Development of Formative Assessment Instruments Misconception Check to Analyze the Conception of Thermodynamics in High School Students

Clarinta Alvani[1], Ridwan Efendi[1]*, & Rizki Zakwandi[1]
[1]Physics Education Study Program, Universitas Pendidikan Indonesia, Indonesia
*Corresponding author: ridwanefendi@upi.edu

Received: 15th September 2025; **Accepted**: 27th November 2025; **Published**: 8th December 2025
DOI: https://dx.doi.org/10.29303/jpft.v11i2.10220

Abstract – *The demands of formative assessment in the kurikulum merdeka should ideally be able to diagnose students' conceptual patterns in depth. However, common assessment practices are still limited to instruments that only measure correct or incorrect answers, so they cannot reveal conceptual understanding. The gap between these demands and reality has led to the development of more targeted instruments, especially for complex thermodynamics material. This study aims to develop a formative assessment instrument in the form of a misconception check to analyze high school students' conceptions of thermodynamics. The method used is quantitative with an instrument development approach based on the Mardapi model. The developed instrument is a formative assessment tool in the form of a misconception check with a multiple-choice format, with answer options designed to represent various categories of conceptions. The instrument was tested on 262 students from three high schools, and its validity was evaluated through content validity, construct validity, and readability tests. After a series of evaluations, it was found that 3 items were invalid and were eliminated, leaving 33 items that were suitable for use. This instrument has a unidimensionality value of 21%, an average Aiken's V coefficient of 0.96, and reliability of 0.92. The results of the study indicate that the developed formative misconception check assessment instrument is feasible and effective for analyzing the conceptions held by high school students on thermodynamics material. Therefore, this instrument can help teachers specifically analyze students' conceptions and design targeted learning.*

Keywords: *Formative Assessment; Misconception Check; Conception; Thermodynamics.*

INTRODUCTION

The success of education depends heavily on three main pillars: curriculum, learning, and assessment(Aditomo, 2024) . The Merdeka Curriculum emphasizes the integration of learning and assessment, placing formative assessment as an integral part of the learning cycle. This approach is in line with strategies such as *Teaching at the Right Level (*TaRL) and *Backward Design* (Wiggins & McTighe, 2005) , which prioritize the achievement of learning objectives and assessment adjustments to ensure that all students achieve a deep understanding of concepts. So far, educational assessment has focused more on summative assessment (*Assessment of Learning)* to measure the final learning outcomes. However, the Merdeka Curriculum encourages a paradigm shift to formative assessment (*Assessment as Learning)* that is oriented towards providing feedback and continuous improvement of the learning process (Schuwirth & Van Der Vleuten, 2011) .

In learning, especially physics, students build new knowledge based on their experiences and understanding (Kiray & Simsek, 2021) . This understanding is referred to as conception (Dewi & Ibrahim, 2019) . However, misconceptions often occur, namely discrepancies between individual understanding and scientific concepts (Saputri et al., 2021) . Therefore, effective formative assessment must be able to reveal students' conceptions so that educators can design appropriate learning strategies (Aufschnaiter & Alonzo, 2018).

This stage is in line with the scientific concept (Saputri et al., 2021) .

The importance of formative assessment has been recognized in theory, but in practice, there is a significant gap. In fact, the formative assessments conducted by teachers are not yet optimal. Interviews with physics teachers and direct observations show that formative assessments are often only conducted orally in class or through homework assignments without in-depth discussion. A study by (Suherly et al., 2023) shows that only 40% of teachers conduct formative assessments in the form of quizzes or assignments, and only 20% provide feedback to students. Other research results also indicate that teachers do not yet have a complete understanding of the requirements of the Merdeka Curriculum, as well as difficulties in designing assessment instruments (Liliawati et al., 2022) .

Commonly used assessment instruments have limitations. Conventional multiple-choice tests often only measure correct or incorrect answers, without recognizing patterns of errors or misconceptions (Chandrasegaran et al., 2007) . The journal Bhaw et al. (2024) also highlights the weakness of conventional multiple-choice questions, namely the lack of effectiveness of distractors, which can make questions too difficult or unreliable. The conventional scoring system (*dichotomous scoring)* only gives a score of 1 for correct answers and 0 for incorrect or unanswered questions. The main weakness of this system is that it cannot accommodate the partial knowledge that students may have (Burfitt, 2017) .

Meanwhile, essay tests, although effective in revealing misconceptions as stated by Resbiantoro et al.(2022) , are impractical to implement on a large scale because they require a long time to assess (Sadler, 1998) . As a result, students are often assessed as lacking creativity and unable to analyze physics concepts because educators only rely on questions from textbooks (Wulandari et al., 2023) . This limitation hinders educators in identifying students' conceptions and misconceptions, even though mastery of correct conceptions is crucial in the Merdeka Curriculum, especially in physics subjects such as thermodynamics, which has many applications in everyday life.

To address this gap, this study aims to develop a more effective formative assessment instrument. Referring to *the Classroom Assessment Techniques (*CAT) concept proposed by Cross & Angelo, the misconception check instrument can be a solution. This method is specifically designed to reveal common misconceptions among students. Previous research by Holbeck et al. (2014) shows that the use of *misconception checks* can improve online learning and provide better information for educators.

Although previous studies have identified various tools used to analyze misconceptions (Resbiantoro et al., 2022) and demonstrated the effectiveness of assessment, there are still gaps in the development of practical, informative instruments that can be used to analyze concepts in depth. This study attempts to fill this gap by developing a formative *misconception check* assessment instrument in a partial multiple-choice format specifically designed to analyze student conceptions.

Partial multiple choice in assessment method reviews, Frary (1989) reported a method in which choices are weighted and students receive scores according to their choices. Students learn several aspects of a concept before becoming fully competent and can be described as having partial knowledge of the concept.

The answer choices in this instrument not only serve as distractors, but are also designed to present various types of conceptions that students may have. Furthermore, this study will categorize student conceptions into five levels, namely: *scientific conception, almost scientific conception, misconception, lucky guess,* and *non-understanding of a* concept-(Derya Kaltakci, 2012; Jannah & Rahmi, 2020; Kiray & Simsek, 2021) . The development of this formative *misconception check* assessment instrument is expected to provide a practical yet informative tool for educators to identify and address students' conceptions more effectively.

RESEARCH METHODS

This study adopted a quantitative method with an instrument development approach that refers to the Mardapi model (Mardapi, 2020) . The aim was to create a *misconception check-type* formative assessment instrument to analyze the concepts of senior high school (SMA) students on thermodynamics material. This development procedure involved several steps, namely: (1) Compiling test specifications, (2) Writing test questions, (3) Reviewing test questions, (4) Conducting test trials, (5) Analyzing test items, (6) Revising test items, and (7) Assembling the test as shown in the following Figure 1.

The research participants consisted of 262 students from three high schools in Bandung City who were selected using *stratified random sampling* based on their 2024 new student admission report card (PPDB) scores, which were high, medium, and low. The sample demographics are presented in Table 1.

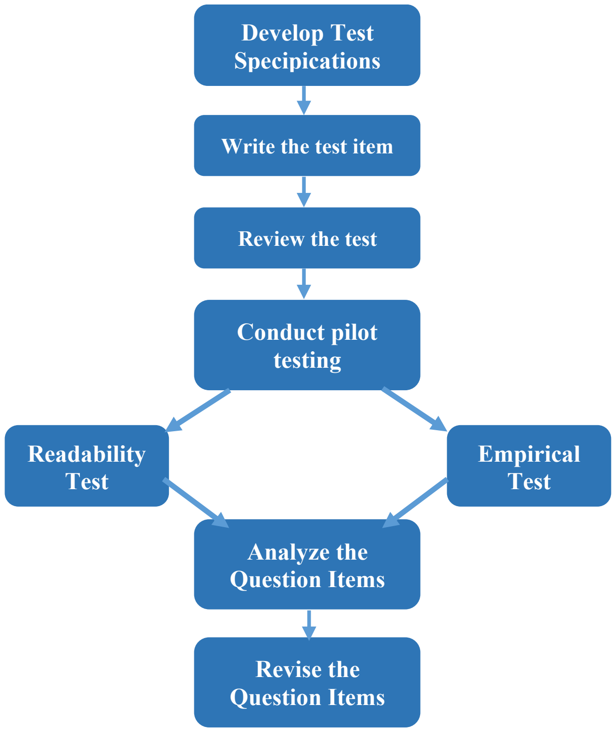


Figure 1. Research Design Flowchart

Table 1. Sample Demographics

No	Aspect	High School Grade		
		High	Middle	Low
1.	Gender			
	Boys	46	30	67
	Girls	21	41	57
2.	Ages (Years)			
	16-17	67	71	-
	17-18	-	-	124

The data collection procedure was carried out through instrument testing after undergoing expert validation and readability testing. Data analysis was conducted quantitatively to evaluate the feasibility and effectiveness of the instrument. The analysis included content validity and readability testing using Aiken's V Index, with the following formula:

$$V = \frac{\Sigma s}{n(c-1)} \quad (1)$$

$$s = r - l_0 \quad (2)$$

Explanation:
V = validity coefficient
n = number of validators
c = highest rating
r = score given by validators
l_0 = lowest score

The validity coefficient (V) value obtained from the subsequent calculation will be interpreted by matching it to the Aiken's V index table. In this study, there were 5 validators with a rating category of 1-5 (5 categories), so the validity coefficient (V) value must be V > 0.80 to be considered valid with a p value of 0.040 or a 40% *error* probability. And for the readability test based on the number of validators and the probability *of error (*p). In this study, there were 11 validators with assessment categories 1-5 (5 categories), so the validity coefficient (V) value must be V > 0.70 to be considered valid with a p value of 0.035 or a 35% probability of *error*.

Empirical data analysis uses the *Rasch* model with *Winsteps* software. This *Rasch* model analysis includes a unidimensionality test, a reliability test *(item reliability)*, an item quality test *(category function, item polarity,* and *item fit),* and a reliability test *(item reliability).* The results will be interpreted in the following table:

The unidimensionality test or prerequisite test is used to ensure that the test measures what it is supposed to measure. The results are interpreted based on the raw variance explained by the measure.

Table 2. Unidimensionality Value Criteria

Raw variance explained by measure (%)	**Criteria**
20 < *Rve* ≤ 40	Met
40 < *Rve* ≤ 60	Suitable
60 < *Rve* ≤ 100	Excellent

(Sumintono & Widhiarso, 2015)

Table 3. Criteria for Unexplained Variance in Contrast

Unexplained Variance in Contrast (%)	**Criteria**
< 3	Exceptional
3 – 5	Very Good
5 – 10	Good
10 -15	Fair
> 15	Good

(Sumintono & Widhiarso, 2015)

To reinforce the unidimensionality test results, the analysis was also reviewed using two additional indicators. Category Function was used to ensure that each answer choice functioned effectively in distinguishing ability levels. Meanwhile, Correlation Order verified the suitability between the difficulty level of the questions and the correlation of student abilities, which reinforced the overall validity of the instrument.

After conducting the prerequisite test, a reliability test was conducted to measure the consistency and reliability of the test results. This test produced Person Reliability, Item Reliability, and Cronbach Alpha (KR-20).

Table 4. Interpretation of Reliability Test

Statistics	**Index Value**	**Criteria**
Item and Pearson Reliability	< 0.67	Low
	0.67 – 0.80	Moderate
	0.81–0.90	Good
	0.91 – 0.94	Very Good
	> 0.94	Very Good
Cronbach Alpha (KR-20)	< 0.50	Low
	0.50 – 0.60	Moderate
	0.61–0.70	Good
	0.70 – 0.80	High
	> 0.80	Very High

In addition to reliability indicators, Rasch Model analysis also displays the Separation value. This value is important because it shows the instrument's ability to distinguish the level of difficulty of the items. The higher the Separation value, the better the instrument is at identifying groups of items. The number of groups identified can be calculated using a formula.

$$H = \frac{[(4 \times separation) + 1]}{3} \quad (3)$$

A validity test is conducted for each item to assess its quality. This test is obtained from the item fit order and can be seen from the outfit mean square (MNSQ) value, outfit Z-Standard (ZSTD), and point measure correlation (PT Measure Corr).

Table 5. *Item Fit* Criteria

Indicator	Acceptable Values
Outfit MNSQ	0.5 < MNSQ < 1.5
ZSTD *Outfit*	-2.0 < ZSTD < +2.0
Pt *Measure Corr*	0.4 < Pt *Measure Corr* < 0.85

(Sumintono & Widhiarso, 2015)

The results of each criterion are then interpreted based on the *fit-statistic* value criteria according to Sumintono & Widhiarso (2015) in Table 6 below.

Table 6. Interpretation *of Fit-Statistic Item*

Criteria	Description
All three indicators are met	Very Suitable
Two of the three indicators are met	Suitable
One of the three indicators is met	Less suitable
None of the indicators are met	Not compliant

(Sumintono & Widhiarso, 2015)

In addition, the level of difficulty of *the* items *(item measure)* and *item maps* are used to map the level of difficulty of the items to the abilities of the students. These item maps can be divided into five interpretation zones to identify the level of difficulty in more detail, ranging from *very hard, hard, medium, easy,* and *very easy.*

- ***Very Hard***
  These items are located at the top of the map and are only answered by students with the highest abilities.
- ***Hard***
  These items are located above the average scale and can only be answered by students with above-average abilities.
- ***Medium***
  This item is located around the midpoint of the logit scale, effective for distinguishing students with average abilities.
- ***Easy***
  This item is located below the average of the logit scale. Students with abilities below average to average can generally answer these items correctly.
- ***Very Easy***
  This item is located at the bottom of the items. They have a very low (large negative) logit value. These questions can be answered correctly by almost all students.

To assess the validity of the test instrument, the information function (TIF) and *Standard Error of Measurement* (SEM) are used. The information function measures how well the instrument measures specific abilities (Sumaryanta, 2021) , while SEM addresses unavoidable errors in measurement. The relationship between the two is inversely proportional; an increase in information correlates with a decrease in SEM, indicating an increase in precision (Retnawati, 2020; Setiawati et al., 2013) . To see the suitability of the test with the students' abilities based on the information function and SEM, it can be classified as follows:

Table 7. Classification of Ability Estimation

Ability Range (θ)	Category
-4 to -2.5	Very Low
-2.5 to -1	Low
-1 to 1	Moderate
1 to 2.5	High
2.5 to 4	Very High

RESULTS AND DISCUSSION

Results

The characteristics of the *misconception check* instrument were analyzed through content validity, readability testing, and data analysis using the *Rasch* model with the help of Winsteps.

Content validity testing was conducted to evaluate the extent to which the items could represent thermodynamics material.

Table 8. Content Validity Test Results

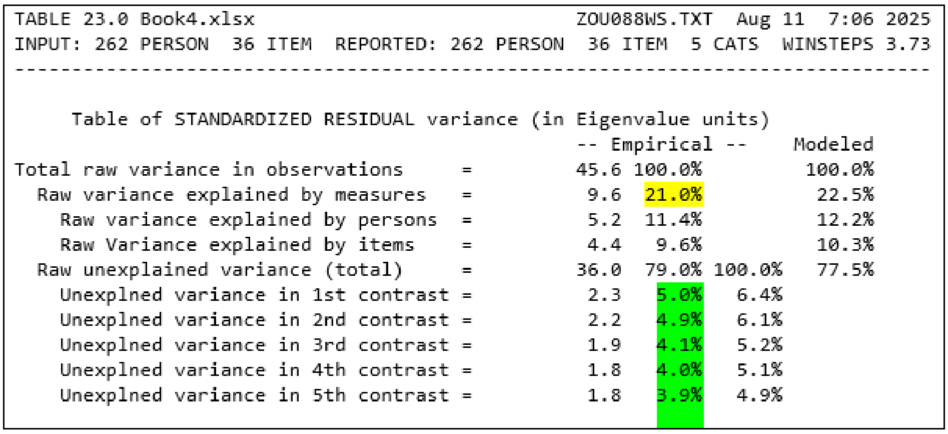
Question	$\Sigma(S)$	N (C-1)	V	Note
1A	20	20	1.00	Very High
1B	20	20	1.00	Very High
2A	20	20	1.00	Very High
2B	20	20	1.00	Very High
3A	20	20	1.00	Very High
3B	20	20	1.00	Very High
4A	20	20	1.00	Very High
4B	20	20	1.00	Very High
5A	20	20	1.00	Very High
5B	20	20	1.00	Very High
6AB*	20	20	1.00	Very High
7A	20	20	1.00	Very High
7B	20	20	1.00	Very High
8A	20	20	1.00	Very High
8B	20	20	1.00	Very High
9A	20	20	1.00	Very High
9B	20	20	1.00	Very High
10A	20	20	1.00	Very High
10B	20	20	1.00	Very High
11AB*	20	20	1.00	Very High
12A	20	20	1.00	Very High
12B	20	20	1.00	Very High
13A	20	20	1.00	Very High
13B	20	20	1	Very High
14A	20	20	1	Very High
14B	20	20	1	Very High
15A	20	20	1	Very High
15B	20	20	1	Very High
16AB*	20	20	1	Very High
17AB*	20	20	1	Very High
18A	20	20	1.0	Very High
18B	20	20	1.00	Very High
19A	20	20	1.00	Very High
19B	20	20	1.00	Very High
20A	20	20	1.00	Very High
20B	20	20	1.00	Very High
Overall average	**720**	**720**	1.00	**Very High**

Readability tests were conducted to ensure that the language and format of the instruments were easily understood by students.

Table 8. Readability Test Results

Question	$\Sigma(S)$	N (C-1)	V	Note
1A	41	44	0.93	Very High
1B	44	44	1.00	Very High
2A	43	44	0.97	Very High
2B	44	44	1.00	Very High
3A	44	44	1.00	Very High
3B	42	44	0.95	Very High
4A	42	44	0.95	Very High
4B	42	44	0.95	Very High
5A	40	44	0.90	Very High
5B	41	44	0.93	Very High
6AB*	44	44	1.00	Very High
7A	42	44	0.95	Very High
7B	44	44	1.00	Very High
8A	41	44	0.93	Very High
8B	42	44	0.95	Very High
9A	42	44	0.95	Very High
9B	42	44	0.95	Very High
10A	44	44	1.00	Very High
10B	44	44	1.00	Very High
11AB*	42	44	0.95	Very High
12A	43	44	0.97	Very High
12B	42	44	0.95	Very High
13A	40	44	0.90	Very High
13B	42	44	0.95	Very High
14A	44	44	1.00	Very High
14B	44	44	1.00	Very High
15A	38	44	0.86	Very High
15B	41	44	0.93	Very High
16AB*	44	44	1.00	Very High
17AB*	44	44	1.00	Very High
18A	41	44	0.93	Very High
18B	41	44	0.93	Very High
19A	44	44	1.00	Very High
19B	42	44	0.95	Very High
20A	44	44	1.00	Very High
20B	44	44	1.00	Very High
Overall average	**1528**	**1584**	**0.96465**	**Very High**

Unidimensionality is a crucial characteristic that assumes that the instrument measures only a single construct. The results are as follows:



```
TABLE 23.0 Book4.xlsx                              ZOU088WS.TXT  Aug 11  7:06 2025
INPUT: 262 PERSON  36 ITEM  REPORTED: 262 PERSON  36 ITEM  5 CATS  WINSTEPS 3.73
--------------------------------------------------------------------------------

     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                               -- Empirical --    Modeled
Total raw variance in observations     =        45.6 100.0%        100.0%
  Raw variance explained by measures   =         9.6  21.0%         22.5%
    Raw variance explained by persons  =         5.2  11.4%         12.2%
    Raw Variance explained by items    =         4.4   9.6%         10.3%
  Raw unexplained variance (total)     =        36.0  79.0% 100.0%   77.5%
    Unexplned variance in 1st contrast =         2.3   5.0%   6.4%
    Unexplned variance in 2nd contrast =         2.2   4.9%   6.1%
    Unexplned variance in 3rd contrast =         1.9   4.1%   5.2%
    Unexplned variance in 4th contrast =         1.8   4.0%   5.1%
    Unexplned variance in 5th contrast =         1.8   3.9%   4.9%
```

Figure 2. *Unidimensionality* Test Results

After passing the *unidimensionality* test, an *item-person* map test was conducted, which is a key feature of *Rasch* analysis that provides a visual representation of the characteristics of the instrument. This allows us to see the distribution of question difficulty levels and student abilities simultaneously. The results are as follows:

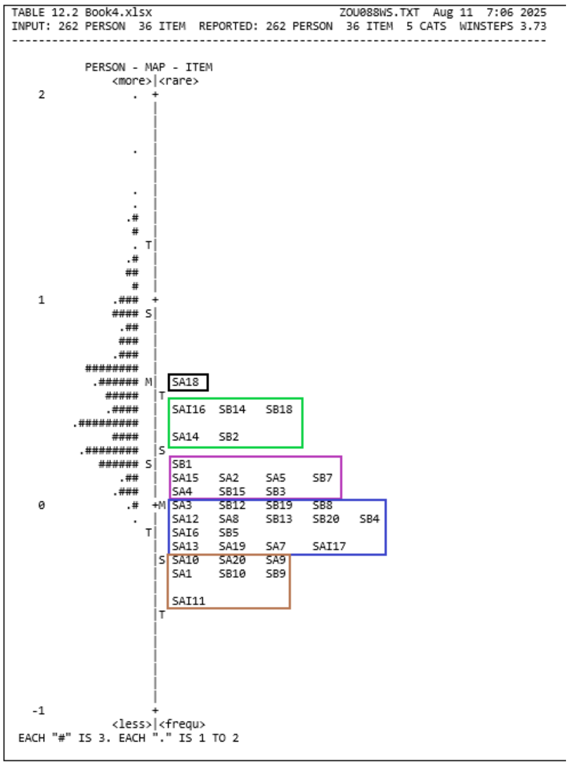
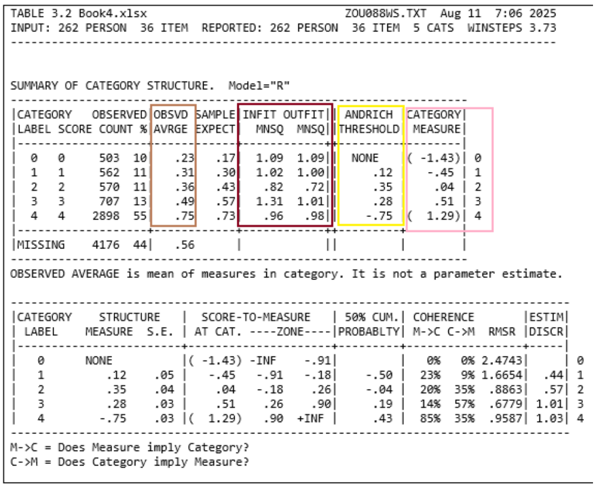


Figure 3. *Person Item Map* of 36 Questions

Next, a *category function* analysis was conducted to test whether each answer option on the multiple-choice instrument functioned effectively and had a logical sequence.



```
TABLE 3.2 Book4.xlsx                              ZOU088WS.TXT  Aug 11  7:06 2025
INPUT: 262 PERSON  36 ITEM  REPORTED: 262 PERSON  36 ITEM  5 CATS  WINSTEPS 3.73
--------------------------------------------------------------------------------


SUMMARY OF CATEGORY STRUCTURE.  Model="R"
-------------------------------------------------------------------
|CATEGORY   OBSERVED|OBSVD SAMPLE|INFIT OUTFIT|| ANDRICH |CATEGORY|
|LABEL SCORE COUNT %|AVRGE EXPECT|  MNSQ  MNSQ||THRESHOLD| MEASURE|
|-------------------+------------+------------++---------+--------|
|  0   0     503  10|   .23    .17|  1.09  1.09||  NONE   |( -1.43)| 0
|  1   1     562  11|   .31    .30|  1.02  1.00||    .12  |   -.45 | 1
|  2   2     570  11|   .36    .43|   .82   .72||    .35  |    .04 | 2
|  3   3     707  13|   .49    .57|  1.31  1.01||    .28  |    .51 | 3
|  4   4    2898  55|   .75    .73|   .96   .98||   -.75  |(  1.29)| 4
|-------------------+------------+------------++---------+--------|
|MISSING    4176  44|   .56      |            ||         |        |
-------------------------------------------------------------------
OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

--------------------------------------------------------------------------------
|CATEGORY    STRUCTURE    |  SCORE-TO-MEASURE   | 50% CUM.| COHERENCE       |ESTIM|
| LABEL    MEASURE   S.E. | AT CAT. ----ZONE----|PROBABLTY| M->C C->M  RMSR |DISCR|
|-------------------------+---------------------+---------+-----------------+-----|
|   0      NONE           |( -1.43) -INF   -.91|         |   0%   0% 2.4743|     | 0
|   1         .12    .05 |   -.45   -.91   -.18|    -.50 |  23%   9% 1.6654|  .44| 1
|   2         .35    .04 |    .04   -.18    .26|    -.04 |  20%  35%  .8863|  .57| 2
|   3         .28    .03 |    .51    .26    .90|     .19 |  14%  57%  .6779| 1.01| 3
|   4        -.75    .03 |(  1.29)   .90  +INF |     .43 |  85%  35%  .9587| 1.03| 4
--------------------------------------------------------------------------------
M->C = Does Measure imply Category?
C->M = Does Category imply Measure?
```

Figure 4. *Category Function* Test Results

To test the effectiveness of the answer options, an analysis of the probability curve of students choosing each option was conducted.

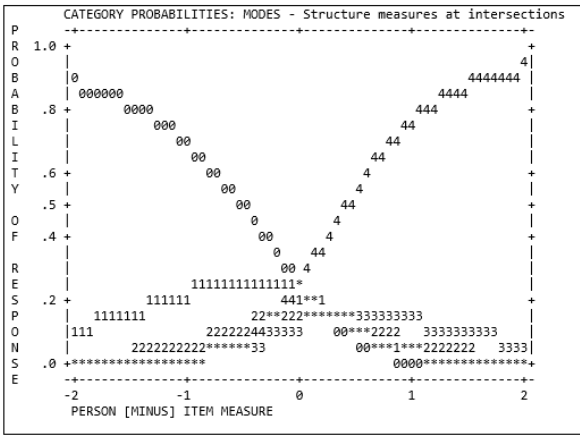


Figure 5. *Category Probability* Curve

The formative assessment data was analyzed using *Item Polarity*. This analysis aimed to examine the consistency of each item with the overall measurement scale.

TABLE 26.1 Book4.xlsx ZOU088WS.TXT Aug 11 7:06 2025
INPUT: 262 PERSON 36 ITEM REPORTED: 262 PERSON 36 ITEM 5 CATS WINSTEPS 3.73

PERSON: REAL SEP.: 1.31 REL.: .63 ... ITEM: REAL SEP.: 3.34 REL.: .92

ITEM STATISTICS: CORRELATION ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
18	236	129	.58	.06	1.71	6.2	1.81	6.2	.07	.44	17.2	20.3	SA18
14	390	129	-.06	.07	1.07	.6	1.56	3.0	.15	.32	25.0	29.6	SA12
24	410	133	-.04	.07	.80	-1.8	.92	-.4	.16	.34	34.6	29.2	SB4
32	296	133	.44	.06	.84	-2.0	.90	-1.0	.21	.43	27.8	20.8	SB14
27	406	133	-.02	.07	.85	-1.3	.83	-1.1	.24	.34	34.6	27.8	SB8
6	343	129	.15	.06	1.01	.1	1.08	.7	.25	.37	26.6	21.7	SA2
25	424	133	-.11	.07	.88	-.9	.93	-.3	.26	.32	28.6	31.7	SB5
2	929	262	-.48	.07	1.09	.7	1.08	.5	.27	.23	58.2	56.0	SAI11
16	294	129	.35	.06	1.31	3.3	1.34	2.9	.27	.41	9.4	18.5	SA14
5	438	129	-.35	.09	1.20	1.2	1.10	.5	.28	.25	45.3	43.8	SA1
26	377	133	.12	.07	1.05	.6	1.24	1.7	.29	.37	18.8	22.2	SB7
36	419	133	-.08	.07	1.27	2.0	1.24	1.3	.29	.33	27.8	30.7	SB20
29	458	133	-.33	.09	.99	.0	.90	-.4	.30	.27	51.1	48.7	SB10
30	398	133	.02	.07	.86	-1.3	.85	-1.0	.31	.35	31.6	27.6	SB12
4	851	262	-.19	.05	1.22	2.1	1.16	1.2	.32	.30	36.4	34.6	SAI17
15	413	129	-.19	.08	1.00	.1	.89	-.5	.34	.29	30.5	33.0	SA13
10	413	129	-.19	.08	.96	-.3	.96	-.2	.35	.29	35.2	33.0	SA7
11	395	129	-.09	.07	.96	-.3	.89	-.6	.35	.31	32.8	29.7	SA8
17	354	129	.10	.07	1.05	.5	1.05	.4	.35	.36	24.2	21.7	SA15
19	411	129	-.17	.08	1.02	.2	.90	-.5	.36	.29	41.4	32.9	SA19
21	355	133	.21	.06	1.29	2.9	1.20	1.5	.36	.39	17.3	20.8	SB1
20	422	129	-.24	.08	.92	-.5	.79	-1.1	.36	.28	39.8	35.2	SA20
13	430	129	-.29	.08	1.09	.6	.84	-.7	.37	.27	43.8	40.1	SA10
3	544	262	.48	.04	.94	-.9	.92	-1.1	.38	.43	21.1	20.3	SAI16
28	457	133	-.32	.09	1.07	.5	.98	.0	.40	.27	48.9	48.6	SB9
7	374	129	.01	.07	.92	-.7	.95	-.3	.41	.34	28.9	26.9	SA3
12	422	129	-.24	.08	.77	-1.7	.73	-1.5	.41	.28	43.0	35.2	SA9
34	288	133	.47	.06	1.05	.6	1.08	.8	.41	.43	12.8	21.0	SB18
8	368	129	.04	.07	.84	-1.6	.78	-1.7	.41	.34	23.4	23.2	SA4
9	347	129	.13	.07	1.07	.8	1.02	.2	.42	.37	15.6	21.8	SA5
1	831	262	-.13	.05	1.10	1.1	.96	-.2	.42	.31	32.2	31.8	SAI6
23	388	133	.07	.07	.98	-.2	.89	-.7	.47	.36	28.6	25.7	SB3
35	409	133	-.03	.07	.96	-.3	.80	-1.2	.49	.34	33.1	29.2	SB19
22	325	133	.33	.06	.63	-4.9	.61	-4.1	.49	.41	33.8	20.8	SB2
33	384	133	.09	.07	.83	-1.6	.77	-1.6	.49	.37	21.1	22.1	SB15
31	416	133	-.07	.07	.95	-.4	.77	-1.4	.51	.33	31.6	30.6	SB13
MEAN	428.2	145.6	.00	.07	1.02	.1	.99	.0			30.9	29.6	
S.D.	144.6	41.2	.25	.01	.19	1.8	.23	1.7			10.8	8.9	

Figure 6. *Correlation Order* Results

Instrument reliability was conducted to analyze the performance of the instrument and students in greater depth. Next, *summary statistics* will present important data related to reliability, fit, and separation, which are crucial for understanding the overall quality of the instrument.

TABLE 3.1 Book4.xlsx ZOU088WS.TXT Aug 11 7:06 2025
INPUT: 262 PERSON 36 ITEM REPORTED: 262 PERSON 36 ITEM 5 CATS WINSTEPS 3.73

SUMMARY OF 261 MEASURED (NON-EXTREME) PERSON

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	58.8	20.0	.57	.20	1.02	.1	.99	.1
S.D.	9.7	.0	.35	.05	.27	.9	.31	.9
MAX.	77.0	20.0	1.71	.47	2.03	2.4	1.93	2.6
MIN.	35.0	20.0	-.06	.15	.37	-3.1	.35	-3.0

REAL RMSE .22 TRUE SD .28 SEPARATION 1.31 PERSON RELIABILITY .63
MODEL RMSE .20 TRUE SD .29 SEPARATION 1.44 PERSON RELIABILITY .67
S.E. OF PERSON MEAN = .02

MAXIMUM EXTREME SCORE: 1 PERSON
VALID RESPONSES: 55.6% (APPROXIMATE)

SUMMARY OF 262 MEASURED (EXTREME AND NON-EXTREME) PERSON

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	58.8	20.0	.58	.20				
S.D.	9.8	.0	.40	.11				
MAX.	80.0	20.0	3.52	1.76				
MIN.	35.0	20.0	-.06	.15	.37	-3.1	.35	-3.0

REAL RMSE .24 TRUE SD .32 SEPARATION 1.31 PERSON RELIABILITY .63
MODEL RMSE .23 TRUE SD .32 SEPARATION 1.41 PERSON RELIABILITY .67
S.E. OF PERSON MEAN = .02

PERSON RAW SCORE-TO-MEASURE CORRELATION = .91 (approximate due to missing data)
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .34 (approximate due to missing data)

SUMMARY OF 36 MEASURED (NON-EXTREME) ITEM

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	428.2	145.6	.00	.07	1.02	.1	.99	.0
S.D.	144.6	41.2	.25	.01	.19	1.8	.23	1.7
MAX.	929.0	262.0	.58	.09	1.71	6.2	1.81	6.2
MIN.	236.0	129.0	-.48	.04	.63	-4.9	.61	-4.1

REAL RMSE .07 TRUE SD .24 SEPARATION 3.34 ITEM RELIABILITY .92
MODEL RMSE .07 TRUE SD .24 SEPARATION 3.47 ITEM RELIABILITY .92
S.E. OF ITEM MEAN = .04

UMEAN=.0000 USCALE=1.0000
ITEM RAW SCORE-TO-MEASURE CORRELATION = -.53 (approximate due to missing data)
5220 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 12387.76 with 4921 d.f. p=.0000
Global Root-Mean-Square Residual (excluding extreme scores): 1.2423

Figure 7. *Summary Statistics* Results

To evaluate the quality of each item individually, an *item fit* analysis was conducted. This analysis focused on three main indicators: *Outfit Mean Square* (MNSQ), *Outfit Z-Standard* (ZSTD), and *Point Measure Correlation (Pt Mean Corr).* An item can be considered valid if it meets two of the three categories.

TABLE 10.1 Book4.xlsx ZOU088WS.TXT Aug 11 7:06 2025
INPUT: 262 PERSON 36 ITEM REPORTED: 262 PERSON 36 ITEM 5 CATS WINSTEPS 3.73

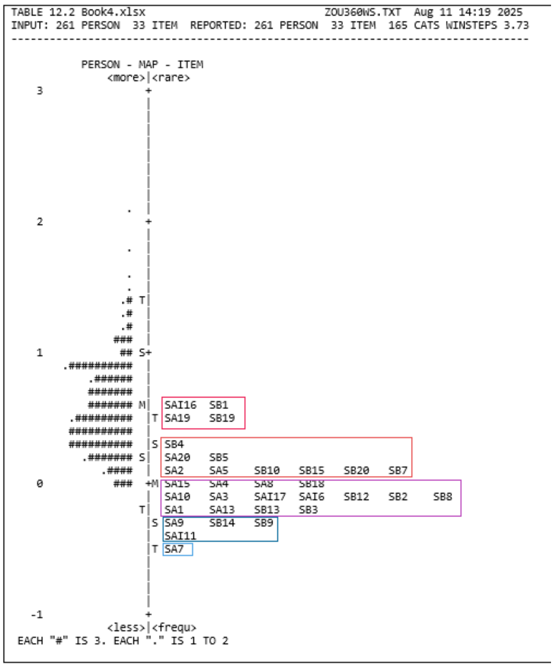
PERSON: REAL SEP.: 1.31 REL.: .63 ... ITEM: REAL SEP.: 3.34 REL.: .92

ITEM STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
18	236	129	.58	.06	1.71	6.2	1.81	6.2	A .07	.44	17.2	20.3	SA18
14	390	129	-.06	.07	1.07	.6	1.56	3.0	B .15	.32	25.0	29.6	SA12
16	294	129	.35	.06	1.31	3.3	1.34	2.9	C .27	.41	9.4	18.5	SA14
21	355	133	.21	.06	1.29	2.9	1.20	1.5	D .36	.39	17.3	20.8	SB1
36	419	133	-.08	.07	1.27	2.0	1.24	1.3	E .29	.33	27.8	30.7	SB20
26	377	133	.12	.07	1.05	.6	1.24	1.7	F .29	.37	18.8	22.2	SB7
4	851	262	-.19	.05	1.22	2.1	1.16	1.2	G .32	.30	36.4	34.6	SAI17
5	438	129	-.35	.09	1.20	1.2	1.10	.5	H .28	.25	45.3	43.8	SA1
1	831	262	-.13	.05	1.10	1.1	.96	-.2	I .42	.31	32.2	31.8	SAI6
2	929	262	-.48	.07	1.09	.7	1.08	.5	J .27	.23	58.2	56.0	SAI11
13	430	129	-.29	.08	1.09	.6	.84	-.7	K .37	.27	43.8	40.1	SA10
6	343	129	.15	.06	1.01	.1	1.08	.7	L .25	.37	26.6	21.7	SA2
34	288	133	.47	.06	1.05	.6	1.08	.8	M .41	.43	12.8	21.0	SB18
28	457	133	-.32	.09	1.07	.5	.98	.0	N .40	.27	48.9	48.6	SB9
9	347	129	.13	.07	1.07	.8	1.02	.2	O .42	.37	15.6	21.8	SA5
17	354	129	.10	.07	1.05	.5	1.05	.4	P .35	.36	24.2	21.7	SA15
19	411	129	-.17	.08	1.02	.2	.90	-.5	Q .36	.29	41.4	32.9	SA19
15	413	129	-.19	.08	1.00	.1	.89	-.5	R .34	.29	30.5	33.0	SA13
29	458	133	-.33	.09	.99	.0	.90	-.4	r .30	.27	51.1	48.7	SB10
23	388	133	.07	.07	.98	-.2	.89	-.7	q .47	.36	28.6	25.7	SB3
11	395	129	-.09	.07	.96	-.3	.89	-.6	p .35	.31	32.8	29.7	SA8
10	413	129	-.19	.08	.96	-.3	.96	-.2	o .35	.29	35.2	33.0	SA7
35	409	133	-.03	.07	.96	-.3	.80	-1.2	n .49	.34	33.1	29.2	SB19
7	374	129	.01	.07	.92	-.7	.95	-.3	m .41	.34	28.9	26.9	SA3
31	416	133	-.07	.07	.95	-.4	.77	-1.4	l .51	.33	31.6	30.6	SB13
3	544	262	.48	.04	.94	-.9	.92	-1.1	k .38	.43	21.1	20.3	SAI16
25	424	133	-.11	.07	.88	-.9	.93	-.3	j .26	.32	28.6	31.7	SB5
20	422	129	-.24	.08	.92	-.5	.79	-1.1	i .36	.28	39.8	35.2	SA20
24	410	133	-.04	.07	.80	-1.8	.92	-.4	h .16	.34	34.6	29.2	SB4
32	296	133	.44	.06	.84	-2.0	.90	-1.0	g .21	.43	27.8	20.8	SB14
30	398	133	.02	.07	.86	-1.3	.85	-1.0	f .31	.35	31.6	27.6	SB12
27	406	133	-.02	.07	.85	-1.3	.83	-1.1	e .24	.34	34.6	27.8	SB8
8	368	129	.04	.07	.84	-1.6	.78	-1.7	d .41	.34	23.4	23.2	SA4
33	384	133	.09	.07	.83	-1.6	.77	-1.6	c .49	.37	21.1	22.1	SB15
12	422	129	-.24	.08	.77	-1.7	.73	-1.5	b .41	.28	43.0	35.2	SA9
22	325	133	.33	.06	.63	-4.9	.61	-4.1	a .49	.41	33.8	20.8	SB2
MEAN	428.2	145.6	.00	.07	1.02	.1	.99	.0			30.9	29.6	
S.D.	144.6	41.2	.25	.01	.19	1.8	.23	1.7			10.8	8.9	

Figure 8. Analysis of the Quality of Each Item

Invalid items were excluded from the analysis, leaving 33 items to be analyzed to determine their level of difficulty *(item measure)*. This analysis aimed to group the items to provide a clearer picture of the test's characteristics. This grouping was interpreted through item distribution maps (*person item maps).*



```
TABLE 12.2 Book4.xlsx                                 ZOU360WS.TXT  Aug 11 14:19 2025
INPUT: 261 PERSON  33 ITEM  REPORTED: 261 PERSON  33 ITEM  165 CATS WINSTEPS 3.73
--------------------------------------------------------------------------------

          PERSON - MAP - ITEM
              <more>|<rare>
   3                +
                    |
                    |
                    |
                    |
                    |
                    |
                    |
                    |
                  . |
   2                +
                    |
                  . |
                    |
                  . |
                  . |
                 .# T|
                 .# |
                 .# |
                ### |
   1             ## S+
        .########## |
            .###### |
            ####### |
            ####### M|  SAI16  SB1
         .#########  |T SA19   SB19
         ########## |
         ##########  |S SB4
          .####### S|  SA20   SB5
              .#### |  SA2    SA5    SB10   SB15   SB20   SB7
   0            ###  +M SA15   SA4    SA8    SB18
                    |  SA10   SA3    SAI17  SAI6   SB12   SB2    SB8
                   T|  SA1    SA13   SB13   SB3
                    |S SA9    SB14   SB9
                    |  SAI11
                    |T SA7
                    |
                    |
                    |
                    |
  -1                +
              <less>|<frequ>
 EACH "#" IS 3. EACH "." IS 1 TO 2
```

Figure 9. *Person Item Map* of 33 Items

Based on the results of the item map analysis, the items were identified as having varying levels of difficulty. To reinforce the findings in the person item maps, the data will then be analyzed by matching it with *the measure order.*

TABLE 13.1 Book4.xlsx ZOU360WS.TXT Aug 11 14:19 2025
INPUT: 261 PERSON 33 ITEM REPORTED: 261 PERSON 33 ITEM 165 CATS WINSTEPS 3.73

PERSON: REAL SEP.: 1.26 REL.: .61 ... ITEM: REAL SEP.: 3.48 REL.: .92

ITEM STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM	G
18	232	128	.60	.06	1.57	5.3	1.72	5.5	.00	.47	17.2	18.1	SB1	0
3	540	261	.56	.05	1.04	.6	1.01	.2	.41	.43	18.8	24.3	SAI16	0
32	296	133	.50	.07	1.16	1.6	1.15	1.5	.23	.40	29.3	30.5	SB19	0
16	290	128	.48	.06	1.19	2.1	1.26	2.1	.27	.43	8.6	17.5	SA19	0
21	355	133	.30	.06	1.10	1.1	1.02	.2	.39	.43	16.5	19.2	SB4	0
22	325	133	.19	.07	.87	-1.4	.84	-1.5	.52	.40	36.8	28.8	SB5	0
17	350	128	.17	.06	1.01	.2	1.02	.2	.39	.38	24.2	22.0	SA20	0
6	339	128	.15	.07	1.12	1.2	1.16	1.4	.25	.37	27.3	24.6	SA2	0
33	384	133	.09	.07	.86	-1.3	.78	-1.5	.50	.38	24.1	23.4	SB20	0
26	377	133	.08	.07	1.13	1.3	1.29	1.9	.27	.40	19.5	21.2	SB10	0
23	388	133	.07	.07	.94	-.6	.85	-.9	.46	.39	27.1	24.1	SB7	0
30	398	133	.07	.08	.99	.0	.98	-.1	.34	.35	33.8	34.2	SB15	0
9	343	128	.06	.06	.95	-.6	.89	-.8	.47	.41	17.2	14.6	SA5	0
31	416	133	.04	.07	.82	-1.5	.64	-1.9	.51	.36	41.4	37.9	SB18	0
8	364	128	-.01	.07	.92	-.8	.84	-1.3	.46	.36	21.1	22.1	SA4	0
11	391	128	-.03	.07	.96	-.3	.89	-.6	.39	.33	32.8	32.4	SA8	0
15	409	128	-.05	.07	.91	-.6	.83	-.8	.39	.31	30.5	33.5	SA15	0
1	827	261	-.05	.05	.88	-1.4	.75	-1.7	.46	.35	38.7	34.5	SAI6	0
4	847	261	-.07	.05	1.06	.6	1.06	.4	.31	.33	41.8	38.2	SAI17	0
7	370	128	-.07	.07	.92	-.9	.92	-.5	.45	.37	28.9	23.7	SA3	0
19	407	128	-.10	.07	.95	-.4	.84	-.8	.41	.32	41.4	32.7	SB2	0
24	410	133	-.10	.09	1.11	.8	1.14	.9	.17	.32	36.8	38.0	SB8	0
27	406	133	-.10	.08	1.09	.7	1.03	.2	.26	.33	39.8	38.8	SB12	0
13	426	128	-.14	.08	.90	-.6	.69	-1.4	.42	.30	46.9	40.7	SA10	0
20	418	128	-.17	.08	.92	-.5	.80	-1.0	.41	.30	39.8	35.5	SB3	0
5	434	128	-.18	.08	.99	.0	.93	-.2	.31	.28	45.3	43.5	SA1	0
14	386	128	-.19	.07	1.14	1.3	1.52	3.1	.13	.34	22.7	28.9	SA13	0
28	457	133	-.24	.08	.93	-.4	.95	-.1	.36	.30	53.4	49.5	SB13	0
12	418	128	-.30	.08	.87	-.9	.80	-1.1	.47	.29	39.8	33.3	SA9	0
25	424	133	-.31	.08	1.07	.6	1.09	.6	.24	.33	24.8	30.7	SB9	0
29	458	133	-.32	.09	1.03	.2	.92	-.3	.30	.28	52.6	49.7	SB14	0
2	925	261	-.44	.07	.95	-.3	.96	-.2	.31	.26	57.5	54.8	SAI11	0
10	409	128	-.48	.08	.94	-.5	.92	-.4	.39	.32	33.6	30.2	SA7	0
MEAN	430.9	146.1	.00	.07	1.01	.1	.98	.0			32.4	31.2		
S.D.	148.5	42.7	.26	.01	.14	1.3	.22	1.5			11.7	9.6		

Figure 10. Measure Order *Test Results*

The process of creating information function curves and SEM began by exporting TIF data from Winsteps to Excel. Then, the SEM value was calculated using the formula $SEM = 1\sqrt{I}$, where I is the information value. This data was then visualized in a scatter plot with the X-axis as the measure (ability) and the Y-axis as the TIF and SEM values.

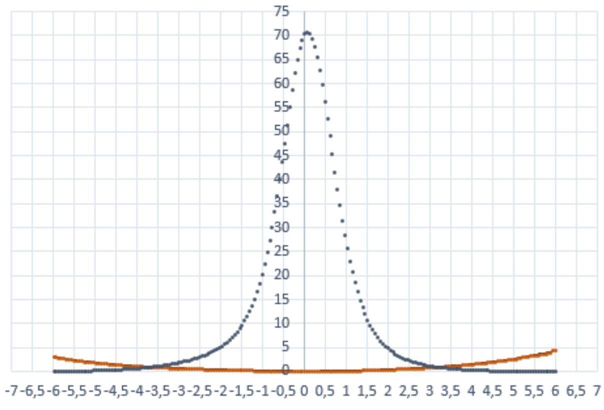


Figure 11. TIF and SEM Curves

Discussion

Based on the results of the above analysis, this formative assessment instrument in the form of *a misconception check* shows strong and reliable characteristics. The content validation of the instrument has been tested with a high Aiken's V coefficient (above 0.88) from experts, consisting of three physics lecturers and two physics teachers. The results can be seen in Table 6, showing that each item has been carefully evaluated and considered relevant and representative of the thermodynamics material taught at the high school level (Aiken, 1985) .

In addition, the results of the readability test, which can be seen in Table 7, show excellent results. Testing of 11 students produced an average Aiken's V coefficient of 0.96, which is well above the minimum value. This value proves that this instrument is easy to understand in terms of language and format by students, so it can be used without linguistic barriers.

The use of the Rasch model through Winsteps software, an approach that is highly relevant for analyzing dichotomous or polytomous data (Boone & Noltemeyer, 2017) , further reinforces the quality of the instrument. Unidimensionality analysis with

PCA shows that the instrument consistently measures a single construct, namely students' conceptual understanding of thermodynamics. This is evidenced by *a raw variance* of 21%, which meets the minimum requirement of 20%, and an *unexplained variance* value in the range of 3% to 5%, indicating excellent criteria.

Although disordered thresholds were found in the *category function* analysis ( ), where each response category did not fully function in a logical order due to students with higher comprehension abilities sometimes tending to choose categories that should have been chosen by students with lower abilities, or vice versa, this could also have occurred because the instrument was not tested on a larger sample size, resulting in a lack of varied responses.

This indicates that the logical order of answer options does not fully function, but this instrument is still reliable (Engelhard & Wind, 2017) . The Infit and Outfit MNSQ values for each category are within an acceptable range (0.5 to 1.5), indicating that the data as a whole remains consistent with the Rasch model and that the students' response patterns do not deviate significantly (Bond & Fox, 2013) . This analysis is very important because it provides unique insights into the *measure* values of each incorrect answer category. Category 4 (correct answers) has *a measure* of 1.29 (the highest level of difficulty), while the other categories represent different types of misconceptions, ranging from *almost scientific concepts* (category 3 with *a measure* of 0.51) to *non-understanding of a concept* (category 0 with *a measure* of -1.43). This underscores that incorrect answers are as important as correct answers in diagnosing misconceptions and designing appropriate learning interventions (Backhaus, 2024; Derya Kaltakci, 2012; Jannah & Rahmi, 2020; Kiray & Simsek, 2021) .

The *Item-Person* Map visualizes the alignment between the difficulty level of test items and the abilities of 262 students. The distribution of student abilities is concentrated in the range of 0 to +1.5 logits, which is in line with the distribution of the items. This map allows educators to identify the "concept zones" of students and determine which concepts are the most difficult (Cross & Angelo, 1993; Leonard, 2024) . In addition, the category probability curve shows an ideal pattern: the probability of choosing the correct answer (category 4) increases as the learner's ability increases, while the probability of choosing *a distractor* (categories 0, 1, 2, 3) decreases. This pattern proves that each item functions well in distinguishing learners based on their ability levels, although the *disordered thresholds* indicate the need for revision of some items in the future.

Technically, the reliability of the instrument is very good with an *item* reliability value of 0.92(Sumintono & Widhiarso, 2015) , indicating strong internal consistency. *Item fit* analysis shows that 33 of the 36 items are valid because they meet at least two of the three criteria set (*Outfit MNSQ, Outfit ZSTD,* and *Pt Mean Corr)*. The distribution of item difficulty levels, ranging from very easy to very difficult (divided into five categories), shows that this instrument is capable of measuring a wide spectrum of student abilities. Thus, although some improvements may be necessary, such as revising invalid items and testing on a larger sample, this instrument is, overall, a valid and reliable tool for identifying students' conceptions and misconceptions.

Based on a comprehensive analysis, the developed *Misconception Check* formative assessment instrument has strong characteristics for analyzing students' conceptions of thermodynamics.

First, these characteristics are supported by strong content validity, a characteristic that has been confirmed through expert assessment using Aiken's V index. The results of the analysis show that the 36 comprehensively developed items represent the scope of the material, construction, and language, so that they can be used as an accurate and relevant assessment tool (Aiken, 1985) . In addition, readability tests also reinforce the feasibility of this instrument. All items have an Aiken's V coefficient value above the minimum value set, with an overall average of 0.96, which is classified as "Very High" (Sumintono & Widhiarso, 2015) . This high readability ensures that students' responses purely reflect their understanding, rather than being influenced by difficulties in interpreting ambiguous questions.

Second, the characteristics of this instrument are reinforced by Rasch model analysis. The unidimensionality test proves that the items consistently measure a single construct, namely thermodynamic concepts, so that each finding can be interpreted specifically. The *item-person* distribution map also shows the distribution of items in accordance with the distribution of student abilities, ensuring that this instrument is capable of identifying concepts at various levels of understanding. Furthermore, *the category function* proved to work well. The analysis shows that each response category (scores 0 to 4) has a high probability in sequential ability ranges, confirming that the designed polytomous scale functions as intended.

Third, evidence of characteristics also comes from *the correlation order of* the items and reliability. The analysis results show that 35 of the 36 items support each other in measuring the same construct uniformly. In addition, the very high *item* reliability value of 0.92 (Sumintono & Widhiarso, 2015) , is a strong argument for the instrument's feasibility, as it shows that the items are very consistent and reliable. Finally, *item fit* analysis confirmed this feasibility, with 33 of the 36 items having good results and items that did not meet the criteria being eliminated. Thus, this instrument is a robust, consistent, and reliable tool that is suitable as a basis for pedagogical decision-making.

The validity of the *Misconception Check* formative assessment instrument was evaluated through TIF curve and SEM analysis. The results of this analysis show that the instrument has a good level of validity and reliability for use.

It can be seen from the *Test Information Function* (TIF) curve that it has a peak or highest information point on the map of around 70.0 logit. This shows that the test information of the instrument is greatest when used and tested on students who have abilities of around 70.0 logit. The second cut-off point of the curve is at teta -3.5 and +3.2, indicating that *the Misconception Check* formative assessment test instrument on thermodynamics material is reliable for determining the level of students' conceptions from a range of -3.5 with very low abilities to +3.2 with very high abilities.

CONCLUSION

Based on the research results, the developed formative assessment instrument, Misconception Check, is proven to be valid, reliable, and suitable for diagnosing high school students' conceptions of thermodynamics. This is supported by strong content validity and readability tests that show that the questions are easy to understand. Analysis using the Rasch model further reinforced these characteristics, such as unidimensionality, which proved that the instrument measured only a single construct. Although there were slight discrepancies in some items, overall the data produced was very reliable. Of the 36 items, 33 were proven to be of high quality and suitable, with excellent reliability of 0.92 and varying levels of difficulty, making it an effective and reliable tool. In practical terms, this instrument can be used by teachers to provide targeted feedback and design appropriate learning, which can ultimately improve students' conceptual understanding. Theoretically, this research contributes to the literature on formative assessment in physics education, particularly in the use of the Rasch model to ensure instrument quality. These results reinforce the framework for developing assessment instruments that can measure and identify conceptions. For optimization, it is recommended to conduct a broader sample test to improve generalization and develop usage guidelines for educators to interpret the results effectively.

ACKNOWLEDGMENT
We would like to thank SMAN 24 Bandung, SMAN 11 Bandung, and SMAN 16 Bandung for giving the researchers the opportunity to conduct this study.

REFERENCES

Aditomo, A. (2024). *Panduan Pembelajaran dan Asesmen Pendidikan Anak Usia Dini, Pendidikan Dasar, dan Pendidikan Menengah Edisi Revisi Tahun 2024*.

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings, educational and psychological measurument. *Educational and Psychological Measurement*, *45*(1), 131–142.

Aufschnaiter, C. von, & Alonzo, A. C. (2018). Foundations of formative assessment: Introducing a learning progression to guide preservice physics teachers' video-based interpretation of student thinking. *Applied Measurement in Education*, *31*(2), 113–127. https://doi.org/10.1080/08957347.2017.1408629

Backhaus, A. (2024). *Diagnosing Misconceptions*. Carpentries. https://carpentries.github.io/lesson-development-training/misconceptions-mcqs.html

Bhaw, N., Kriek, J., & Rampho, G. (2024). The Use of Multiple-choice Questions as an Assessment Tool in First-year University Physics Modules. *Journal of Education and Practice*, *July*, 0–3. https://doi.org/10.7176/jep/16-1-07

Bond, T. G., & Fox, C. M. (2013). Applying the Rasch Model. In *Applying the Rasch Model*. Routledge. https://doi.org/10.4324/9781410614575

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education*, *4*(1). https://doi.org/10.1080/2331186X.2017.1416898

Burfitt, J. (2017). Partial credit in multiple-choice items. *40 years on: We are still learning! Proceedings of the 40th Annual Conference of the Mathematics Education Research Group of Australasia*, 117–124.

Chandrasegaran, A. L., Treagust, D. F., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, *8*(3), 293–307. https://doi.org/10.1039/B7RP90006F

Cross, K. P., & Angelo, T. A. (1993). Classroom assessment techniques: A handbook for faculty. *The National Center for Research to Improve Post-*

secondary Teaching and Learning. http://www.scopus.com/inward/record.url?eid=2-s2.0-84873961428&partnerID=tZOtx3y1

Derya Kaltakci. (2012). Development and Application of a Four-Tier Test to Assess Pre-Service Physics Teachers Misconceptions About Geometrical Optics. In *Middle East Technical University* (Nomor September). Middle East Technical University.

Dewi, S. Z., & Ibrahim, T. (2019). Pentingnya Pemahaman Konsep Untuk Mengatasi Miskonsepsi Dalam Materi Belajar IPA di Sekolah Dasar. *Jurnal Pendidikan UNIGA*, *13*(1), 130–136. http://dx.doi.org/10.52434/jpu.v17i1.2 553

Engelhard, G., & Wind, S. (2017). Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments. In *Invariant Measurement with Raters and Rating Scales: Rasch Models for Rater-Mediated Assessments*. https://doi.org/10.4324/9781315766829

Holbeck, R., Bergquist, E., & Lees, S. (2014). Classroom Assessment Techniques: Checking for Student Understanding in an Introductory University Success Course. *Journal of Instructional Research*, *3*, 38–42. http://files.eric.ed.gov/fulltext/EJ1127694.pdf

Jannah, R., & Rahmi, I. (2020). Pengembangan E- Diagnostic Four Tier Test Untuk Mengidentifikasi Miskonsepsi Peserta Didik. *Natural Science*, *6*(2), 151–160. https://ejournal.uinib.ac.id/jurnal/index.php/naturalscience/article/view/1721

Kiray, S. A., & Simsek, S. (2021). Determination and Evaluation of the Science Teacher Candidates' Misconceptions About Density by Using Four-Tier Diagnostic Test. *International Journal of Science and Mathematics Education*, *19*(5), 935–955. https://doi.org/10.1007/s10763-020-10087-5

Leonard, D. (2024). *28 Ways to Quickly Check for Understanding From sketching comics to drafting tweets, these fun—and fast—ways to check for understanding are creative and flexible.* edutopia. https://www.edutopia.org/article/quick-ways-to-check-for-understanding/

Liliawati, W., Efendi, R., Purwana, U., & Muslim. (2022). Meningkatkan Konsepsi Asesmen Guru Fisika SMA Melalui Program Penguatan Kompetensi. *Online Submission*, *7*, 69–74. http://electronicportfolios.com/portfolios/njedgenet.pdf

Mardapi, D. (2020). *Teknik Penyusunan Instrumen Tes dan Non Tes*. Parama.

Resbiantoro, G., Setiani, R., & Dwikoranto. (2022). A Review of Misconception in Physics: The Diagnosis, Causes, and Remediation. *Journal of Turkish Science Education*, *19*(2), 403–427. https://doi.org/10.36681/tused.2022.128

Retnawati, H. (2020). *Validitas Reliabilitas & Karakteristik Butir*. Parama.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *International Journal of Phytoremediation*, *21*(1), 77–84. https://doi.org/10.1080/0969595980050104

Saputri, L., Maison, M., & Kurniawan, W. (2021). Pengembangan Four-Tier Diagnostic Test Berbasis Website untuk Mengidentifikasi Miskonsepsi pada Materi Suhu dan Kalor. *Jurnal Ilmiah Teknologi Informasi Asia*, *15*(1), 61. https://doi.org/10.32815/jitika.v15i1.563

Schuwirth, L. W. T., & Van Der Vleuten, C. P. M. (2011). Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher*, *33*(6), 478–485. https://doi.org/10.3109/0142159X.2011.565828

Setiawati, F. A., Mardapi, D., & Azwar, S. (2013). Penskalaan Teori Klasik Instrumen Multiple Intelligences Tipe Thurstone Dan Likert. *Jurnal Penelitian dan Evaluasi Pendidikan*, *17*(2), 259–274. https://doi.org/10.21831/pep.v17i2.1699

Suherly, T., Azizahwati, A., & Rahmad, M. (2023). Kemampuan Pemahaman Konsep Awal Siswa dalam Pembelajaran Fisika : Analisis Tingkat Pemahaman pada Materi Fluida Dinamis. *Jurnal Paedagogy*, *10*(2), 494. https://doi.org/10.33394/jp.v10i2.7239

Sumaryanta. (2021). Teori Tes Klasik dan Teori Respon Butir: Konsep dan Contoh Penerapannya. In *Cetakan Pertama* (Vol. 15, Nomor 2).

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan RASCH Pada Assessment Pendidikan*. Trim Komunikata.

Wiggins, G., & McTighe, J. (2005). *Understanding by Design* (D. Russel (ed.)). Julie Houtz.

Wulandari, K., Khoiroh, M., & Prihatiningtyas, S. (2023). Pengembangan Instrumen Penilaian Formatif Materi Usaha Dan Energi Pada Mata Pelajaran Fisika SMA/MA. *Diffraction*, *5*(1), 1–7. https://doi.org/10.37058/diffraction.v5i1.6216