Application of the RASCH model to analyze Critical Thinking Skills Instruments (CTSI) on static fluid concept

Lingga Thursina Fajriyati*, Duden Saepuzaman, Lina Aviyanti, Winny Liliawati

Department of Physics Education, Universitas Pendidikan Indonesia, Dr. Setiabudi Street, No.229, Bandung, 40154, Indonesia

*Corresponding author, email: lingga.t1@upi.edu

Article History
Received: 26 August 2025
Revised: 9 May 2026
Accepted: 13 June 2026

Keywords
Critical thinking skills
Instrument
Rasch model
Static fluid

Abstract
This study aims to analyze the reliability, validity of item analysis, and estimation of respondents' abilities in assessing students' critical thinking skills (CTS) instrument on static fluid material using the Rasch model. There are five indicators of the CTS instrument, namely Basic Clarification (BCL), Decision Basis (TBD), Inference (INF), Advanced Clarification (ACL), and Assumption and Integration (SIN), with a total of 20 items. The Rasch model was chosen because it provides an in-depth analysis of item characteristics, respondents' abilities, and allows the identification of imprecise or biased items. Although the theoretical background is presented briefly due to the methodological focus of this study, previous studies have primarily emphasized improving CTS through learning interventions rather than developing standardized measurement instruments. The research method used is quantitative with a descriptive design using Win steps 3.73 software for data analysis. The sample consisted of 215 high school students in grades XI and XII majoring in science, consisting of 120 female students and 95 male students in West Java. The results of the study showed that the test instrument had high reliability with an individual reliability of 0.73 and an item reliability of 0.92, as well as good internal consistency with a Cronbach's Alpha of 0.78. The validity of this instrument also met the acceptance criteria of the Rasch model, with MNSQ infit and MNSQ outfit values ranging from 0.5 to 1.5, indicating acceptable model fit. Therefore, the CTS test instrument is suitable for use in high school physics education. This study contributes to addressing the limited availability of validated CTSCTS instruments specifically designed for static fluid material at the high school level using Rasch analysis.

Fajriyati, L. T., Saepuzaman, D., Aviyanti, L., & Liliawati, W. (2026). Application of the RASCH model to analyze Critical Thinking Skills Instruments (CTSI) on static fluid concept. *Momentum: Physics Education Journal*, *10*(1), 47-66. https://doi.org/10.21067/mpej.v10i1.12442

1. Introduction

21st-century skills are essential for science learning today. One of the necessary skills is critical thinking. Critical thinking is a reflective and rational way of thinking that focuses on decisions about what to believe or do (Ennis, 2023; Skjølberg et al., 2023; Wilson, 2024). Additionally, critical thinking is a structured and disciplined process of actively and skillfully conceptualizing, applying, analyzing, synthesizing, and evaluating information gathered through observation, experience, reflection, reasoning, or communication as a guide for beliefs and actions (Paul & Elder, 2020; Irwanto, 2023; Kelsey & Starzec, 2024). This skill helps produce effective and targeted decisions or understanding (Halpern, 2014; Swiecki et al., 2019; Turan, 2019; Scott et al., 2021). Previous research indicates that many physics teachers have not effectively integrated learning strategies to develop students' critical thinking skills and abilities (Prafitasari et al., 2021; Pereira et al., 2023; Almeddine & Bashir, 2024). Based on a literature synthesis on 21st-century skills such as critical thinking, there remains a lack of research focusing on these skills in the context of physics (Tiruneh et al., 2017; Lei & Kathleen, 2019; Jamil et al., 2024).

The improvement of students' critical thinking skills must be continuously assessed as an indicator of success in the teaching process and as a basis for reflection to improve the quality of the learning process. One method for assessing critical thinking skills is through testing. To obtain data on critical thinking skills, a high-quality instrument is required. Many existing instruments are often not thoroughly tested for validity and reliability, resulting in inaccurate and unreliable evaluation outcomes (Amirzadeh et al., 2024). The Rasch model is more recommended for use in determining the quality of a test instrument (Dijkers & Millis, 2020; Killip et al., 2022; Mukhibin, et al., 2024). The Rasch model is part of Item Response Theory, which is included in Modern Test Theory or Latent

doi: 10.21067/mpej.v10i1.12442

ISSN: 2548-9135

Trait Theory. This theory is designed to measure latent psychological or educational attributes, such as critical thinking skills. Unlike Classical Test Theory, Modern Test Theory allows for the separation of test takers' abilities (latent traits) from item characteristics, thereby producing more objective, accurate, and invariant measurements across samples and testing instruments (van der Linden, 2018; Wilson, 2023; Embretson & Reise, 2025).). The Rash Model uses a one-parameter approach (IPL) focused on the difficulty level of test items and allows for the evaluation of the fit between empirical data and the theoretical model (fit statistics) as well as the detection of potentially biased items. Meanwhile, Classical Test Theory (CTT) has several fundamental weaknesses, such as the dependence of reliability and validity estimates on each item (Bichi & Talib, 2018; Stemler & Naples, 2021; Fergadiotis et al., 2023). Additionally, CTT assumes that all items contribute equally to the total score, meaning it does not accurately reflect the reality of measuring critical thinking skills. Therefore, the use of the Rasch Model is more suitable for developing modern instruments such as critical thinking skills and physics lessons.

The Rasch model is capable of producing objective and independent measurements from both samples and items, exchanging raw scores into a linear scale that is easy to interpret, and assisting in validating and improving instruments through the detection of unsuitable items (Wind & Hua, 2022; Yıldırım Hoş, H., & Uysal Saraç, M, 2023). The Rasch model provides researchers with the opportunity to conduct more in-depth and comprehensive analyses, including detecting items that are inappropriate or biased, as well as evaluating the internal consistency of an instrument (Planinic et al., 2019; Samsudin et al., 2021; Liu et al., 2023). Therefore, the application of this model can improve the accuracy and reliability of evaluation results while providing a better understanding of the instrument's performance across various contexts.

The comprehensive use of the Rasch Model, which serves not only to evaluate but also to improve the quality of assessment instruments (Setyawarno et al., 2025). The results of this study contribute significantly to physics education by providing more accurate and reliable evaluation instruments to measure students' critical thinking skills. It can make a significant contribution to the field of physics education by providing more accurate and reliable evaluation tools to measure students' reflective thinking abilities. This study aims to analyze the use of the Rasch Model in the development of instruments for measuring critical thinking skills in static fluid physics material. The purpose of this study is to analyze the use of the Rasch Model in the development of instruments for measuring critical thinking skills in static fluid physics material.

The selection of static fluid material in this study has strong scientific urgency. Concepts in static fluids such as hydrostatic pressure, Pascal's principle, and Archimedes' principle are abstract physics concepts (Loverude et al., 2003; Kaltakci-Gurel et al., 2017). A lack of understanding of these concepts can directly impact students' critical thinking skills in solving problems related to static fluid phenomena (Wieman & Holmes, 2015; Setiawan & Faoziyah, 2020). The urgency of selecting this topic in this study is also reinforced by previous research discussing the development of CTS in physics lessons. The problem-solving-based teaching approach to improve CTS is general in the context of science (Tiruneh et al., 2017). Additionally, the inquiry-discovery learning model significantly improves students' critical thinking skills in physics compared to conventional methods, by inspiring independent problem-solving, deeper understanding, improved cognitive performance, and reduced misconceptions, particularly in static fluid material (Hudha & Batlolona, 2017). CTS is also important in physics learning for problem-solving and concept understanding, as guided inquiry learning encourages students to analyze, synthesize, and evaluate information, thus enhancing their understanding and application of physics principles (Maknun, 2020).

The difference between this study and previous studies is that it focuses on developing CTS instruments for static fluid material using Rasch analysis. While previous studies generally focused on improving CTS through the application of specific learning models, this study emphasizes the development of standardized measurement instruments (Putra et al., 2023; Ismail et al., 2022). The novelty of this study lies in the application of the Rasch Model to the development and analysis of CTS test items, which allows for a more objective and comprehensive evaluation of validity, reliability, item difficulty, and student ability distribution. Furthermore, this study addresses the limited availability of CTS instruments specifically designed for static fluids at the high school level and systematically validated using Rasch analysis.

Based on the objectives of the study and comparison with previous studies, the formulation of the research questions includes: 1) To what extent does the critical thinking skills test instrument on statistical fluid material meet the validity and reliability criteria based on the Rasch Model analysis?, 2) How can the application of the Rasch Model improve the quality of students' critical thinking skills test instruments on statistical fluid material in physics lessons?, 3) How does the Rasch Model analysis identify potentially biased items in the test instrument?.

2. Method

This study uses quantitative methods and descriptive design to determine students' critical thinking skills in learning Physics. This method systematically alanyzes data by measuring students' abilities and skills using structured instruments (Creswell, 2014; Worachak et al., 2023; Juandi et al., 2024). The instrument consisted of 20 multiple-choice items covering five CTS indicators: Basic Clarification (BCL), Decision Basis (TBD), Inference (INF), Advanced Clarification (ACL), and Conjecture and Integration (SIN). All items were scored dichotomously, with 1 for a correct answer and 0 for an incorrect answer.

Content validity was established through expert assessment involving physics education experts who evaluated items for relevance, indicator alignment, and language clarity. Based on their suggestions, several items were revised to improve clarity and contextual accuracy before administration. A pilot test was also conducted, and items that did not meet the criteria were revised or removed before final data collection.

Data analysis was conducted using Winsteps 3.73 software based on the Rasch model. This model was chosen because it provides a comprehensive analysis of item characteristics, respondent abilities, and allows for the detection of inappropriate or biased items (Boone, Staver, & Yale, 2014; Natanael et al., 2023). The analysis included item reliability, person reliability, Cronbach's Alpha, item fit statistics (infit and outfit MNSQ), estimation of respondent ability, and Differential Item Functioning (DIF).

The population in this study were high school students in grades XI and XII of science specialization in West Java. The sample was selected using random sampling technique to ensure data representation in the population (Taherdoost, 2016; Etikan & Bala, 2017). Participants came from high schools in West Java including Bandung, Bekasi, Tasikmalaya and Cirebon consisting of 215 participants with a female sample of 120 students and a male sample of 95 students. The geographical distribution of participants is shown in Figure 1 and the sample distribution is shown in Table 1.

Figure 1. Geographic Distribution of Participants

Table 1. Sample Distribution

Gender	Person	Presentation
Female	120	55.8%
Male	95	44.2%
Total	215	100%

Students are asked to complete a test consisting of questions designed to measure critical thinking skills. The test instrument in this study tested critical thinking skills about static fluid concept consisting of 20 multiple choice questions. Scoring is complete using a rubric that refers to critical thinking indicators, with each question given a score of 0-1 for incorrect or correct answers (Donoghue et al., 2022; Fadillah et al., 2023; Zhang et al., 2023; Tasçi, 2024). The question indicators are adjusted to the indicators of the aspects and sub-aspects of CTS according to Ennis (2011), as shown in Table 2 below:

Table 2. Aspect and CTS Sub-Aspect

No	CTS Aspect	CTS Sub-Aspect	Item
1	Basic Clarification (BCL)	Formulating a question.	1, 2
		Analyzing arguments.	3, 4, 5
		Asking and answering clarification questions.	6, 7
2	The Bases for a decision (TBD)	Considering the credibility of a source.	8
		Observing and considering observation results.	9
3	Inference (INF)	Making and considering deductions.	10, 11
		Making and considering inductions.	12, 13
		Making and considering value judgments.	14
4	Advanced Clarification (ACL)	Identifying and considering definitions.	15, 16
		Referring to unstated assumptions.	17, 18
5	Supposition and integration (SIN)	Considering and thinking logically, premises, reasons, assumptions, positions, and other suggestions.	19
		Integrating other skills and dispositions in making and defending a decision.	20

As a concrete illustration of the types of questions that have been developed, Figure 2 shows an example of an item that presents the aspect of Basic Clarification (BCL), with the sub-aspect of formulating a question, located in item number 1.

Take a look at the following picture!

A construction worker uses nails to drive through wood. He realizes that the pointed end of the nail makes the nailing process easier than if it were blunt.

Based on the information above, the most appropriate critical question to develop is:

A. How to calculate the pressure exerted by the nail on the wood?
B. What is the relationship between the surface area of the nail tip and the pressure generated?
C. Why do nails with pointed tips penetrate wood more easily than blunt ones?
D. How does pressure affect the speed and ease of nailing into wood?
E. What are the factors that affect the pressure exerted by the nail on the wood surface?

Figure 2. Example of Critical Thinking Skills Test Questions Static Fluid Concept

Before the test instrument is widely used, systematic steps need to be taken to ensure that the instrument has an adequate level of validity and reliability. The instrument development process begins with a literature study, which examines relevant theories and research results to formulate indicators and constructs to be measured. Next is the preparation of the initial draft of the instrument based on the indicators that have been formulated, then conducting a limited trial to 50 students in class XI. At this stage, the items of the instrument were evaluated from the aspects of readability, language clarity, content conformity with indicators, as well as allowing bias or ambiguity. The data from the initial trial were then analyzed to assess the validity of the items and the reliability of the instrument. If items were found to be invalid, inconsistent, or not in accordance with the indicators,

the instrument was revised either by repairing, replacing, or eliminating problematic items. After revision, the instrument was tested again in a pilot test on a larger sample. The test results are used to conduct more in-depth statistical analysis such as Rasch Model analysis to ensure that the instrument is suitable for use in the main research. By following these stages or cycles systematically, it can ensure that the instruments used are truly capable of producing valid, reliable, and objective data in the context of quantitative research (Boone, Staver, & Yale, 2014; Tutz, 2023). The reliability and consistency of the instrument were analyzed using the Rasch model through the reliability coefficient and respondent fit statistics (fit person) (Linacre, 2020). The test instrument reliability criteria in the Rasch Model according to Sumnintono & Widhiarso (2015), are shown in Table 3 as follows:

Table 3. Interpretation of Reliability Value

Reliability Value (Person/Item)	Interpretation
$r > 0.94$	Special
$0.91 \leq r \leq 0.94$	Very good
$0.80 \leq r < 0.90$	Good
$0.67 \leq r < 0.80$	Simply
$r < 0.67$	Weak

(Sumintono & Widhiarso, 2015)

The validity of the instrument in the Rasch Model is seen by the Cronbach's Alpha value, which identifies the overall reliability and reflects the relationship between students and the questions given (Tavakol & Dennick, 2011). The reliability criteria for test instruments in the Rasch Model according to Sumnintono & Widhiarso (2015), shown in Table 4.

Table 4. Interpretation of Cronbach's Alpha Value

Cronbach's Alpha Value	Interpretation
$\alpha \geq 0.80$	Very good
$0.70 \leq \alpha < 0.80$	Good
$0.60 \leq \alpha < 0.70$	Simply
$0.50 \leq \alpha < 0.60$	Bad
$\alpha < 0.50$	Very bad

(Sumintono & Widhiarso, 2015)

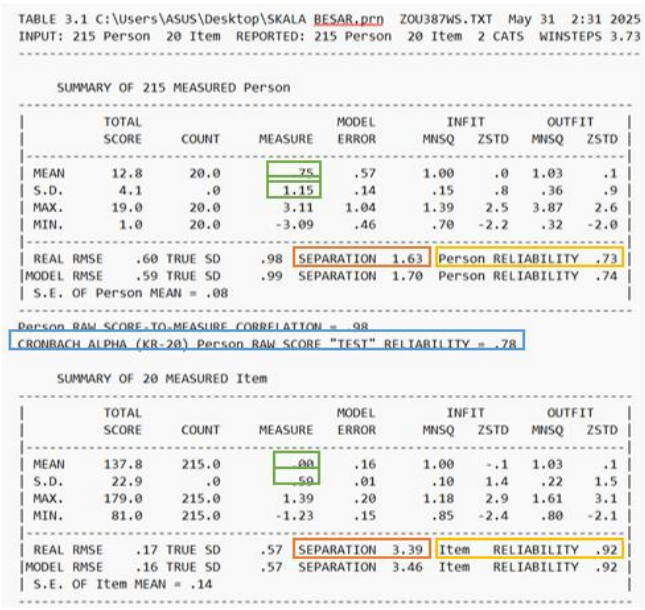
In addition, this study also analyzed the Wright Map, the measurement of student ability (Person Measure), and the suitability of student answers to the Rasch model through Person Fit indicators such as Outfit MNSQ, ZSTD, and Point Measure correlation. The Wright Map helps illustrate the distribution of student ability and question difficulty on the same scale. Person Fit is used to identify responses that do not fit the ideal Rasch pattern (Abdulridah et al., 2022; Handayani et al., 2023; Humphry, 2023; Yamada, 2025), while Person Measure is used to assess student ability. Respondents' abilities are estimated based on item response patterns, resulting in logit scores that are objective and comparable (Wright & Stone, 1999; Laliyo et al., 2020; Altun et al., 2021).

3. Results and Discussion

3.1. Reliability and Consistency

3.1.1. Reliability (Variability) Analysis

In the context of the Rasch Model, reliability not only reflects internal consistency, but also reflects the relationship between students' abilities and skills and the characteristics of a given item. Therefore, reliability analysis is very important to ensure that the instrument can truly be used in objective, accurate and precise measurement. The results of the instrument reliability analysis are shown in Figure 3.



```
TABLE 3.1 C:\Users\ASUS\Desktop\SKALA BESAR.prn  ZOU387WS.TXT  May 31  2:31 2025
INPUT: 215 Person  20 Item  REPORTED: 215 Person  20 Item  2 CATS  WINSTEPS 3.73

     SUMMARY OF 215 MEASURED Person

|          TOTAL                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-------------------------------------------------------------------------------|
| MEAN      12.8      20.0         .75     .57      1.00     .0   1.03     .1 |
| S.D.       4.1        .0        1.15     .14       .15     .8    .36     .9 |
| MAX.      19.0      20.0        3.11    1.04      1.39    2.5   3.87    2.6 |
| MIN.       1.0      20.0       -3.09     .46       .70   -2.2    .32   -2.0 |
|-------------------------------------------------------------------------------|
| REAL RMSE    .60 TRUE SD    .98  SEPARATION  1.63  Person RELIABILITY  .73   |
|MODEL RMSE    .59 TRUE SD    .99  SEPARATION  1.70  Person RELIABILITY  .74   |
| S.E. OF Person MEAN = .08                                                     |

Person RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .78

     SUMMARY OF 20 MEASURED Item

|          TOTAL                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-------------------------------------------------------------------------------|
| MEAN     137.8     215.0         .00     .16      1.00    -.1   1.03     .1 |
| S.D.      22.9        .0         .59     .01       .10    1.4    .22    1.5 |
| MAX.     179.0     215.0        1.39     .20      1.18    2.9   1.61    3.1 |
| MIN.      81.0     215.0       -1.23     .15       .85   -2.4    .80   -2.1 |
|-------------------------------------------------------------------------------|
| REAL RMSE    .17 TRUE SD    .57  SEPARATION  3.39  Item   RELIABILITY  .92   |
|MODEL RMSE    .16 TRUE SD    .57  SEPARATION  3.46  Item   RELIABILITY  .92   |
| S.E. OF Item MEAN = .14                                                       |
```

Figure 3. Reliability Analysis

Based on Figure 3 summary results of measurements on 215 respondents and there are 20 items of question instruments using Rasch analysis, it can be concluded that the instrument has a fairly good quality and reflects the consistency of the instrument in differentiating the ability of respondents (Bond & Fox, 2015; Linacre, 2020). This is supported by the results in Table 4, where the person reliability value of 0.73 indicates that participants' abilities are measured consistently and can be able to identify differences between individuals quite well. Table 4 follows. Reliability Analysis.

Table 4. Reliability Analysis

	Logit Mean (SD)	Separation	Reliability	Cronbach's Alpha
Person	0.75 (1.15)	1.63	0.73	0.78
Item	0.00 (0.59)	3.39	0.92	

A person separation value of 1.63 indicates that the instrument is able to group students into two different ability levels, which is important for the purpose of diagnosis and classification of participants' abilities (Wilson, 2018; Linacre, 2020). This means that differences in student ability can be detected quite clearly by the instrument used. Meanwhile, for the items, the reliability reached 0.92, which falls into the very high category. This shows that each item in the instrument makes a consistent contribution in measuring the intended construct, namely critical thinking skills. The average logit item of 0.00 with a standard deviation of 0.59 also indicates that the difficulty level of the questions is spread equally according to the distribution of student abilities, so that the instrument is not biased towards students with low or high abilities (Lin et al., 2021; Yalinkilic & Gul, 2023). In addition, the Cronbach Alpha coefficient of 0.78 strengthens the internal consistency between items, indicating that the items in the instrument are well related and measure the same thing, namely critical thinking skills (Zhang et al., 2019; Wolfs et al., 2023).

Therefore, the results of this reliability analysis indicate that the developed instrument is classified as reliable and valid, and is suitable for use to measure student abilities objectively, stably and consistently, in accordance with the Rasch Model approach. The Rasch Model itself emphasizes the importance of the relationship between item characteristics and individual abilities in producing measurements that are invariant, that is, the measurement results are not influenced by a particular population or by a particular item. Therefore, high reliability in the Rasch Model not only reflects the stability of the data, but is also an important indicator that an instrument has fair and generalizable qualities. In other words, this instrument is not only able to provide consistent results, but also accurate and meaningful, so that it can be used to measure student abilities in a larger and more diverse context.

3.1.2. Individual Suitability Level

The level of individual fit is used to evaluate whether a test taker's response pattern to items is appropriate based on an ideal measurement model. If the student has a high ability, it is expected to answer easy and medium questions correctly. Conversely, if there is a mismatch between the student's response pattern and his/her ability, it indicates a misfit. For this purpose, the Person Fit Order table in the Winstep 3.73 output is used, especially in the "Item Fit Statistic" section, which shows statistical values such as Infit Mean Square (MNSQ) and Outfit MNSQ. The ideal values of these statistics are in the range of 0.5 to 1.5. Values outside this range may indicate that students are answering inconsistently or showing unnatural patterns (Sumintono & Widhiarso, 2015; Meijer & Sijtsma, 2001). Therefore, the level of individual fit can also be used to detect possible cheating or technical errors during test taking (Student, 2022; Balta & Dogan, 2024). The results of students' Person Fit Order analysis are shown in Figure 4.

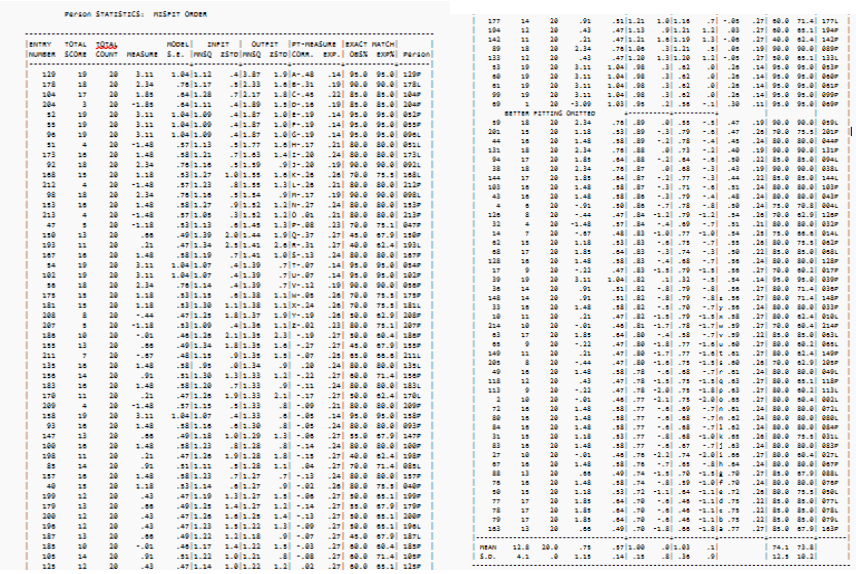


Figure 4. Person Fit Order

Based on Figure 4, the results of the analysis of the level of individual suitability for critical thinking skills (Sumintono et al., 2015). The distribution of student in the unsuitable category is shown in Table 5.

Table 5. Student distribution that misfit the model

Student Ability Level	Person	%
Doesn't meet 2 indicators (Outfit MNSQ & ZSTD)	163P, 0,79L, 078L, 077L	1.86
Doesn't meet 2 indicators (Outfit MNSQ & PT Corr.)	129P, 178L, 104P, 204P, 052P, 055P, 096L, 051L, 173L, 092L, 168L, 212P, 098L, 153P, 213P	6.97

This distribution is then visualized in Figure 5, to clarify the proportion of participants classified as fit or misfit based on the predefined categories.

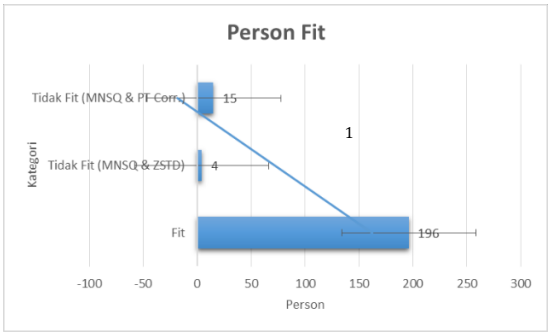


Figure 5. Diagram Person Fit Order

Based on Table 5 which is visualized in Figure 5 about the results of the analysis of the level of individual fit using the Rasch Model, it is known that there are a number of students answering instrument items that are not in accordance with the expectations of the model. This mismatch is identified based on three main indicators, namely Outfit Mean Square (MNSQ), Z-standardized (ZSTD), and Point Measure Correlation (PT Corr.). There were 4 students who did not meet the two Outfit MNSQ and ZSTD indicators including 163P, 0.79L, 078L, 077L, equivalent to 1.86% of the total respondents. This shows that the answers of these students are very deviant and inconsistent, most likely due to a lack of understanding of the items or answering randomly.

Meanwhile, there were 15 students including 129P, 178L, 104P, 204P, 052P, 055P, 096L, 051L, 173L, 092L, 168L, 212P, 098L, 153P, 213P who did not meet the two indicators of Outfit MNSQ and PT Corr., equivalent to 6.97% of the overall respondents. A low PT Corr. value of <0.20 indicates that students' responses to the items are not positively correlated with the level of ability being measured, identifying the possibility of answering not seriously or not according to actual ability (Bond & Fox, 2015; Köhler & Hartig, 2017).

From the results of the analysis, the overall number of students who showed misfit in two categories was 19, meaning that a small proportion of the sample did not provide reliable responses based on the model used. This result is important to consider in data interpretation as it can affect the validity of the findings and decisions made from the instrument (Sumintono & Widhiarso, 2015).

The level of individual fit or person fit in the Rasch Model is an important indicator to measure how well a student's response pattern is to the ability item being measured. Misfit occurs when students give answers that are inconsistent with the general pattern expected by the model, which may indicate cheating ignorance (Meijer & Sijtsma, 2001; Yalinkilic & Gul 2023). Misfit can compromise the validity of the measure results as inappropriate responses can provide biased ability estimates (Bond & Fox, 2015; Köhler & Hartig, 2017). Paying attention to person fit not only maintains measurement accuracy and fairness, but also obtains diagnostic information about respondents' behavior so that teaching can be more targeted.

3.2. Validity

3.2.1. Instrument Validity

The validity of item instruments in the Rasch Model is indicated by the ideal Infit and Outfit MNSQ values close to 1.0, with an acceptable range between 0.5 to 1.5. In addition, for ZSTD the ideal value is 0.0 and the tolerance is in the range of -2.0 to +2.0. These values indicate that the item is as expected and does not need revision (Sumintono & Widhiarso, 2015).

Instrument validity in the Rasch Model also reflects the extent to which items are able to consistently measure the intended construct. Infit and Outfit MNSQ values that are within the ideal range indicate that participants' responses to the items fit the Rasch Model, without significant deviation. ZSTD values between -2.5 and +2.0 indicate that response variation does not deviate from model expectations (Sumintono & Widhiarso, 2015; Bond et al, 2020).

In addition, item and person reliability indicators also support validity. High item reliability reflects that the items are able to distinguish participants based on their abilities in a stable manner. For this reason, good person reliability can show that the data generated are quite consistent and can be trusted in the interpretation of measurement results (Zou & Bolt, 2023; Avinç & Doğan, 2024). Therefore, the validity of the instrument is not only seen from the fit statistic value, but also from the overall consistency and fit of the items to the model. The results of the instrument validity analysis are shown in Figure 6.

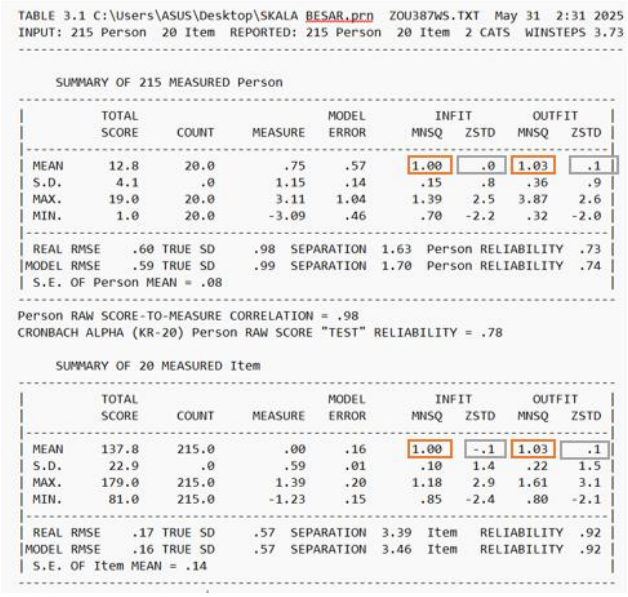


TABLE 3.1 C:\Users\ASUS\Desktop\SKALA BESAR.prn ZOU387WS.TXT May 31 2:31 2025
INPUT: 215 Person 20 Item REPORTED: 215 Person 20 Item 2 CATS WINSTEPS 3.73

SUMMARY OF 215 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	12.8	20.0	.75	.57	1.00	.0	1.03	.1
S.D.	4.1	.0	1.15	.14	.15	.8	.36	.9
MAX.	19.0	20.0	3.11	1.04	1.39	2.5	3.87	2.6
MIN.	1.0	20.0	-3.09	.46	.70	-2.2	.32	-2.0

REAL RMSE .60 TRUE SD .98 SEPARATION 1.63 Person RELIABILITY .73
MODEL RMSE .59 TRUE SD .99 SEPARATION 1.70 Person RELIABILITY .74
S.E. OF Person MEAN = .08

Person RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .78

SUMMARY OF 20 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	137.8	215.0	.00	.16	1.00	-.1	1.03	.1
S.D.	22.9	.0	.59	.01	.10	1.4	.22	1.5
MAX.	179.0	215.0	1.39	.20	1.18	2.9	1.61	3.1
MIN.	81.0	215.0	-1.23	.15	.85	-2.4	.80	-2.1

REAL RMSE .17 TRUE SD .57 SEPARATION 3.39 Item RELIABILITY .92
MODEL RMSE .16 TRUE SD .57 SEPARATION 3.46 Item RELIABILITY .92
S.E. OF Item MEAN = .14

Figure 6. Instrument Validity

Based on Figure 6, the Rasch analysis results show that 215 respondents have an average value of Infit MNSQ = 1.00 and Outfit MNSQ = 1.03. These values are within the acceptable range of 0.5 to 1.5, indicating that the participants' responses are relatively in accordance with the Rasch Model and no items should be revised or deleted even though there is a slight insignificant variation. As for the 20 items, the average values of Infit MNSQ = 1.00 and Outfit MNSQ = 1.03 were close to the ideal value of 1.0 indicating that the items were valid.

3.2.2. Construct Validity

When defining constructs, an important thing to consider is that the constructs must theoretically be unidimensional from low to high levels (Wei et al., 2012; Wang et al. 2025). In this study, the measured construct is students' critical thinking skills on static fluid material. Static fluid material was chosen because this material has been taught in class XI semester 1, so that class XI during semester 2 and XII have also received the material because a test can be given if the student has learned the material. Static fluid material tests can be given at two grade levels, namely grade XI and XII so that they can provide a comprehensive picture of critical thinking skills.

Construct validity in the Rasch Model refers to the instrument actually measuring the theoretical construct and is tested through three main approaches including unidimensionality, item fit statistics, and point-measure correlation. Unidimensionality is tested using residual analysis, where a second component eigenvalue smaller than 2.0 indicates that all items measure one main construct. Fit statistics such as Outfit and Infit MNSQ that were within the range of 0.5-1.5 and ZSTD between -2.0 to +2.0 indicated that the items did not deviate from the model. In addition, the point-measure correlation value above 0.3 indicates that each item contributes positively and consistently to the measurement of the intended construct, so the instrument is considered constructively valid (Sumintono & Widhiarso, 2015). The results of the construct validity analysis are shown in Figure 7.

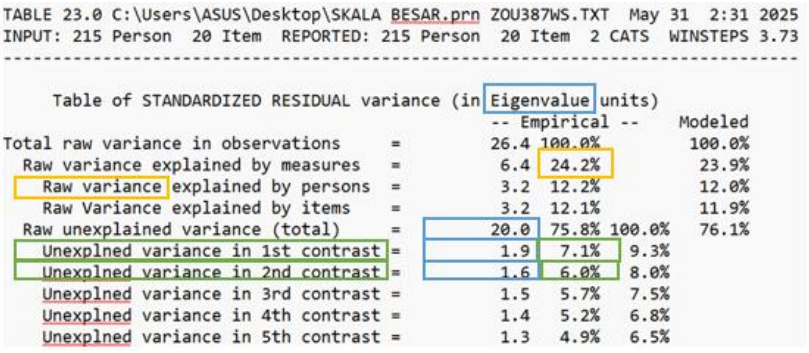


TABLE 23.0 C:\Users\ASUS\Desktop\SKALA BESAR.prn ZOU387WS.TXT May 31 2:31 2025
INPUT: 215 Person 20 Item REPORTED: 215 Person 20 Item 2 CATS WINSTEPS 3.73

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

	Eigenvalue	-- Empirical --		Modeled
Total raw variance in observations =	26.4	100.0%		100.0%
Raw variance explained by measures =	6.4	24.2%		23.9%
Raw variance explained by persons =	3.2	12.2%		12.0%
Raw Variance explained by items =	3.2	12.1%		11.9%
Raw unexplained variance (total) =	20.0	75.8%	100.0%	76.1%
Unexplned variance in 1st contrast =	1.9	7.1%	9.3%	
Unexplned variance in 2nd contrast =	1.6	6.0%	8.0%	
Unexplned variance in 3rd contrast =	1.5	5.7%	7.5%	
Unexplned variance in 4th contrast =	1.4	5.2%	6.8%	
Unexplned variance in 5th contrast =	1.3	4.9%	6.5%	

Figure 7. Construct Validity

Based on Figure 7 Rasch analysis results, the construct validity of this instrument can be declared good. This is indicated by the raw variance explained by measures value of 24.2%, which exceeds the minimum standard of 20%, so that the items in this instrument substantially measure one main construct well (Boone et al., 2014; Sumintono & Widhiarso, 2015; Linacre, 2022).

Unidimensionality in the context of Rasch analysis refers to the assumption that all items in an instrument measure the same underlying dimension or construct (Embretson & Reise, 2000; DeVellis, 2017; Bond et al., 2021). If unidimensionality is met, then participants' scores can be meaningfully interpreted as representing one main ability, rather than a combination of several other dimensions. In the Rasch Model, unidimensionality is evaluated through Principal Component Analysis (PCA) residual analysis. An instrument is considered unidimensional if the raw variance explained by measures ≥ 20%, and the unexplained variance in the 1st contrast < 15% and eigenvalue < 3.0 (Linacre, 2020; Sumintono & Widhiarso, 2015).

In addition, the value of unexplained variance in the 1st contrast was recorded at 7.1% with an eigenvalue of 1.9, both of which are still below the critical limit (maximum 15% and eigenvalue < 3.0), indicating that there are no additional dimensions strong enough to indicate multidimensionality. Although the total unexplained variance was 75.8%, this value is still acceptable in the context of Rasch analysis if the variance explained meets the minimum criteria. Therefore, it can be concluded that this instrument has good construct validity as it is generally able to measure one main construct consistently. Therefore, this instrument has good construct validity as it is generally able to measure one main construct consistently and accurately.

3.3. Question Item Analysis

3.3.1. Item Parameter Analysis

Item parameter analysis is an important aspect of the Rasch Model as it allows for the evaluation of item difficulty as well as the accuracy of the instrument in discriminating between participants' abilities. The level of item difficulty is indicated through the measure value, where the ideal value is in the range -2.0 to +2.0 logits. This range indicates that the items are suitable for effectively measuring variations in student ability. Items that have measure values outside this range are considered too easy or too difficult, making them less informative for measurement (Sumintono & Widhiarso, 2015; Boone et al., 2014; Bond & Fox, 2015). The results of the item parameters analysis are shown in Figure 8.

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
15	81	215	1.39	.16	1.12	1.8	1.32	2.5	.33	.44	66.0	70.9	S15
18	98	215	.98	.15	.96	-.7	.89	-1.1	.49	.45	68.8	69.3	S18
17	106	215	.80	.15	.96	-.6	.96	-.3	.48	.45	69.3	69.2	S17
20	111	215	.68	.15	1.18	2.9	1.32	3.1	.29	.45	63.3	69.2	S20
14	126	215	.32	.16	.85	-2.4	.80	-2.1	.56	.45	79.5	70.6	S14
16	134	215	.13	.16	1.08	1.2	1.03	.3	.39	.44	67.4	71.7	S16
6	135	215	.10	.16	.85	-2.4	.85	-1.4	.56	.44	79.1	71.9	S6
4	138	215	.03	.16	.94	-.9	.85	-1.4	.50	.44	72.1	72.4	S4
13	141	215	-.05	.16	.98	-.2	.93	-.6	.46	.44	73.0	73.0	S13
5	144	215	-.13	.16	.85	-2.2	.80	-1.6	.55	.44	79.1	73.7	S5
9	144	215	-.13	.16	.89	-1.5	.90	-.8	.52	.44	77.2	73.7	S9
19	146	215	-.18	.16	1.11	1.4	1.28	2.0	.34	.43	71.6	74.2	S19
10	147	215	-.21	.16	1.01	.2	.98	-.1	.43	.43	74.9	74.4	S10
2	148	215	-.24	.16	1.02	.3	.97	-.2	.42	.43	73.5	74.6	S2
1	149	215	-.26	.16	.99	-.1	.89	-.8	.45	.43	74.0	74.9	S1
12	149	215	-.26	.16	.97	-.4	.92	-.5	.46	.43	76.7	74.9	S12
7	157	215	-.49	.17	1.08	1.0	1.19	1.2	.35	.42	74.0	77.0	S7
11	157	215	-.49	.17	1.06	.7	1.33	2.0	.35	.42	75.8	77.0	S11
8	166	215	-.76	.18	.93	-.7	.88	-.6	.45	.40	81.4	79.7	S8
3	179	215	-1.23	.20	1.11	1.0	1.61	2.3	.24	.37	85.1	84.4	S3
MEAN	137.8	215.0	.00	.16	1.00	-.1	1.03	.1			74.1	73.8	
S.D.	22.9	.0	.59	.01	.10	1.4	.22	1.5			5.3	3.6	

Figure 8. Item Parameters

Based on Figure 8, the analysis shows that the most difficult item is item S15 with a measure value of +1.39 logits, while the easiest item is item S3 with a measure value of -1.23 logits, resulting in a difficulty range of 2.62 logits. This range indicates that the items cover a good variety of difficulty levels from very easy to very difficult, which is important to reflect the diverse abilities of test takers (Giguère et al., 2023; Prihartono et al., 2021). Furthermore, the measurement accuracy of each item is checked through the Standard Error of Measurement (SEM) value. The ideal SEM is less than 0.5

logits, with a category <0.3 logits considered excellent. In this analysis, the SEM of all items ranged from 0.16 to 0.20 logits, which means that the items have high accuracy in estimating students' abilities. Low SEM values indicate good measurement reliability and the ability to distinguish high and low ability participants consistently (Sumintono & Widhiarso, 2015; Linacre, 2021). The number of respondents also affects the accuracy of parameter estimation. In this study, the number of respondents of 215 was considered ideal enough to obtain stable and reliable item parameter estimates (Bond & Fox, 2015; Linacre, 2021). Therefore, it can be concluded that all items in this instrument have difficulty and accuracy parameters that meet the criteria in the Rasch model, so they can be used to measure student abilities thoroughly and fairly.

3.3.2. Item Fit with Rasch Model

Data is said to fit the Rasch model if it meets three main indicators, namely: the Outfit Mean Square (MNSQ) value is in the range of 0.5 to 1.5 (ideally close to 1.0), the Z-standardized fit statistic (ZSTD) value is between -2.0 to +2.0 (ideally close to 0.0), and the Point Measure Correlation (PT-Measure Corr.) value is above 0.3 and positive (Sumintono & Widhiarso, 2015; Bond & Fox, 2015; Linacre, 2020). These three indicators show the extent to which each item in the instrument contributes validly to the measurement of respondents' abilities. The analysis results for these variables are shown in Figure 9.

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
3	179	215	-1.23	.20	1.11	1.0	1.49	2.0	A	.24	.37	85.1	84.4	S3
11	157	215	-.49	.17	1.06	.7	1.33	2.0	B	.35	.42	75.8	77.0	S11
20	111	215	.68	.15	1.18	2.9	1.32	2.0	C	.29	.45	63.3	69.2	S20
15	81	215	1.39	.16	1.12	1.8	1.32	2.0	D	.33	.44	66.0	70.9	S15
19	146	215	-.18	.16	1.11	1.4	1.28	2.0	E	.34	.43	71.6	74.2	S19
7	157	215	-.49	.17	1.08	1.0	1.19	1.2	F	.35	.42	74.0	77.0	S7
16	134	215	.13	.16	1.08	1.2	1.03	.3	G	.39	.44	67.4	71.7	S16
2	148	215	-.24	.16	1.02	.3	.97	-.2	H	.42	.43	73.5	74.6	S2
10	147	215	-.21	.16	1.01	.2	.98	-.1	I	.43	.43	74.9	74.4	S10
1	149	215	-.26	.16	.99	-.1	.89	-.8	J	.45	.43	74.0	74.9	S1
13	141	215	-.05	.16	.98	-.2	.93	-.6	j	.46	.44	73.0	73.0	S13
12	149	215	-.26	.16	.97	-.4	.92	-.5	i	.46	.43	76.7	74.9	S12
17	106	215	.80	.15	.96	-.6	.96	-.3	h	.48	.45	69.3	69.2	S17
18	98	215	.98	.15	.96	-.7	.89	-1.1	g	.49	.45	68.8	69.3	S18
4	138	215	.03	.16	.94	-.9	.85	-1.4	f	.50	.44	72.1	72.4	S4
8	166	215	-.76	.18	.93	-.7	.88	-.6	e	.45	.40	81.4	79.7	S8
9	144	215	-.13	.16	.89	-1.5	.90	-.8	d	.52	.44	77.2	73.7	S9
14	126	215	.32	.16	.85	-2.4	.80	-2.0	c	.56	.45	79.5	70.6	S14
5	144	215	-.13	.16	.85	-2.2	.80	-1.6	b	.55	.44	79.1	73.7	S5
6	135	215	.10	.16	.85	-2.4	.85	-1.4	a	.56	.44	79.1	71.9	S6
MEAN	137.8	215.0	.00	.16	1.00	-.1	1.03	.1				74.1	73.8	
S.D.	22.9	.0	.59	.01	.10	1.4	.22	1.5				5.3	3.6	

Figure 9. Item Fit with Rasch Model

Based on Figure 9, it can be seen that all items have MNSQ and ZSTD Outfit values that are within the ideal limits, which are between 0.5 to 1.5 for MNSQ and -2.0 to +2.0 for ZSTD. This indicates that there are no items that are noisy or do not fit the model to the extreme, so there are no items that overfit or underhit significantly. This means that all items are within the tolerance limits and do not need to be deleted or modified (Boone et al., 2014; Müller, 2020). In addition, the PT-Measure Corr. values of all items were above 0.3 and all were positive, indicating that the items made a consistent contribution to the overall measurement of student ability.

A good item fit to the Rasch Model is very important because it shows that the items measure the same construct and strengthens the validity of the instrument (Hambleton et al., 1991; Guler et al., 2022). Therefore, it can be concluded that the instruments used in this study as a whole have met the requirements of the Rasch Model fit and are suitable for use to measure student abilities reliably and validly.

3.3.3. Different Item Functioning (DIF)

In Differential Item Functioning (DIF) analysis, an item is said to contain bias if a significant p-value (Prob.) <0.05 indicates that the difference in response between groups is not caused by differences in ability, but by the characteristics of the item itself (Zumbo, 1999; Boone et al., 2014; Hagquist & Andrich, 2017). DIF analysis in the Rasch Model is conducted to ensure that the instrument is measurement fairness and does not favor certain groups (Zumbo, 1999; Holland & Wainer, 2012). The results of the analysis for DIF are shown in Figure 10.

DIF class specification is: DIF=$S1W1

Person CLASSES	SUMMARY DIF CHI-SQUARE	D.F.	PROB.	BETWEEN-CLASS MEAN-SQUARE	t=ZSTD	Item Number	Name
3	4.1359	2	.1242	.9138	.2445	1	S1
3	2.6773	2	.2588	.3827	-.4887	2	S2
3	1.6639	2	.4317	.3388	-.5753	3	S3
3	3.7247	2	.1528	.8666	.1936	4	S4
3	4.2857	2	.1152	.9718	.3049	5	S5
3	1.2317	2	.5372	.2118	-.8783	6	S6
3	4.0082	2	.1325	.8852	.2138	7	S7
3	.8884	2	.6393	.0700	-1.4302	8	S8
3	4.3221	2	.1131	.4234	-.4139	9	S9
3	1.9734	2	.3692	.3876	-.4793	10	S10
3	4.3602	2	.1110	.4223	-.4158	11	S11
3	1.2369	2	.5358	.0837	-1.3542	12	S12
3	3.3415	2	.1853	.6667	-.0459	13	S13
3	.3780	2	.8284	.0854	-1.3454	14	S14
3	3.5193	2	.1700	5.6212	2.6675	15	S15
3	3.4194	2	.1781	.6053	-.1289	16	S16
3	3.0138	2	.2224	2.8236	1.5736	17	S17
3	4.5004	2	.1000	6.2585	2.8619	18	S18
3	1.5440	2	.4586	.3523	-.5479	19	S19
3	3.2520	2	.2072	1.3704	.6656	20	S20

Figure 10. Different Item Functioning (DIF)

Based on Figure 10, the research findings indicate that all items on the statistical fluid instrument exhibited no indication of gender bias, as they had a DIF probability value greater than 0.05. This result indicates that the items performed fairly across both male and female students, so the probability of correct answers was determined by student ability, not gender. From a Rasch model perspective, this condition indicates measurement invariance, or measurement invariance, namely the instrument's ability to maintain the same measurement meaning across different groups (Navarro-González et al., 2024). DIF analysis is also an important approach to ensuring instrument fairness because it can identify the possibility of items favoring certain groups. These results align with research by Afifa et al. (2024), which demonstrated that items without gender-specific DIF can measure student ability more objectively and consistently without affecting specific demographic characteristics.

The absence of gender bias in the statistical fluid instrument also indicates that the item construction was well-designed so that the context, language, and question representation did not favor the experiences of one gender. This is crucial in the development of educational science instruments because biased items can distort estimates of student ability and reduce the validity of test result interpretations. According to Plouffe et al. (2021), items detected with gender bias have the potential to benefit certain groups and therefore need to be revised or eliminated to maintain fairness. Furthermore, DIF-free instruments reflect the objective and invariant characteristics of Rasch measurement, namely the instrument's ability to produce stable ability estimates across different respondent groups (Boone et al., 2014; Hagquist & Andrich, 2017; Linacre, 2020). Therefore, the results of this study reinforce the merits of statistical fluid instruments for use in educational assessments because they are capable of measuring students' conceptual abilities more accurately, fairly, and reliably.

3.3.4. Scalogram

Scalogram analysis, which has its roots in the Guttman scaling method, is essential in evaluating the construct validity and reliability of a measurement instrument. Scalograms are used to look at the pattern of participants' responses to items and determine whether the responses follow a certain hierarchical pattern according to the Guttman model. In the Guttman model, it is assumed that each item has a different level of difficulty, and if a participant is able to correctly answer a more difficult item, then that participant should be able to correctly answer an easier item (Guttman, 1944; Engelhard, 2009; Antino et al., 2020). Therefore, the scalogram can help identify whether a participant's answer pattern matches the expectations of the hierarchical model, as well as detect deviant or random responses. This method is also closely related to Rasch analysis, where scalograms can be used as a visual way to verify data from Rasch model estimation (Bond & Fox, 2015). Scalograms allow observers to directly see participants' consistency in answering, identifying respondents with perfect, slightly deviant, or random patterns (Meijer & Sijtsma, 2001; Engelhard, 2009; Dirlik & Kartal, 2022). The results of the analysis are shown in Figure 11, Figure 12 and Figure 13.

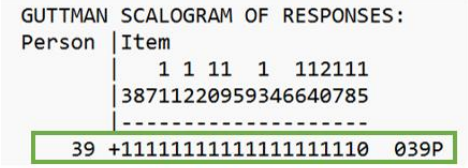


Figure 11. Perfect Pattern

Based on Figure 11, participant 039P showed a perfect answer pattern. No deviations were found. All her responses conformed to the hierarchical pattern of the Guttman model, characterized by a consistent and consecutive row of 1s. The data from this participant can be used as a reference to compare other participants' patterns. This perfect pattern identifies high consistency and thorough understanding of the items presented (Guttman, 1944).

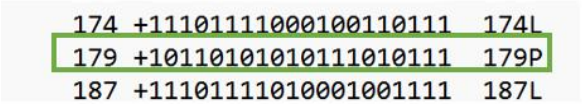


Figure 12. Slightly Deviating Pattern

Based on Figure 12, participant 179P showed a pattern that deviated slightly from the ideal pattern. This is indicated by 1 (correct) after a row of 0 (incorrect), which should not occur in a Guttman pattern. This small deviation could be due to the participant's inaccuracy when answering, but overall it is still acceptable in Rasch-based assessment (Guttman, 1944; Meijer & Sijtsma, 2001; Engelhard, 2009).

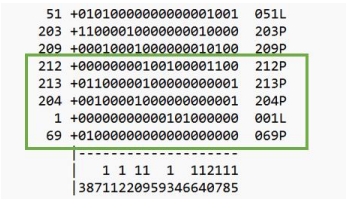


Figure 13. Random Pattern

Based on Figure 13, answer patterns such as participants 212P, 213P, 204P, 001L and 069P show illogical or overly random responses. They do not reflect the hierarchy of the question, which is most likely due to a lack of consistency in answering or guesswork. This pattern indicates that participants did not understand the content of the question or answered randomly, and their responses cannot be used for valid inference (Guttman, 1944; Meijer & Sijtsma, 2001; Bond & Fox, 2015). Scalogram analysis is important in evaluating participants' responses to items in a measurement instrument. By identifying perfect, deviant, and random patterns, it is possible to assess the quality of the data and the validity of the test results in greater depth.

3.4. Estimated Ability of Respondents

3.4.1. Capability Parameters

In the analysis, the ability parameters in the Rasch Model are expressed in logit units with a range between -3 to +3 logits. A positive logit value indicates that a person has above-average ability, while a negative logit value indicates below-average ability. The higher the logit value, the greater the probability of a participant answering an item correctly, because the ability level is greater than the difficulty of the item (Sumintono & Widhiarso, 2015; Bond & Fox, 2015).

The Rasch model positions participants and items on the same scale, enabling direct interpretation of participants' abilities in relation to the difficulty level of the questions (Boone et al., 2014; Sumintono & Widhiarso, 2015; Bond & Fox, 2015). This creates a measurement system that is invariant and objective, unaffected by the distribution of raw scores. The results of the analysis of the highest and lowest ability parameters are shown in Figure 14

Person 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%	Person
39	19	20	3.11	1.04	.82	.1	.32	-.5	.54	.14	95.0	95.0	039P
52	19	20	3.11	1.04	1.09	.4	1.87	1.0	-.19	.14	95.0	95.0	052P
53	19	20	3.11	1.04	.98	.3	.62	.0	.26	.14	95.0	95.0	053P
212	4	20	-1.48	.57	1.23	.8	1.55	1.3	-.26	.21	80.0	80.0	212P
213	4	20	-1.48	.57	1.05	.3	1.52	1.2	.01	.21	80.0	80.0	213P
204	3	20	-1.85	.64	1.11	.4	1.89	1.5	-.16	.19	85.0	85.0	204P
1	2	20	-2.32	.76	1.08	.3	1.16	.5	-.01	.16	90.0	90.0	001L
69	1	20	-3.09	1.03	.95	.2	.56	-.1	.30	.11	95.0	95.0	069P
MEAN	12.8	20.0	.75	.57	1.00	.0	1.03	.1			74.1	73.8	
S.D.	4.1	.0	1.15	.14	.15	.8	.36	.9			12.5	10.2	

Figure 14. Highest and Lowest Ability Parameters

Participants with code 039P had the highest ability among all participants, as indicated by a logit value of +3.11. This logit value shows that these participants were able to answer high-difficulty items with a total score of 19 out of a maximum of 20, indicating an almost perfect level of mastery of the material. Conversely, participant with code 069P has the lowest ability among all participants, as indicated by a logit score of -3.09. This logit score indicates that the participant experienced significant difficulty in answering the items on the instrument, further supported by a total score of 1 out of a maximum of 20, indicating very low mastery of the tested material.

The results of the ability parameter analysis using the Rasch Model provide an objective and standardized picture of the participants' level of mastery of the material being tested. The extreme difference in logit scores between participants 039P and 069P reflects significant variation in ability within the population, thereby providing a strong basis for decision-making in education, such as implementing specialized interventions, improving assessment tools, or adjusting the difficulty level of questions. To support visual understanding of the distribution of participants' abilities, Figure 15 shows a scatter plot. This scatter plot illustrates the distribution of participants' abilities based on logit values, which show the spread of ability levels from very low to very high.

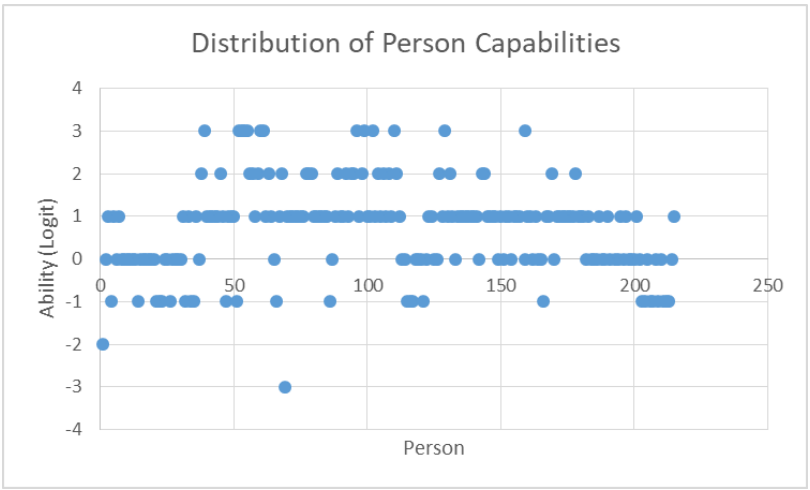


Figure 15. Distribution of Person Capabilities

In line with previous findings stating that Rasch analysis can fairly map students' abilities (Boone et al., 2014; Sumintono & Widhiarso, 2015), it can also provide a strong basis for developing data-driven learning policies. Educators can identify students who require additional support, such as developing more targeted remedial programs, and adapting instructional materials to varying levels of ability. Additionally, these findings can be utilized in evaluating the quality of assessment instruments to ensure that the test items effectively measure students' abilities. Therefore, the Rasch Model not only impacts measurement aspects but also contributes to the overall improvement of educational quality.

3.4.2. Wright Map

In the analysis of student ability, it can be done by visualizing data from WINSTEP 3.73 software through the Variable Maps display or better known as the Wright Map. This visualization illustrates the distribution of student ability levels as well as the difficulty of the questions on the same scale, namely the logit scale. This mapping allows educators to see the position of students based on their level of understanding, ranging from high to low categories, as well as the extent to which the questions given are challenging for students. Thus, Wright Map becomes an important diagnostic tool to evaluate students' abilities and provide more targeted interventions. When it comes to critical thinking skills, the information from Wright Map is very relevant (Boone et al., 2014; Sumintono & Widhiarso, 2015; Facione, 2020). The results of the Wright Map analysis are shown in Figure 16.

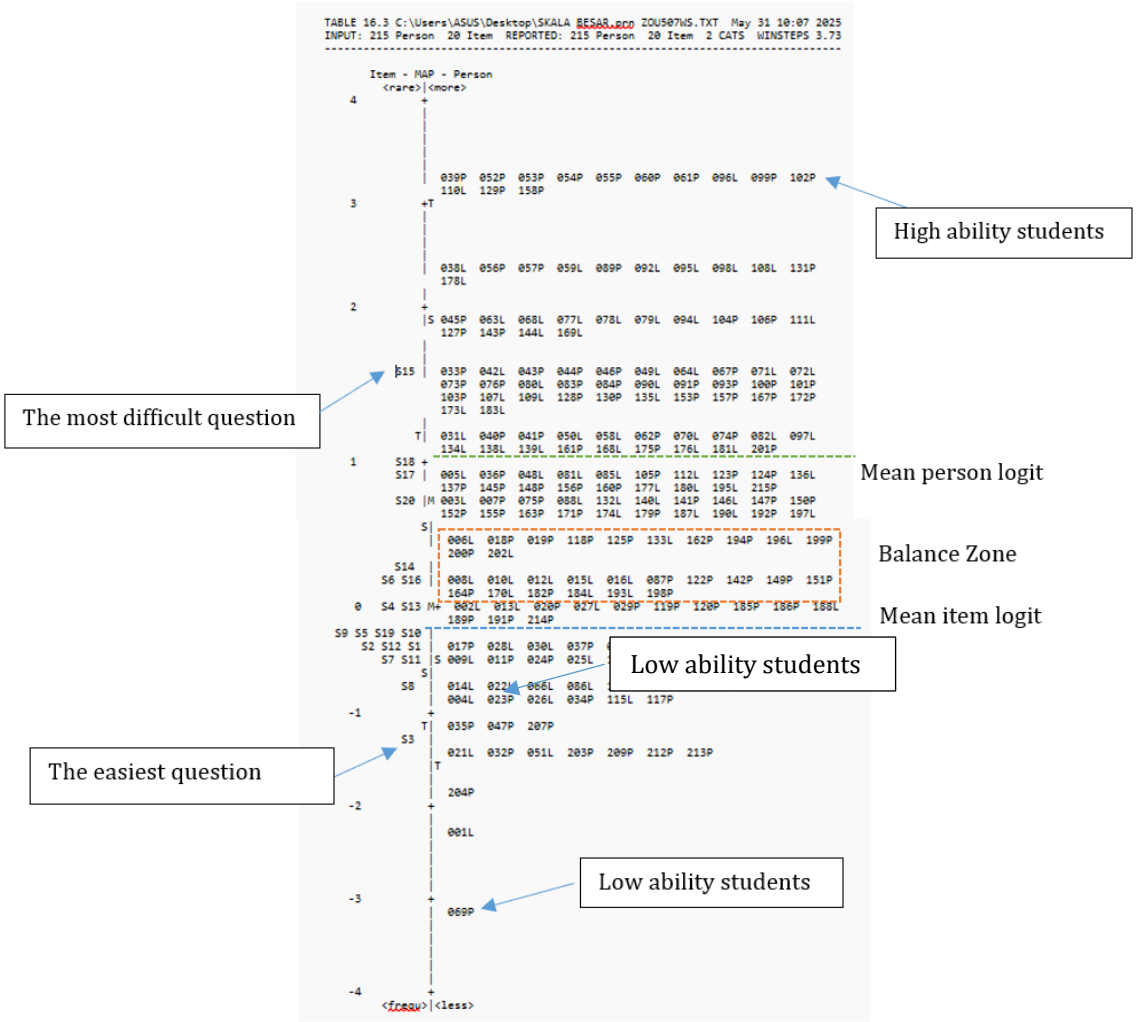


Figure 16. Wright Map

Figure 16 provides a comprehensive visual representation of the distribution of student ability (person) and item difficulty on the same logit scale. The vertical axis shows logit units representing item difficulty on the left side and student ability on the right side. Students with the highest ability (smartest) are found at the top of the student respondents with codes 039P, 052P, 053P, 054P, 055P, 060P, 061P, 096L, 099P, 102P, 110L, 129P, and 158P. Meanwhile, the students with the lowest ability (weakest) were at the bottom of the student respondents with code 069P. The most difficult items are located at the top of the left side with question code S15, and the easiest items are at the bottom of the left side with question code S3. Thus, this map shows the extent to which students' abilities are proportional to the level of difficulty of the questions.

The midpoint on the Wright Map, referred to as the balance zone, is the position where average student ability meets average item difficulty (Xie & Lim, 2015; Bond, Yan, & Heene, 2021; Chien, Lin, & Wang, 2022; Andrich, 2023). The midpoint is of course important because it is an indicator of the

extent to which the question matches the ability of the majority of students (Giguère et al., 2023). On the map, there are three problems including S6, S14, S16 and twenty-six students including 006L, 008L, 010L, 012L 015L, 016L, 018P, 019P, 087P, 118P, 122P, 125P, 133P, 142P, 149P, 162P, 164P, 170L, 182P, 184L, 193L, 194P, 196L, 198P, 200P, 202L who are in the balance point, This means that twenty-six students are likely to have abilities that match the three questions and show a balance between student abilities and the level of difficulty of the questions which can then be the basis for understanding the pattern of student understanding of certain questions. If there are some students below this zone, it means that the questions given are too difficult and not proportional to the overall ability of the students. Conversely, if many students are above this zone, it is likely that the questions given are too easy and not challenging enough.

Wright Map analysis is very relevant to the context of developing critical thinking skills. If students are able to answer questions with a high level of difficulty, it is likely that they have reached more complex levels of thinking such as analysis, evaluation, and synthesis. Conversely, students who are only able to do easy problems may still be at a basic level of thinking. Therefore, this map can also be used as a basis for designing adaptive and challenging learning according to students' critical thinking levels (Sumintono & Widhiarso, 2015; Geller & Neumann, 2018; Walsh et al., 2019; Facione, 2020).

4. Conclusion

The results of this study indicate that the critical thinking ability test instrument on static fluid material meets the validity and reliability criteria based on the Rasch Model analysis. The instrument shows that the item reliability is very good (0.92), good internal consistency (Cronbach's Alpha 0.78), and adequate respondent reliability (0.73). In addition, all items meet the fit criteria so that it is valid for measuring students' critical thinking skills. The application of the Rasch Model also improves the quality of the instrument through more accurate item calibration, objective estimation of respondents' abilities, and better evaluation of internal consistency. In addition, the DIF analysis shows that there are no items that experience significant bias between respondent groups because all items have a DIF probability value of more than 0.05, so the instrument is fair and objective. Thus, the instrument is suitable for use to measure students' critical thinking skills in physics learning, especially static fluid material, and contributes to the development of more accurate and bias-free assessment instruments through the application of the Rasch Model.

Author Contributions

All authors have equal contributions to the paper. All the authors have read and approved the final manuscript.

Funding

No funding support was received.

Declaration of Conflicting Interests

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

References

Abdulridah Dhyaaldian, S. M., Hasan Al-Zubaidi, S., A Mutlak, D., Raheem Neamah, N., ALI ALBEER, A. A. M., A Hamad, D., ... & Ghaleb Maabreh, H. (2022). Psychometric evaluation of cloze tests with the Rasch model. *International Journal of Language Testing*, *12*(2), 95-106.

Afifa, M., Khoirunnisa, R., Pratiwi, S. M. V., & Meitaza, D. (2024). Utilizing Rasch Model to Analyze A Gender Gap in Students' Scientific Literacy on Energy. *Jurnal Pendidikan Fisika Indonesia*, *20*(1), 85-95.

Alameddine, M. A., & Bashir, M. M. (2024). *Investigating Strategies for Teaching Critical Thinking in Physics Classrooms. American J Sci Edu Re*: AJSER-202.

Altun, S. A., Büyüköztürk, Ş., & Seheryeli, M. Y. (2021). Validity and Reliability Evidence of Professional Obsolescence Scale According to Different Test Theories. *International Journal of Assessment Tools in Education, 8*(2), 257-278.

Amirzadeh, S., Rasouli, D., & Dargahi, H. (2024). Assessment of validity and reliability of the Feedback Quality Instrument. *BMC Research Notes*, *17*(1), 227.

Andrich, D. (2023). Person–item distribution and the quality of measurement. *Measurement: Interdisciplinary Research and Perspectives, 21*(3), 145–163.

Antino, M., Alvarado, J. M., Asún, R. A., & Bliese, P. (2020). Rethinking the exploration of dichotomous data: Mokken scale analysis versus factorial analysis. *Sociological Methods & Research, 49*(4), 839–867. https://doi.org/10.1177/0049124118769090

Avinç, E., & Doğan, F. (2024). Digital literacy scale: Validity and reliability study with the rasch model. *Education and information Technologies*, 1-47.

Ayub, M. R. S. S. N., Istiyono, E., Munadi, S., Permadi, C., Pattiserlihun, A., & Sudjito, D. N. (2020). Analisa Penilaian Soal Fisika Menggunakan Model Rasch Dengan Program R:-. *Jurnal Sains Dan Edukasi Sains*, *3*(2), 46-52.

Balta, E., & Dogan, C. D. (2024). Investigation of Preknowledge Cheating via Joint Hierarchical Modeling Patterns of Response Accuracy and Response Time. *SAGE Open*, *14*(4), 21582440241297946.

Bichi, A. A., & Talib, R. (2018). Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development. *International Journal of Evaluation and Research in Education*, *7*(2), 142-151.

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

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge. https://doi.org/10.4324/9781315814698

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge. https://doi.org/10.4324/9781003035861

Chien, C. Y., Lin, C. C., & Wang, W. C. (2022). Visual diagnosis of person–item targeting using person–item maps. *Applied Psychological Measurement, 46*(6), 454–470.

Choi, S. W. (2014). A review of the Rasch measurement model in psychometrics. *Frontiers in Psychology*, 5, 1077.

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (4th ed.). SAGE Publications.

Dijkers, M. P., & Millis, S. R. (2020). The template for intervention description and replication as a measure of intervention reporting quality: Rasch analysis. *Archives of rehabilitation research and clinical translation, 2*(3), 100055.

Dirlik, E. M., & Kartal, S. (2022). The comparison of the dimensionality results provided by the automated item selection procedure and DETECT analysis. International Journal of Assessment Tools in Education, 9(4), 808–830. https://doi.org/10.21449/ijate.1059200

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications*. Sage publications.

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

Ennis, R. H. (2023). Critical Thinking Across the Disciplines. *Inquiry: Critical Thinking Across the Disciplines*, 38(2), 1–10.

Etikan, I., & Bala, K. (2017). Sampling and sampling methods. *Biometrics & Biostatistics International Journal*, *5*(6), 00149.

Facione, P. A. (2020). *Critical Thinking: What It Is and Why It Counts* (2020 update). Insight Assessment.

Fadillah, S. M., Ha, M., Nuraeni, E., & Indriyanti, N. Y. (2023). Exploring Confidence Accuracy and Item Difficulty in Changing Multiple-Choice Answers of Scientific Reasoning Test. *Malaysian Journal of Learning and Instruction (MJLI)*, *20*(2), 319-341.

Putra, P. D. A., Sulaeman, N. F., Supeno, & Wahyuni, S. (2023). Exploring students' critical thinking skills using the engineering design process in a physics classroom. *The Asia-Pacific Education Researcher*, *32*(1), 141-149.

Fergadiotis, G., Casilio, M., Dickey, M. W., Steel, S., Nicholson, H., Fleegle, M., ... & Hula, W. D. (2023). Item response theory modeling of the verb naming test. *Journal of Speech, Language, and Hearing Research, 66*(5), 1718-1739.

Giguère, G., Brouillette-Alarie, S., & Bourassa, C. (2023). A look at the difficulty and predictive validity of LS/CMI items with Rasch modeling. *Criminal Justice and Behavior*, *50*(1), 118-138.

Hagquist, C., & Andrich, D. (2017). Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health and quality of life outcomes*, *15*(1), 181.

Halpern, D. F. (2014). *Thought and Knowledge: An Introduction to Critical Thinking* (5th ed.). Psychology Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Handayani, S., Sukmawati, E., & Rahmawati, D. (2023). Analysis of Students' Responses Using Rasch Model in Physics Learning. *Journal of Educational Measurement and Evaluation*, 15(2), 123–132.

Hudha, M. N., & Batlolona, J. R. (2017). How are the physics critical thinking skills of the students taught by using inquiry-discovery through empirical and theorethical overview*?. Eurasia Journal of Mathematics, Science and Technology Education*, 14(2), 691-697.

Humphry, S., Montuoro, P., & Maxwell, C. (2024). Cumulative ordering as evidence of construct validity for assessments of developmental attributes. *Journal of Psychoeducational Assessment, 42*(1), 60-73.

Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. Routledge.

Irwanto, I. (2023). Improving Preservice Chemistry Teachers' Critical Thinking and Science Process Skills Using Research-Oriented Collaborative Inquiry Learning. *Journal of Technology and Science Education*, *13*(1), 23-35.

Ismail, S. N., Muhammad, S., Omar, M. N., & Shanmugam, K. S. (2022). THE PRACTICE OF CRITICAL THINKING SKILLS IN TEACHING MATHEMATICS: TEACHERS'PERCEPTION AND READINESS. *Malaysian Journal of Learning and Instruction*, *19*(1), 1-30.

Jamil, M., Hafeez, F. A., & Muhammad, N. (2024). Critical thinking development for 21st century: Analysis of Physics curriculum. *Journal of Social & Organizational Matters*, *3*(1), 1-10.

Juandi, T., Kaniawati, I., Samsudin, A., & Riza, L. (2024). Prospective teachers' perception of critical and reflective thinking skills on modern physics: Rasch Analysis. *Journal for the Education of Gifted Young Scientists*, *12*(3), 137-150.

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 & Technological educaTion*, *35*(2), 238-260.

Kamilah, D. S., Muki, B. G., Aviyanti, L., & Suhandi, A. (2025). Review of misconceptions in physics among Indonesian high school students: Diagnosis, causes, and remediation. *Momentum: Physics Education Journal, 9*(1), 144–162. https://doi.org/10.21067/mpej.v9i1.11056

Kelsey Hall, E. D., & Starzec, K. (2024). Using an Interrupted Case Study to Engage Undergraduates' Critical Thinking Style and Enhance Content Knowledge. *Journal on Empowering Teaching Excellence, Spring 2024*, 46.

Killip, S. C., MacDermid, J. C., Wouters, R. M., Sinden, K. E., Gewurtz, R. E., Selles, R. W., & Packham, T. L. (2022). Rasch analysis of the brief Michigan Hand Questionnaire in patients with thumb osteoarthritis. *BMC Musculoskeletal Disorders*, *23*(1), 551.

Köhler, C., & Hartig, J. (2017). Practical significance of item misfit in educational assessments. *Applied Psychological Measurement*, *41*(5), 388-400.

Laliyo, L. A. R., Tangio, J. S., Sumintono, B., Jahja, M., & Panigoro, C. (2020). Analytic Approach of Response Pattern of Diagnostic Test Items in Evaluating Students' Conceptual Understanding of Characteristics of Particle of Matter. *Journal of Baltic Science Education*, *19*(5), 824-841.

Lin, J., Li, H., & Wang, Y. (2021). Analyzing item difficulty and test reliability in educational measurement: A Rasch model approach. *Journal of Educational Measurement*, 58(2), 120-135.

Linacre, J. M. (2020). *A user's guide to Winsteps: Rasch-model computer programs*. Winsteps.com.

Liu, J., Sun, M., Liu, Z., & Xu, Y. (2023). Pre-Service Teachers' Instructional Innovation Capabilities: A Many-Faceted Rasch Model Analysis. *SAGE Open*, 13(2), 21582440231218802. https://doi.org/10.1177/21582440231218802

Lei, K., and Kathleen, M. (2019). The gap in research on critical thinking skills in physics education. *Physics Education Research*, 15(3), 234-245.

Loverude, M. E., Kautz, C. H., & Heron, P. R. (2003). Helping students develop an understanding of Archimedes' principle. I. Research on student understanding. *American Journal of Physics*, *71*(11), 1178-1187.

Mukhibin, A., Rusyid, H. K., Lutfi, A., Herman, T., & Utomo, D. A. S. (2023). An Analysis of Students' Mathematical Self-Efficacy Instruments Using Rasch Model. *Indonesian Journal of Mathematics Education*, 6(2), 72–80. https://doi.org/10.31002/ijome.v6i2.994

Müller, M. (2020). Item fit statistics for Rasch analysis: can we trust them? *Journal of Statistical Distributions and Applications*, 7(5).

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25(2), 107–135.

Moore, M., & Gordon, P. C. (2015). Reading ability and print exposure: Item response theory analysis of the author recognition test. *Behavior research methods*, 47, 1095-1109.

Natanael, Y., Salsabilla, R., Aulia, D., Khoirunnisa, D., Munawar, H. N., & Hidayat, N. S. (2023). Rasch Rating Scale Model: Bias Detection and Validation Test of Indonesian-Adolescent Life Satisfaction Scale. *ResearchGate*.

Navarro-González, M. C., Padilla, J. L., & Benítez, I. (2024). Analyzing measurement invariance for studying the gender gap in educational testing: A mixed studies systematic review. *European Journal of Psychological Assessment*.

Paul, R., and Elder, L. (2020). *Critical Thinking: Tools for Taking Charge of Your Learning and Your Life* (4th ed.). Rowman & Littlefield.

Pereira, V. V., Samsudin, A., & Utama, J. A. (2023). STUDENT WORKSHEETS OF PBL AND PROBING PROMPTING TECHNIQUE ON CRITICAL THINKING SKILLS. *Journal of Teaching and Learning Physics*, *8*(2), 89-98.

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), 020111.

Plouffe, R. A., Kowalski, C. M., Tremblay, P. F., Saklofske, D. H., Rogoza, R., Di Pierro, R., & Chahine, S. (2021). Gender Differences or Gender Bias?. *European Journal of Psychological Assessment*.

Prafitasari, F., Sukarno, S., & Muzzazinah, M. (2021). Integration of critical thinking skills in science learning using blended learning system. *International Journal of Elementary Education*, *5*(3), 434-445.

Samsudin, A., Cahyani, P. B., Rusdiana, D., Efendi, R., Aminudin, A. H., & Costu, B. (2021). Development of a Multitier Open-Ended Work and Energy Instrument (MOWEI) Using Rasch Analysis to Identify Students' Misconceptions. *Cypriot Journal of Educational Sciences*, *16*(1), 16-32.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial*. Trim Komunikata.

Setiawan, D., & Faoziyah, N. (2020). Development of a five-tier diagnostic test to reveal the student concept in fluids. *Physics Communication*, *4*(1), 6-13.

Setyawarno, D., Maryati, & Natadiwijaya, I. F. (2025). Promoting a valid question model for measuring computational thinking skills based on confirmatory factor analysis and Rasch model. *Cogent Education*, *12*(1), 2505339.

Scott, I. A., Hubbard, R. E., Crock, C., Campbell, T., & Perera, M. (2021). Developing critical thinking skills for delivering optimal care. *Internal Medicine Journal*, *51*(4), 488-493.

Skjølberg, K. H., Trysnes, I., & Furrebø, E. F. (2023). Is the Coronavirus Created by the Government to Control Us? Critical Thinking and Conspiracy Beliefs among Norwegian Youth in Upper Secondary Schools. *Journal of Social Science Education, 22*(4), n4.

Student, S. R. (2022). Appraising Traditional and Purpose-built Person Fit Statistics' Power to Detect Cheating. *Chinese/English Journal of Educational Measurement and Evaluation*/ 教育测量与评估双语期刊, 3(1), 3.

Stemler, S. E., & Naples, A. (2021). Rasch measurement v. item response theory: Knowing when to cross the line. *Practical Assessment, Research & Evaluation*, *26*, 11.

Swiecki, Z., Ruis, A. R., Gautam, D., Rus, V., & Williamson Shaffer, D. (2019). Understanding when students are active-in-thinking through modeling-in-context. *British journal of educational technology, 50*(5), 2346-2364.

Taherdoost, H. (2016). Sampling methods in research methodology; how to choose a sampling technique for research. *International Journal of Academic Research in Management*, 5(2), 18–27.

Tasçi, G. (2024). Development of a Protein Concept Inventory: A Proposal for Item Scoring and Responding. *Science Insights Education Frontiers, 23*(2), 3755-3777.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Tiruneh, D. T., De Cock, M., Weldeslassie, A. G., Elen, J., & Janssen, R. (2017). Measuring critical thinking in physics: Development and validation of a critical thinking test in electricity and magnetism. *International Journal of Science and Mathematics Education,* 15, 663-682.

Turan, U., Fidan, Y., & Yıldıran, C. (2019). Critical thinking as a qualified decision-making tool.

Tiruneh, D. T., De Cock, M., & Elen, J. (2018). Designing learning environments for critical thinking: examining effective instructional approaches. *International journal of science and mathematics education*, *16*, 1065-1089.

Tutz, G. (2023). Unidimensionality in Rasch Models: Efficient item selection and hierarchical clustering methods based on marginal estimates. *arXiv preprint arXiv:2309.00553*.

Van der Linden, W. J. (Ed.). (2018). *Handbook of item response theory: Three volume set*. CRC press.

Walsh, C., Quinn, K. N., Wieman, C., & Holmes, N. G. (2019). Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking. *Physical Review Physics Education Research*, *15*(1), 010135.

Wang, X., & Chen, L. (2020). Validity and reliability of assessment tools using Rasch analysis in healthcare research. *Nursing Research*, 69(3), 213-221.

Wang, J., & Tam, T. (2025). *Bringing generalized status back in: Cross-national evidence for a unidimensional measure*. *Social Indicators Research*. https://doi.org/10.1007/s11205-025-03595-w

Wieman, C., & Holmes, N. G. (2015). Measuring the impact of an instructional laboratory on the learning of introductory physics. *American Journal of Physics*, 83(11), 972-978.

Embretson, S. E., & Reise, S. P. (2025). *Item response theory: Foundations for psychologists and social scientists*. Routledge.

Wilson, M. (2023). *Constructing measures: An item response modeling approach*. Routledge.

Wilson, K., & Defianty, M. (2024). The critical challenge for ELT in Indonesia: Overcoming barriers in fostering critical thinking in testing-oriented countries. *TESOL in Context*, *33*(1), 82-96.

Wind, S., & Hua, C. (2022). *Rasch measurement theory analysis in R*. Chapman and Hall/CRC.

Wei, S., Liu, X., Wang, Z., & Wang, X. (2012). Using rasch measurement to develop a computer modeling-based instrument to assess students' conceptual understanding of matter. *Journal of Chemical Education*, 89(3), 335–345.

Wolfs, Z. G., Brand-Gruwel, S., & Boshuizen, H. P. (2023). Assessing Tonal Abilities in Elementary School Children: Testing Reliability and Validity of the Implicit Tonal Ability Test Using Rasch Measurement Model. *SAGE Open*, *13*(3), 21582440231199041.

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials*. Wide Range, Inc.

Xie, Q., Zhong, X., Wang, W. C., & Lim, C. P. (2014). Development of an item bank for assessing generic competences in a higher-education institute: a rasch modelling approach. *Higher Education Research & Development*, *33*(4), 821-835.

Yalinkilic, F., & Gul, S. (2023). Development an achievement test on the subject of "Basic Compounds in the Structure of Living Things". *Science Insights Education Frontiers, 18*(2), 2905-2925.

Yamada, Y., Kobayashi, N., Wagman, P., & Håkansson, C. (2025). Validity and reliability of the Japanese version of the occupational balance questionnaire. *British Journal of Occupational Therapy*, 03080226251329771.

Yıldırım Hoş, H., & Uysal Saraç, M. (2023). A Mixture Rasch Model Analysis of Mathematics Achievement. *Kastamonu Education Journal*, *31*(1), 133–142. https://doi.org/10.24106/kefdergi.1246453

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF). *Ottawa: National Defense Headquarters*, *160*, 53.

Zou, T., & Bolt, D. M. (2023). Person misfit and person reliability in rating scale measures: The role of response styles. *Measurement: Interdisciplinary Research and Perspectives, 21*(3), 167-180.

Zhang, Y., Li, Z., & Zhao, X. (2019). Evaluating internal consistency of instruments: Cronbach's alpha and beyond. *Measurement and Evaluation in Counseling and Development*, 52(1), 42-56. https://doi.org/10.1080/07481756.2018.1486686

Zhang, M., Heffernan, N., & Lan, A. (2023). Modeling and Analyzing Scorer Preferences in Short-Answer Math Questions. *arXiv preprint arXiv:2306.00791*.