48–53. https://doi.org/10.56566/amplitudo.v1i2.9

Caleon, I. S., & Subramaniam, R. (2010). Do students know What they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions. *Research in Science Education*, *40*(3), 313–337. https://doi.org/10.1007/s11165-009-9122

Çelikkanlý, N. Ö., & Kýzýlcýk, H. Þ. (2022). A review of studies about four-tier diagnostic tests in physics education. *Journal of Turkish Science Education*, *19*(4), 1291–1311. https://doi.org/10.36681/tused.2022.175

Dhyaaldian, S. M. A., Kadhim, Q. K., Mutlak, D. A., Neamah, N. R., Kareem, Z. H., Hamad, D. A., Tuama, J. H., & Qasim, M. S. (2022). A comparison of polytomous rasch models for the analysis of c-tests. *International Journal of Language Testing*, *12*(2), 107–117. https://doi.org/10.22034/IJLT.2022.157128

Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-Tier diagnostic test with certainty of response index on the concepts of fluid. *Journal of Physics Conference Series*, *1155*(1). https://doi.org/10.1088/1742-6596/1155/1/012078

Farida, F., Alamsyah, Y. A., Anggoro, B. S., Andari, T., & Lusiana, R. (2024). Rasch measurement validation of an assessment

tool for measuring students' creative problem-solving through the use of ict | validación de una herramienta de evaluación basada en el modelo de rasch para medir la resolución creativa de problemas en estudiantes. *Pixel Bit Revista De Medios Y Educación*, *71*, 83–106. https://doi.org/10.12795/pixelbit.107973

Fauzan, A., Ega, H. M., Sigalingging, J. A., & Nugroho, Y. S. (2021). Analysis of heat gains from flat plate heater measured using multi-axis heat flux sensors. *Evergreen*, *8*(4), 844–849. https://doi.org/10.5109/4742130

Fenditasari, K., Jumadi, Istiyono, E., & Hendra. (2020). Identification of misconceptions on heat and temperature among physics education students using four-tier diagnostic test. *Journal of Physics Conference Series*, *1470*(1). https://doi.org/10.1088/1742-6596/1470/1/012055

Hawkins, R. J., Hawkins, J., Tremblay, B., Wiles, L., & Higgins, K. (2024). Use of the rasch model for fit statistics and rating scale diagnosis for the student perception appraisal. *Journal of Nursing Measurement*, *32*(3), 391–403. https://doi.org/10.1891/JNM-2022-0122

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

Hermita, N., Suhandi, A., Syaodih, E., Samsudin, A., Isjoni, & Rosa, F. (2017). Assessing pre-service elementary school teachers' alternative conceptions through a four-tier diagnostic test on magnetism concepts. *Advanced Science Letters*, *23*(11), 10910–10912. https://doi.org/10.1166/asl.2017.10184

Hikmah, F. N., Sukarelawan, Moh. I., Nurjannah, T., & Djumati, J. (2021). Elaboration of high school student's metacognition awareness on heat and temperature material: wright map in rasch model. *Indonesian Journal of Science and Mathematics Education*, *4*(2), 172–182. https://doi.org/10.24042/IJSME.V4I2.9488

Ismail, M. S., Din, M. S. H., & Jusoh, M. S. (2021). Predictive modelling using Rasch's person-item map: Intrepreting and assessing in Malaysia manufacturing firms. *AIP Conference Proceedings*, *2347*. https://doi.org/10.1063/5.0052973

Istiyono, E. (2022). Diagnostic tests as an important pillar in today's physics learning: four-tier diagnostic test a comprehensive diagnostic test solution. *Journal of Physics Conference Series*, *2392*(1). https://doi.org/10.1088/1742-6596/2392/1/012001

Istiyono, E., Fenditasari, K., Ayub, M. R. S., Saepuzaman, D., & Dwandaru, W. S. B. (2024). An eight-category partial credit model as very appropriate for four-tier diagnostic test scoring in physics learning. *AIP Conference Proceedings*, *2622*(1). https://doi.org/10.1063/5.0133862

Istiyono, E., Shanti, M. R. S., Saepuzaman, D., Dwandaru, W. S. B., & Zakwandi, R. (2022). A four tier web based assessment with eight categories diagnostic. *Proceedings International Conference on Education and Technology ICET*, *2022-Octob*, 190–196. https://doi.org/10.1109/ICET56879.2022.9990847

Istiyono, E., Sunu Brams Dwandaru, W., 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), 371–385. https://doi.org/10.12973/eu-jer.12.1.371

Jumadi, J., Sukarelawan, M. I., & Kuswanto, H. (2023). An investigation of item bias in the four-tier diagnostic test using Rasch model. *International Journal of Evaluation and Research in Education*, *12*(2), 622–629. https://doi.org/10.11591/ijere.v12i2.22845

Kaltakci-Gurel, D., Eryilmaz, A., & McDermott, L. C. (2017). Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics. *Research in Science and Technological Education*, *35*(2), 238–260. https://doi.org/10.1080/02635143.2017.1310094

Li, X., Su, Z., Wang, L., Li, J., & Diao, Y. (2024). Psychometric evaluation of the pictorial scale of perceived movement skill competence in Chinese children: An item response theory analysis. *Child Care Health and Development*, *50*(4). https://doi.org/10.1111/cch.13275

Munggarani, M. E., Supriyati, Y., & Astra, I. M. (2021). Identifying high school students' misconceptions using digital four-tier diagnostic tests in distance learning. *Journal of Physics Conference Series*, *2019*(1). https://doi.org/10.1088/1742-6596/2019/1/012016

Nabilah, F. N., Maknun, J., Muslim, M., Samsudin, A., Hasanah, L., & Suhandi, A. (2019). Eleventh-grade student's conceptions about temperature and heat. *Journal of Physics Conference Series*, *1280*(5). https://doi.org/10.1088/1742-6596/1280/5/052067

Oktaviani, N., & Istiyono, E. (2023). Application of partial model in practical performance assesment of physics learning. *AIP Conference Proceedings*, *2556*. https://doi.org/10.1063/5.0130724

Ozdemir, B., & Alshamrani, A. H. (2020). Examining the fairness of language tests across gender with IRT-based differential item and test functioning methods. *International Journal of Learning Teaching and Educational Research*, *19*(6), 27–45. https://doi.org/10.26803/ijlter.19.6.2

Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research*, *15*(2). https://doi.org/10.1103/PhysRevPhysEducRes.15.020111

Rohima, I. E., & Hasbiah, A. W. (2023). Rasch analysis of household waste reduction behavior: case study in sunten jaya, bandung barat regency. *Journal of Community Based Environmental Engineering and Management*, *7*(2), 111–118. https://doi.org/10.23969/JCBEEM.V7I2.10383

Romine, W. L., Schaffer, D. L., & Barrow, L. (2015). Development and application of a novel rasch-based methodology for evaluating multi-tiered assessment Instruments: Validation and utilization of an undergraduate diagnostic test of the water cycle. *International Journal of Science Education*, *37*(16), 2740–2768. https://doi.org/10.1080/09500693.2015.110539

Salele, N., Khan, 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*, *13*, 59886–59910. https://doi.org/10.1109/ACCESS.2025.3555432

Samsudin, A., Azizah, N., Fratiwi, N. J., Suhandi, A., Irwandani, I., Nurtanto, M., Yusup, M., Supriyatman, S., Masrifah, M., Aminudin, A. H., & Costu, B. (2024). Development of DIGaKiT: identifying students'

alternative conceptions by Rasch analysis model. *Journal of Education and Learning (EduLearn)*, *18*(1), 128–139. https://doi.org/10.11591/EDULEARN.V18I1.20970

Saputra, A., & Tania, L. (2024). An Indonesian-version of the short attitude toward mathematics inventory: a voice from pre-service chemistry teacher. *International Journal of Evaluation and Research in Education (IJERE)*, *13*(3), 1908–1916. https://doi.org/10.11591/IJERE.V13I3.27718

Schooler, C. (1968). A Note of Extreme Caution on the Use of. *Source: American Journal of Sociology*, *74*(3), 296–301.

Steyn, R., & de Bruin, G. P. (2020). An investigation of gender-based differences in assessment instruments: A test of measurement invariance. *SA Journal of Industrial Psychology*, *46*. https://doi.org/10.4102/sajip.v46i0.1699

Sukarelawan, M. I., Sriyanto, S., Puspitasari, A. D., Sulisworo, D., & Hikmah, U. N. (2021). Four-Tier Heat and Temperature Diagnostic Test (4T-HTDT) to Identify Student Misconceptions. *JIPFRI (Jurnal Inovasi Pendidikan Fisika Dan Riset Ilmiah)*, *5*(1), 1–8. https://doi.org/10.30599/JIPFRI.V5I1.856

Walfridsson, U., Walfridsson, H., Middeldorp, M. E., Sanders, P., & Årestedt, K. (2022). Validation of the English version of the arrhythmia-specific questionnaire in tachycardia and arrhythmia (ASTA): a Rasch evaluation study. *Journal of Patient Reported Outcomes*, *6*(1). https://doi.org/10.1186/s41687-022-00493-4

Wind, S. A. (2018). Using Guttman errors to explore rater fit in rater-mediated performance assessments. *Methodologica Innovations*, *11*(3). https://doi.org/10.1177/2059799118814396

Xiao, X., Xue, M., & Cheng, Y. (2023). Bayesian partial credit model and its applications in science education. In *Contemporary Trends and Issues in Science Education* (Vol. 57). https://doi.org/10.1007/978-3-031-28776-3_4

Yang, F., Lu, H., Liu, W., & Zhou, H. (2020). Understanding the contributions of microscopic heat transfer to thermal conductivities of liquid aldehydes and ketones by molecular dynamics simulation. *Journal of Chemical Information and Modeling*, *60*(6), 3022–3029. https://doi.org/10.1021/acs.jcim.0c00184

Yu, X., Tang, Q., Qin, C., & Li, Y. (2024). Person-Fit in cognitive diagnostic assessment. *Journal of Psychological Science*, *47*(3), 744–751. https://doi.org/10.16719/j.cnki.1671-6981.20240329

APPENDIX

Appendix A. The conception category

Items	Mean coding	Difficulty index	SU	PU	PN	MC	NU	Dominant
Q01	2.6	3.4	4	40	0	0	16	PU
Q02	1.7333	4.2667	1	18	1	2	38	NU
Q03	2.8667	3.1333	1	52	2	0	5	PU
Q04	2.8333	3.1667	6	41	1	2	10	PU
Q05	2.25	3.75	3	21	10	1	25	NU
Q06	2.5333	3.4667	1	37	7	0	15	PU
Q07	2.5667	3.4333	3	28	13	0	16	PU
Q08	2.1333	3.8667	3	25	3	0	29	NU

Q09	1.9333	4.0667	1	19	6	2	32	NU
Q10	1.8	4.2	0	17	7	0	36	NU
Q11	2.8333	3.1667	10	34	1	0	15	PU
Q12	2.2833	3.7167	1	30	5	3	21	PU
Q13	2.2667	3.7333	0	13	24	2	21	PN
Q14	2.3	3.7	0	30	9	0	21	PU
Q15	2.0667	3.9333	0	24	7	2	27	NU
Q16	1.9333	4.0667	0	24	3	2	31	NU
Q17	2.1333	3.8667	0	27	6	2	25	PU
Q18	2.7	3.3	1	9	40	0	10	PN
Q19	1.8167	4.1833	2	19	0	3	36	NU
Q20	1.8167	4.1833	0	20	4	1	35	NU

Appendix B. Students' conceptual understanding of each question

Person	Gender	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08	Q09	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
01	M	PU	SU	SU	PU	PU	NU	NU	PU	SU	NU	NU	NU	MC	PN	MC	MC	MC	SU	MC	MC
02	M	PU	NU	NU	NU	NU	PN	NU	NU	NU	NU	NU	PU	NU	NU	PU	NU	NU	NU	NU	PU
03	M	PU	PU	PU	PU	PU	PN	NU	PU	PU	NU	NU	PN	NU	PN	NU	NU	NU	PN	NU	PN
04	M	PU	MC	PU	MC	NU	NU	NU	NU	NU	PU	NU	NU	NU	PU	NU	PN	PN	NU	NU	NU
05	M	PU	NU	PU	PU	PU	PU	NU	NU	NU	NU	NU	PU	PU	NU	PU	PU	NU	PN	PU	PU
06	M	NU	PN	PU	PU	NU	NU	NU	NU	PU	NU	NU	PN	NU	NU	NU	NU	PN	NU	NU	NU
07	M	PU	PU	PU	SU	SU	PU	SU	NU	MC	NU	SU	MC	PN	PU	MC	MC	MC	NU	MC	PU
08	F	SU	NU	PU	SU	SU	SU	SU	SU	PU	PU	PU	PU	PU	PU	PN	PU	PU	PN	NU	PU
09	M	PU	PU	PN	PU	PU	PU	PU	PU	NU	PU	SU	PU	PU	PU	PN	PU	PU	PU	SU	NU
10	F	NU	NU	PU	NU	NU	PU	PN	NU	NU	NU	PU	PU	NU	NU	NU	NU	NU	PN	NU	NU
11	M	PU	PU	PN	PU	PU	PU	PU	PU	NU	PU	SU	PU	PU	PU	PN	PU	PU	PU	SU	NU
12	F	PU	NU	PU	PU	NU	NU	PU	NU	PU	PN	PU	NU	PN	NU	NU	NU	NU	PN	NU	NU
13	F	PU	NU	PU	PU	PN	PU	SU	PU	NU	NU	PU	MC	PN	PU	NU	PU	PU	PN	NU	NU
14	F	PU	PU	NU	PU	PN	PU	PU	PU	NU	PN	PU	PU	PU	PU	NU	PU	PU	PN	NU	NU
15	M	NU	NU	PU	NU	PU	PU	PN	NU	PU	PN	SU	MC	NU	PU	NU	PU	PU	PN	NU	NU
16	F	PU	NU	PU	PU	NU	PU	PN	NU	NU	PN	PU	NU	PN	PU	NU	PU	NU	PN	NU	PU
17	M	PU	NU	PU	PU	NU	PU	PU	PU	NU	PU	PU	PU	PN	PU	PU	NU	NU	PN	NU	PU
18	M	PU	NU	PU	PU	NU	PU	PU	PU	NU	PU	PU	PU	PN	PU	PU	PU	PN	PN	PU	PN
19	F	PU	PU	PU	NU	NU	PU	PU	PU	PU	PU	PU	NU	NU	PU	PU	PU	PU	PU	PU	NU
20	M	SU	NU	PU	PU	NU	PU	NU	NU	NU	NU	SU	PU	PU	PU	NU	NU	PU	PN	PU	PU
21	F	PU	PU	PU	PU	PU	PU	PU	NU	NU	NU	PU	PU	NU	NU	PN	NU	NU	PN	NU	NU
22	M	PU	NU	PU	PU	PN	PU	PU	PU	NU	PU	PU	PU	PN	PU	PN	PU	PU	PN	NU	PU
23	F	PU	NU	PU	PU	PU	PU	PU	SU	PU	PN	SU	NU	PN	PU	NU	PU	PU	PN	PU	NU
24	F	PU	NU	PU	PU	NU	PU	PU	PU	NU	NU	NU	PU	PN	NU	NU	NU	PU	PU	PU	NU
25	F	PU	NU	PU	PU	NU	PU	PN	PU	NU	NU	PU	PU	PN	NU	NU	NU	NU	PN	NU	PU
26	M	PU	PU	PU	PU	PU	PU	PU	PU	PU	PU	PU	PU	PU	PU	PU	PU	PU	PN	PU	PU
27	M	NU	NU	PU	PU	NU	PU	PU	NU	PN	NU	NU	NU	PN	NU	NU	NU	PU	PU	NU	NU
28	F	PU	PU	PU	PU	NU	PU	NU	NU	PU	NU	PU	PU	PN	PU	PU	NU	PU	PN	PU	PN
29	F	PU	NU	PU	PU	PU	PU	PU	PU	PU	PU	PU	PU	NU	PU	PU	PU	PU	NU	NU	PU
30	F	NU	NU	PU	PU	PU	PU	PN	PU	PU	NU	PU	PN	PN	PU	PU	PU	PU	PN	NU	NU
31	F	NU	NU	PU	PU	PU	PU	PU	PU	PU	NU	NU	PU	PN	PU	PU	NU	PU	PN	NU	NU
32	M	NU	NU	PU	NU	PU	PN	NU	NU	NU	NU	PU	NU	NU	NU	PU	PU	NU	PN	PU	NU
33	M	PU	PU	NU	PU	PN	PN	PU	NU	PN	PU	NU	NU	PU	PU	NU	NU	PU	PU	NU	NU
34	F	NU	NU	PU	PU	NU	NU	PN	NU	PU	PU	SU	NU	PN	PU	PU	PN	NU	PN	PU	PU
35	M	PU	PU	PU	PU	PN	NU	PU	PU	NU	PU	PN	NU	PU	PU	NU	NU	PU	PN	NU	NU
36	M	PU	NU	PU	NU	NU	NU	PN	NU	MC	NU	SU	NU	NU	NU	PU	NU	PN	PN	MC	PN
37	M	NU	NU	NU	NU	PU	NU	PN	PN	PU	PN	NU	PU	NU	PN	NU	NU	NU	NU	NU	NU
38	M	PU	NU	PU	PU	PU	PU	PU	NU	NU	NU	PU	PU	NU	NU	NU	PU	NU	PN	NU	NU
39	F	PU	NU	PU	SU	PU	PN	PU	PU	NU	NU	PU	PU	PN	PN	PU	NU	PU	PN	NU	PU
40	F	NU	PU	PU	PU	NU	PU	NU	PU	PU	PU	SU	NU	PN	NU	PU	PN	PU	PN	PU	NU
41	M	PU	PU	PU	NU	PN	PU	NU	NU	PN	NU	PU	PU	PN	NU	PU	NU	NU	PU	PU	NU
42	F	NU	NU	PU	NU	NU	NU	PU	NU	NU	NU	PU	NU	PN	PN	PU	NU	NU	PN	NU	NU
43	F	NU	NU	PU	PU	NU	NU	PU	NU	NU	NU	PU	NU	NU	PU	PU	NU	NU	PN	PU	PU
44	F	PU	NU	PU	PU	PU	PN	NU	NU	NU	NU	PU	PU	NU	PN	PU	PU	PU	PN	NU	PU
45	F	SU	NU	PU	SU	PU	NU	PN	SU	PU	PU	SU	PN	MC	NU	NU	PU	PU	PU	NU	PU
46	F	PU	NU	PU	PU	PU	PU	PU	NU	NU	NU	PU	NU	PU	PU	PU	NU	PU	PN	NU	PU
47	F	NU	NU	PU	PU	PN	PU	NU	NU	NU	NU	PU	PU	PN	PU	NU	NU	NU	PN	PU	NU
48	F	PU	NU	PU	PU	NU	PU	NU	NU	PU	NU	PU	PU	PN	PN	NU	NU	PN	NU	PU	NU
49	M	PU	PU	PU	SU	SU	PU	PN	PU	PU	PU	PU	SU	PU	PU	PU	PU	PU	PU	PU	NU
50	M	NU	NU	PU	PU	NU	PN	PN	NU	PN	NU	NU	NU	PN	PN	PN	NU	NU	PN	NU	NU
51	F	PU	PU	PU	PU	NU	PU	PU	PN	NU	NU	PU	PU	NU	NU	PU	PU	NU	PN	PU	NU
52	F	PU	NU	PU	PU	NU	PU	PU	PU	PN	NU	NU	PU	NU	NU	NU	NU	NU	PN	PU	NU
53	F	PU	NU	PU	PU	PN	PU	PU	PU	NU	PU	PU	NU	PU	PU	PU	PU	NU	PN	NU	PU
54	M	PU	PU	NU	PU	PN	NU	PU	NU	PU	NU	PU	NU	NU	NU	PN	NU	PN	NU	NU	PU
55	F	PU	PU	PU	NU	PU	NU	PU	PU	PN	NU	PU	NU	NU	PU	NU	PU	NU	NU	NU	PU
56	M	NU	NU	PU	SU	NU	NU	PU	PU	NU	NU	PU	PU	PN	NU	NU	NU	NU	PN	PU	NU
57	F	PU	NU	PU	PN	PN	PU	PN	NU	NU	NU	NU	PN	NU	NU	NU	NU	PU	PN	NU	NU
58	M	NU	NU	PU	PU	NU	NU	NU	PN	NU	PN	PU	NU	PN	PU	NU	NU	NU	PN	NU	NU
59	M	SU	MC	PU	MC	MC	PU	PN	NU	NU	NU	NU	PU	NU	NU	NU	PU	NU	NU	NU	NU
60	M	PU	PU	PU	PU	PU	PU	NU	PU	NU	NU	PU	PU	PU	PN	PU	NU	PU	PN	NU	NU