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

Vol. 27, No. 1, 2026
Article ID: 1200
Pages: 452-475

Measuring Critical Thinking in Physics: A Rasch Analysis of Instrument Quality and Gender Equivalence

Maria Goreti Halim[1,*], Duden Saepuzaman[1], Lina Aviyanti[1], Judhistira Aria Utama[1], & Abu Nawas[2]

[1]Department of Physics Education, Universitas Pendidikan Indonesia, Indonesia
[2]School of Education, Adelaide University, Australia

*Corresponding email: mariagoretihalim@upi.edu

Received: 05 January 2026 *Accepted: 04 March 2026* *Published: 03 April 2026*

Abstract: This study seeks to evaluate the quality of a Critical Thinking Skills (CTS) instrument for high school students on dynamic fluids, focusing on reliability, item validity, and respondent ability assessment using the Rasch model. This research utilized a quantitative technique with a descriptive design. The research sample comprised 200 11th-grade science students from several high schools in Manggarai Regency, East Nusa Tenggara, Indonesia, of whom 140 were female, and 60 were male. Data were analyzed with the Rasch modeling approach with the assistance of WINSTEPS software version 3.73. The findings indicated that the instrument exhibited generally acceptable psychometric properties, with a Cronbach's Alpha of 0.89 and a reliability of 0.93, indicating strong internal consistency and measurement stability. However, Rasch model analysis revealed that approximately 10% of the items did not fit the model expectations (misfit), around 15% indicated potential gender bias based on Differential Item Functioning (DIF) analysis, and 20,5% of respondents showed misfit response patterns. These results suggest that, while the overall reliability indices were high, certain items and response patterns require further refinement to achieve optimal measurement precision and fairness. The person reliability score of 0.83 indicated that the instrument reliably and accurately differentiated between varied levels of responder competence in assessing critical thinking abilities. In conclusion, this CTS instrument demonstrates overall acceptable measurement quality within the Rasch framework, although several psychometric limitations remain evident. These findings position the instrument as a preliminary yet functional assessment tool for measuring students' critical thinking skills in dynamic fluid topics, while highlighting the importance of continued empirical validation. Future studies are encouraged to expand the range of item difficulty, re-examine items exhibiting misfit and gender-related DIF, and involve more diverse samples to enhance measurement precision, fairness, and generalizability.

Keywords: critical thinking skills, Rasch model, dynamic fluids.

Article's DOI: https://doi.org/10.23960/jpmipa.v27i1.pp452-475

INTRODUCTION

Critical thinking skills have become a central focus in global education, particularly in addressing the challenges of the 21st century, which demand higher-order thinking, rational decision-making, and evidence-based problem-solving (Akhvlediani et al., 2023; Rusmin et al., 2024). Critical thinking is the capacity to critically evaluate information and formulate reasoned conclusions that shape beliefs and actions (Cottrell, 2017; Dwyer et al., 2014; Pasquinelli & Richard, 2023). It transcends rote memorization and encompasses advanced cognitive skills like interpretation, analysis, inference, assessment, explanation, and self-regulation (Facione, 2015).

In line with the demands of 21st-century education, critical thinking is consistently positioned as a core competency alongside creativity, collaboration, and communication

(Thornhill-Miller et al., 2023). In a digital landscape characterized by massive and diverse access to information, students are required not only to acquire knowledge but also to evaluate the credibility of sources, identify bias, and revise their understanding based on empirical evidence. Metacognitive regulation refers to the ability to consciously monitor and evaluate one's own thinking processes. This capacity has been shown to contribute significantly to the development of critical thinking by enhancing cognitive awareness and the strategic management of thinking processes (Rivas, S. F., Saiz, C., & Ossa, 2022). Through metacognitive regulation, students can recognize assumptions, detect logical inconsistencies, and adjust their conclusions to better align with the available evidence (Nunez et al., 2025). Therefore, critical thinking is understood as a reflective and evaluative process that integrates reasoning, evidence-based judgment, and self-regulation in a coherent manner.

In scientific education, critical thinking is vital as science involves not just the collection of information, but a process of knowledge building that necessitates sophisticated cognitive engagement (Morris et al., 2013; Stéphan Vincent-Lancrin, 2024). Critical thinking is also a key indicator of successful science learning, in which students are expected not only to memorize formulas but also to understand concepts, evaluate information, and solve contextual problems (García-Carmona, 2023; Santos, 2017). Studies indicate that augmenting critical thinking skills can enhance students' comprehension of academic disciplines, their capacity to apply information to practical scenarios, and their problem-solving abilities (Abrami et al., 2015; Liu et al., 2025; Razak, 2022; Rusmin et al., 2024). Therefore, critical thinking should be a primary goal in learning processes, especially in subjects like physics, which heavily rely on logical, quantitative, and applied reasoning.

Despite its recognized importance, many students still demonstrate low proficiency in critical thinking, particularly in science education (Arifin et al., 2025). Instructional practices still tend to emphasize procedural problem solving rather than deep conceptual reasoning, which may hinder its development (Arslan, 2010; Hurrell, 2021; Krange & Ludvigsen, 2008; Žakelj & Štemberger, 2025). Moreover, existing assessment instruments inadequately represent higher-order thinking, often prioritizing answer accuracy over reasoning quality. These limitations underscore the need for valid and reliable instruments to measure critical thinking in science, particularly in physics.

Students often perceive physics as a difficult subject due to its demand for deep conceptual understanding and advanced mathematical skills (Badmus & Jita, 2024; Christensen & Thompson, 2010). According to Young & Fredman, dynamic fluids-an area within fluid mechanics that focuses on moving fluids is among the most complex topics, requiring comprehension of models, principles, and fundamental laws such as Newton's laws and the conservation of energy (Young & Freedman, 2012). This topic is not only abstract but also highly applicable in real-world contexts, such as blood flow, aerodynamics, and piping systems. A superficial, mechanical understanding of this subject is inadequate; rather, an educational approach that fosters critical thinking about the concepts is essential.

In the context of dynamic fluids, critical thinking is necessary to connect theoretical concepts with everyday phenomena, evaluate experimental data, and construct logical scientific arguments (Bondarev et al., 2013; Etkina & Planinšiè, 2015). Students must be encouraged to analyze cause-and-effect relationships arising from changes in variables in fluid systems, explain the physical mechanisms underlying phenomena, and assess the assumptions embedded in existing mathematical models. These processes cannot be

optimized if instruction is limited to delivering formulas and calculation procedures without providing opportunities for exploration and reasoning. Hence, assessment approaches that can reveal students' critical thinking skills in a specific and structured manner are crucial for supporting meaningful science learning (Zuhaida et al., 2022).

Although several studies have developed assessment instruments in physics education, many remain general in scope and not specifically tailored to the conceptual demands of dynamic fluids. Instruments that do not align closely with the cognitive characteristics of a particular topic may fail to capture the authentic manifestation of students' critical thinking. Therefore, developing a topic-specific instrument is essential to ensure that the measurement accurately reflects the complexity of reasoning required to understand dynamic fluid topics.

A valid and reliable critical thinking test instrument is vital to objectively and comprehensively evaluate students' abilities. Various researchers have developed such instruments, based on existing frameworks, to measure students' levels of critical thinking in science education contexts (Sya'Bandari et al., 2018; Tiruneh et al., 2017). However, studies that focus on developing critical-thinking instruments specifically for dynamic fluids remain limited. Existing work primarily develops critical-thinking instruments for fluids in general or targets other constructs, such as problem-solving and cognitive learning outcomes in dynamic fluids (Alyaumusyifa et al., 2024; Nurdini et al., 2020; Sanjaya et al., 2024). This underscores the need for more context-specific tools that align with the subject matter's attributes to assess students' critical thinking with greater accuracy and depth.

In several previous studies, test instrument analysis has generally relied on the Classical Test Theory (CTT) approach due to its procedural simplicity and practical accessibility (Hambleton & Jones, 1993). Although CTT has been widely used for instrument evaluation, it has several limitations. CTT is highly dependent on the sample used, so its reliability and validity estimates may vary across different respondent characteristics. Furthermore, CTT does not permit the direct placement of student ability and item difficulty on a common measurement scale, limiting the interpretability of their relationship, as item difficulty indices are not located on a unified latent metric (Wright, 1999; Wright & Masters, 1982). Because CTT relies primarily on aggregated total scores, it provides limited diagnostic information at the individual and item levels, a limitation that becomes particularly critical when assessing higher-order constructs such as critical thinking in physics. These considerations have encouraged the adoption of modern measurement approaches, particularly the Rasch model, which offers more objective and stable parameter estimation across different samples.

The Rasch model enables the empirical validation of instrument functioning by placing item difficulty and student ability on a common logit scale, thereby allowing more precise interpretation of measurement results (Bond & Fox, 2015). Developing Rasch-based instruments for assessing critical thinking can improve assessment quality and help identify students' thinking weaknesses across specific skill indicators (Halimatun Sa et al., 2020; Kassiavera et al., 2024; Sujatmika et al., 2025). However, limited research has examined the item functioning of critical thinking instruments within specific physics domains, such as dynamic fluid concepts, using Rasch analysis to map both item quality and student ability within a unified measurement framework. Therefore, this study analyzes the item characteristics of a critical thinking skills exam on dynamic fluids using the Rasch model to develop a valid, reliable instrument that objectively maps both item quality and student ability.

Based on the background, literature review, and identified research gap, this study seeks to comprehensively examine the quality and characteristics of a critical-thinking test instrument for dynamic fluids. Therefore, the research questions of this study are as follows:

1. How is the quality of the developed critical thinking test instrument on dynamic fluids?
2. How do the characteristics of the test items and students' abilities reflect the measurement of critical thinking skills on dynamic fluids?
3. Does the developed critical thinking test instrument function fairly across different gender groups?

METHOD

Participants

The study population comprised all 11th-grade science students from several senior high schools in Manggarai Regency, East Nusa Tenggara. The research sample consisted of 200 11th-grade science students, including 140 females and 60 males. The sampling method used was purposive sampling, based on the following criteria: schools that had adopted training on dynamic fluids, had sufficient digital learning resources, and expressed willingness to participate in the study. Prior to data collection, coordination and permission were obtained directly from the physics teachers at each participating school. The test was administered online using Google Forms under teacher supervision during scheduled class activities. Before accessing the test link, students were informed about the study's purpose and assured that their responses would be used solely for research and would not affect their academic grades. Students completed the test individually, without discussion or external assistance, to ensure the authenticity of responses. Upon submission, responses were automatically recorded, exported to a spreadsheet, and coded using a dichotomous scoring system.

Research Design

This research utilized a quantitative methodology including a descriptive design. A descriptive study design seeks to systematically describe the attributes of a population without altering the variables under study (Nwabuko, Iwu, Njoku, & Nwamoh, 2024; Slater & Hasson, 2024). By collecting numerical data, this design enables the identification of patterns, prevalence, and distribution of the phenomena under study, while also providing an initial foundation for understanding trends and relationships among variables (Chen, 2021; Price & Lovell, 2018). In this study, the descriptive design was used to systematically, factually, and accurately describe the characteristics of critical thinking skill test items and the distribution of student abilities. Instrument validation was conducted employing the Rasch Model methodology, utilizing WINSTEPS software version 3.73, which facilitates a comprehensive study of item features and individual student proficiency levels.

Instrument

The instrument used in this study was a critical-thinking skills test consisting of 20 multiple-choice questions. Each item was developed based on Facione's critical thinking skills framework, which identifies six core skills: interpretation, analysis, evaluation, inference, explanation, and self-regulation (Facione, 2015). In this study, five indicators were selected for measurement, while self-regulation was excluded.

The exclusion of the self-regulation indicator was grounded in both theoretical and methodological considerations. Self-regulation is a metacognitive ability that is reflective and intrapersonal in nature; therefore, it is more appropriately measured with non-test

instruments, such as reflective questionnaires or interviews. In contrast, this study used an objective test with a dichotomous scoring system, which is better suited to assessing cognitive skills observable in students' response choices. The distribution of the test items across each critical thinking indicator is presented in Table 1.

Table 1. Domains of critical thinking skills in the instrument

No	Critical Thinking Skills Indicator	Sub-Indicator of Critical Thinking Skills	Item Number
1	Interpretation	a) Categorize	1
		b) Decode significance	2,3
		c) Clarify meaning	4
2	Analysis	a) Examine ideas	5
		b) Identify arguments	6
		c) Identify reasons and claims	7
3	Evaluation	a) Assess credibility of claims	8
		b) Evaluate arguments	9
4	Inference	a) Query evidence	10, 11
		b) Conjecture alternatives	12, 13
		c) Draw logically valid or justified conclusions	14, 15
5	Explanation	a) State results	16, 17
		b) Justify procedures	18, 19
		c) Present arguments	20

To establish clear construct–content alignment, each critical thinking indicator was operationalized within authentic dynamic fluid contexts. The interpretation indicator required students to classify observable flow phenomena, such as laminar and turbulent flow. It also required them to interpret relationships between cross-sectional area and velocity using the continuity principle. The analysis indicator was implemented through items that asked students to examine competing claims, identify valid arguments, and distinguish scientifically sound reasoning from misconceptions in contexts such as Torricelli's theorem, ideal fluid assumptions, and aerodynamic applications.

The evaluation indicator required students to assess the credibility of claims about Bernoulli's principle and fluid-measurement devices. For instance, students determined whether velocity in a venturimeter could be obtained without pressure measurements. The inference indicator involved drawing logical conclusions from quantitative data and experimental discrepancies. This included predicting pressure–velocity changes in varying pipe diameters and explaining deviations between theoretical and measured results. Finally, the explanation indicator required students to select scientifically coherent reasoning in technological applications, including carburetors, spray devices, aerodynamic systems, and medical infusion mechanisms. Through this structured alignment, the instrument measured higher-order reasoning within dynamic fluid content rather than procedural computation alone. The instrument development process was carried out through four main stages, as follows:

1. Construct Formulation
   This stage involved identifying the indicators and sub-indicators of critical thinking based on Facione's theoretical framework and organizing them into a test blueprint aligned with the content of dynamic fluids.

2. Item Development
   At this stage, multiple-choice items were developed based on the established blueprint. Each item was designed to elicit higher-order thinking skills in accordance with the targeted indicator.
3. Content Validation
   Content validation was conducted through expert judgment involving two physics education lecturers. The evaluation covered the alignment of items with theoretical indicators, content accuracy, item construction, distractor quality, and language clarity. Revisions were made based on the experts' feedback until the instrument met the criteria for content validity.
4. Limited Pilot Testing
   A limited pilot test was conducted with one Grade 11 science class of 35 students, whose characteristics were similar to those of the research sample. The pilot data were analyzed using the Rasch Model to evaluate empirical validity and instrument reliability, including item fit statistics, item difficulty levels, item and person reliability, and internal consistency. Items that did not meet the established criteria were revised before being administered in the main study.

To provide a concrete illustration of the developed test items, the following presents an example item representing the interpretation indicator, specifically the sub-indicator of decoding significance, corresponding to item number 3.

This study employed a dichotomous scoring method for the critical thinking abilities exam, awarding a score of 1 for right responses and 0

Observe the following image of a clean water system in a three-story house.

A three-story house uses a rooftop water tank to distribute water. When faucets on different floors are opened at the same time, the water flows much faster on the first floor than on the second and third floors. After checking the system, no pipe leaks are found. However, the water is not distributed evenly, and the upper floors receive slower water flow.
Based on this situation, which solution would be the most effective to ensure a more even water distribution to all floors?

A. Use larger-diameter pipes on the upper floors to increase water pressure according to the continuity principle.
B. Install booster pumps in the pipes leading to the second and third floors to increase water pressure and flow rate.
C. Raise the position of the water tank to increase gravitational pressure so that water flows more evenly to all floors.
D. Install a control valve in the pipe leading to the lower floor to reduce its flow rate and increase pressure to the upper floors.
E. Use smoother pipes with a larger diameter to reduce friction losses and improve water flow to the upper floors.

Figure 1. Example of a critical thinking skills test item on the topic of dynamic fluids

for wrong ones. This scoring approach was implemented to enhance data processing and quantitative analysis, enabling researchers to objectively assess mastery of critical thinking abilities based on the number of correct responses. The total score obtained by each participant reflects both their overall achievement in critical thinking skills and their performance on each specific indicator.

Data Analysis

The students' exam outcomes were evaluated employing the Rasch Model, facilitated by Winsteps software version 3.73. This study encompassed the evaluation of the instrument's validity and reliability, concept validity, item fit to the Rasch model, Differential Item Functioning (DIF), item difficulty, and student ability, all assessed on a unified logit scale. The Rasch Model facilitates the detection of misunderstandings and specific deficiencies in critical-thinking markers that students have yet to acquire. The analytical results underpin the assessment of the instrument's quality and provide a comprehensive overview of students' critical-thinking abilities regarding dynamic fluids.

Item dependability in Rasch analysis denotes the degree to which the consistency of item difficulty levels corresponds with the Rasch model. A high item dependability score indicates that the hierarchy of item difficulty is consistent and trustworthy across various responder groups. The requirements for test instrument dependability in Rasch modeling, as articulated by Sumintono & Widhiarso (Sumintono & Widhiarso, 2015), are delineated in Table 2 below:

Table 2. Interpretation of reliability value

Reliability Value (Person/Item)	Interpretation
$r > 0.94$	Special
$0.91 \leq r \leq 0.94$	Very good
$0.81 \leq r < 0.90$	Good
$0.67 \leq r < 0.80$	Simply
$r < 0.67$	Weak

Rasch modeling was used to evaluate the instrument's validity, specifically for measuring students' cognitive capacities. In Rasch modeling, validity is assessed using Cronbach's alpha, which indicates the instrument's overall reliability and reflects the degree to which test items consistently correlate with student responses. A high Cronbach's Alpha score indicates that the instrument exhibits strong homogeneity and reliably measures the desired construct. Table 3 below delineates the conditions for interpreting Cronbach's Alpha values within the context of Rasch modeling (Sumintono & Widhiarso, 2015):

Table 3. 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

This study also conducted a crucial analysis, including the scalogram, Wright map, and person fit, assessed using indicators such as Outfit MNSQ, ZSTD, and Point Measure Correlations (PT-Measure Corr). Person-fit analysis is used to identify response inconsistencies (misfit) relative to the measurement model (Walker & Wind, 2020). The Scalogram was applied to analyze students' response patterns by item difficulty level, enabling the identification of response consistency and the sequential grouping of students from the highest to the lowest ability levels (Lestari & Samsudin, 2020). Concurrently, the Wright map illustrates the correlation between participants' ability levels (person measure) and item difficulty levels on a singular logit scale (Ciesla & Yao, 2011; Omolade, 2025; Samsudin et al., 2023).

■ RESULT AND DISCUSSION

Quality of the Developed Instrument

Evaluating the overall quality of the developed instrument is essential to ensure that it

functions as a valid and reliable measure of critical thinking skills within the specific context of dynamic fluids. Given the identified research gap in topic-specific critical thinking instruments, a comprehensive psychometric examination was conducted using the Rasch model. This evaluation includes reliability, validity, and item functioning to determine whether the instrument meets acceptable measurement standards.

Reliability Analysis

Reliability is a crucial aspect in instrument evaluation as it reflects the consistency and stability of measurement results. In the context of assessing critical thinking skills, high reliability ensures that the scores obtained accurately represent students' true abilities rather than being influenced by fluctuations or irregularities in the instrument. A reliability analysis was performed to evaluate the instrument's stability from two perspectives: person dependability and item reliability, using the Rasch Model with WINSTEPS software version 3.73. This method facilitates a thorough assessment of item stability and participant capability within a cohesive and impartial measuring system. Table 4 below presents the quantitative statistics from the reliability analysis.

Table 4. Reliability analysis results

	Logit Mean (SD)	Separation	Reliability	Cronbach's Alpha
Person	0.52	2.19	0.83	
	(1.57)			0.89
Item	0.00	3.71	0.93	
	(0.71)			

The results in Table 4 show a person reliability of 0.83 and an item reliability of 0.93, while Cronbach's Alpha reached 0.89, indicating strong internal consistency. Within the Rasch framework, a person's reliability of 0.83 suggests that the instrument provides stable estimates of students' critical thinking abilities, with observed score variation largely reflecting genuine differences rather than measurement error. The person separation value of 2.19 indicates that the instrument can differentiate students into approximately three ability strata (Farzad et al., 2017). This differentiation is pedagogically valuable in heterogeneous classrooms, as it enables the identification of lower, moderate, and higher levels of reasoning performance in dynamic fluid contexts. Similarly, the high item reliability (0.93) and item separation index (3.71) indicate a well-defined hierarchy of item difficulty. This stability suggests that the relative ordering of item difficulty would likely remain consistent across comparable samples. Consequently, the instrument is appropriate for both diagnostic and evaluative assessment purposes within the Rasch measurement framework.

Instrument Validity Analysis

Instrument validity in Rasch modeling refers to the alignment between empirical data and the mathematical model. An instrument is considered valid if participants' response patterns align with model predictions, indicating that each test item genuinely measures the intended construct. This validity can be evaluated using fit statistics, which reflect whether items and respondents behave consistently with Rasch model assumptions. Table 5 presents the indicators used to assess the instrument's validity.

Table 5. Summary of rasch fit statistics for persons and items

Measure	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
Person	1.00	0.1	1.02	0.1
Item	0.99	-0.2	1.02	0.1

According to Table 5, all Infit and Outfit MNSQ values are within the allowed range of 0.5 to 1.5, and the ZSTD values are between -2.0 and +2.0, demonstrating that the data conform to the Rasch model. The average Infit MNSQ value for items is 0.99, and the Outfit MNSQ is 1.02-both values are very close to the ideal of 1.0, suggesting that the items are valid. Thus, the instrument used in this study demonstrates good internal validity for measuring students' critical thinking skills.

Construct Validity

Construct validity is a crucial aspect of instrument validation, ensuring that all items genuinely represent the intended construct. A prevalent approach to assessing construct validity within the Rasch model framework is to evaluate the assumption of unidimensionality, which pertains to the extent to which the items in an instrument reliably measure a single dimension or construct (Bond & Fox, 2015). A unidimensional instrument will yield valid data for assessing responder ability along a singular underlying variable. The assessment of unidimensionality in the Rasch model may be performed by analyzing two primary indicators from the Principal Component Analysis (PCA) of residuals: the raw variance accounted for by measurements and the unexplained variance in the first comparison, as seen in Table 6.

The analysis findings in Table 6 indicate that the observed raw variance is 37.1%, which roughly corresponds to the model-expected value of 36.7%. The variation accounted for by this measure significantly exceeds 20%, indicating the presence of a sufficiently dominant dimension to encapsulate the targeted construct (Linacre, 2016). The unexplained variance in the initial comparison should not exceed 15%, as elevated values may indicate a secondary dimension or extraneous influences (Wright & Stone, 1999). The measured value is 5.6%, which is considered very good and adequate (Fisher, 2007).

Table 6. PCA of residuals summary

Indicator	Empirical (%)	Model (%)
Raw variance explained by measures	37.1	36.7
Unexplained variance in 1st contrast	5.6	-

Item Fit Analysis

Item fit analysis in the Rasch model is essential to ensure that each item in the instrument measures the intended construct validly and consistently. One of the key indicators is Outfit Mean Square (MNSQ), which is used to detect unexpected responses, particularly from participants with extreme ability levels. According to Wright & Linacre, acceptable ranges for both Outfit and Infit MNSQ are between 0.5 and 1.5. Items with MNSQ values within this range are considered to fit the Rasch model. Values above 1.5 indicate that the item may be inconsistent or confusing, while values below 0.5 suggest the item may be too easy or overly predictable (Wright & Linacre, 1994).

In addition to MNSQ, another indicator used is the Z-standardized fit statistic (ZSTD), which indicates the statistical significance of MNSQ deviations. However, if MNSQ values already fall within the acceptable range, ZSTD values may be disregarded due to their high sensitivity in large sample sizes (Linacre, 2016). Thus, the primary interpretation should focus on the MNSQ values. Furthermore, the Point Measure Correlation (PTMEA CORR) is used to assess the direction and strength of the relationship between item scores and respondents' ability estimates. A good PTMEA

CORR value should be positive and above 0.30, indicating that responses to the item align with the level of measured ability (Bond & Fox, 2015; Linacre, 2002).

By considering these three indicators simultaneously, items with Outfit/Infit MNSQ values between 0.5 and 1.5 and a positive PTMEA CORR above 0.30 can be categorized as fitting the Rasch model. Table 7 presents the outcomes of the item fit analysis, including logit values, Standard Error of Measurement (SEM), Outfit MNSQ, Outfit ZSTD, PTMEA CORR, and the fit status of each test item.

Table 7. Identification of test items not fitting based on rasch model parameters

No	Item Code	Logit	Standard Error Measurement	Outfit MNSQ	Outfit ZSTD	Point Measure Correlation	Remark
1	S20	0.25	0.18	1.65	4.0	0.38	Not Fit (MNSQ > 1.50)
2	S14	1.83	0.19	1.44	1.9	0.46	Fit
3	S17	0.16	0.18	1.32	2.1	0.48	Fit
4	S2	-0.98	0.19	1.30	1.2	0.47	Fit
5	S8	0.87	0.18	1.28	1.9	0.53	Fit
6	S18	-0.06	0.18	1.14	0.9	0.52	Fit
7	S13	0.09	0.18	1.06	0.5	0.53	Fit
8	S15	-0.45	0.18	1.13	0.7	0.54	Fit
9	S3	1.03	0.18	1.10	0.7	0.57	Fit
10	S9	0.90	0.18	1.03	0.3	0.62	Fit
11	S19	-0.35	0.18	1.03	0.2	0.56	Fit
12	S12	-0.03	0.18	0.92	-0.5	0.61	Fit
13	S6	0.03	0.18	0.90	-0.6	0.61	Fit
14	S10	-0.03	0.18	0.94	-0.3	0.61	Fit
15	S4	-0.25	0.18	0.83	-1.0	0.61	Fit
16	S1	-1.13	0.19	0.77	-0.9	0.56	Fit
17	S5	-0.13	0.18	0.73	-1.8	0.66	Fit
18	S11	-0.35	0.18	0.62	-2.5	0.67	Fit
19	S16	-0.32	0.18	0.72	-1.8	0.67	Fit
20	S7	-1.09	0.19	0.49	-2.4	0.65	Not Fit (MNSQ < 0.50)

Based on the recommended tolerance limits, Table 7 shows that item S20 has an MNSQ value of 1.65, which exceeds the upper threshold, indicating an inconsistent response pattern or possible response fatigue. Meanwhile, item S7 has an MNSQ value of 0.49 (below the lower threshold), suggesting that the item is overly predictable or fails to provide meaningful new information. Both items are considered misfitting and should be reviewed to maintain the validity of the critical thinking skills instrument.

A deeper content analysis suggests that the underfit observed for item S20 may be attributable to construct-irrelevant variance. The item presents multiple experimental contexts simultaneously (Venturi tube with manometer, U-tube, two sheets of paper, a ping-pong ball and a hair dryer, and flowing water through a narrow opening). While the question explicitly asks for a quantitative verification of Bernoulli's principle, several distractors still represent valid qualitative demonstrations of pressure differences.

Consequently, the cognitive demand may extend beyond the critical evaluation of experimental design to the identification of implicit linguistic cues, such as the term "quantitative".This increases reading load and introduces additional processing demands unrelated to the intended construct (critical thinking in scientific experimentation).

According to Cognitive Load Theory, excessive extraneous cognitive load, originating from suboptimal task design or unnecessary information, can overload working memory capacity and hinder effective cognitive processing (Paas & Van Merriënboer, 2020). When extraneous load exceeds learners' cognitive capacity, task performance may become unstable and less consistent across ability levels (Jong, 2010). In assessment contexts, this may manifest as erratic response patterns that are not attributable to the latent trait being measured, thereby introducing construct-irrelevant variance. In Rasch measurement terms, such noise may appear as underfit, reflected in elevated Outfit MNSQ values. Therefore, revision of item S20 should focus on reducing extraneous contextual complexity, sharpening the operational definition of "quantitative verification," ensuring stronger alignment with a single dominant latent construct.

In contrast, the overfit observed in item S7 appears to stem from limited cognitive complexity and highly transparent distractors. The stem clearly cues aerodynamic design and high-speed motion, strongly activating prior knowledge of Bernoulli's principle. The correct option is substantially more conceptually aligned with fluid dynamics than the distractors, which focus on engine power, tire type, wheel diameter, or suspension stiffness. This imbalance reduces the competitive functioning of distractors and may lower the item's discrimination capacity. The relatively low difficulty estimate (-1.09 logits) supports the interpretation that the item is comparatively easy. The Outfit MNSQ below 0.50 suggests overpredictability, meaning that respondents' answers closely follow the Rasch model expectation with minimal variability. While overfit is less threatening than underfit, it may indicate redundancy or limited contribution to the overall test information function. There is also a possibility of local dependence if similar aerodynamic contexts are present in other items, which could artificially inflate consistency. To enhance psychometric quality, distractors should be reconstructed to include conceptually plausible alternatives within the context of fluid dynamics, thereby increasing the demand for analytical reasoning and improving discrimination.

Overall, the misfit patterns of S20 and S7 arise from different psychometric mechanisms. Item S20 likely reflects construct-irrelevant variance and potential multidimensionality, leading to underfit. Item S7, on the other hand, reflects overpredictability due to low cognitive demand and weak distractor competition. These findings demonstrate the importance of integrating Rasch fit statistics with substantive theoretical analysis to ensure construct validity, unidimensionality, and optimal item functioning in critical thinking assessment.

Item Parameter Analysis (Item Difficulty)

The item parameter analysis in the Rasch model is performed to assess the difficulty level of each test item. The item's difficulty is indicated by the measure value, shown in logit units, as shown in Table 8.

Table 8. Item difficulty estimates (logit measures) based on the rasch model

Entry Number	Measure	Model S.E.
14	1.83	0.19
3	1.03	0.18
9	0.90	0.18

8	0.87	0.18
20	0.25	0.18
17	0.16	0.18
13	0.09	0.18
6	0.03	0.18
10	-0.03	0.18
12	-0.03	0.18
18	-0.06	0.18
5	-0.13	0.18
4	-0.25	0.18
16	-0.32	0.18
11	-0.35	0.18
19	-0.35	0.18
15	-0.45	0.18
2	-0.98	0.19
7	-1.09	0.19
1	-1.13	0.19
Mean	0.00	0.18
S. D.	0.71	0.00

The analysis findings depicted in Table 8 indicate that the item difficulty ranges from -1.13 logits to 1.83 logits. This range falls within the permissible interval of -2 to +2 logits, as shown by Sumintono & Widhiarso (2015), indicating that none of the items are too easy or too challenging. The most challenging item is S14, possessing a logit value of 1.83, and the least difficult item is S1, having a logit value of -1.13. This indicates that item S14 is the most significant obstacle to pupils, whereas item S1 is the most readily answered correctly. The uniform distribution of measurement results within this range demonstrates the instrument's capacity to accommodate varying participant skills, consistent with effective targeting in item response theory.

The measurement accuracy of each item may be assessed using the standard error of measurement (SEM). The results indicate that the SEM for all items ranges from 0.18 to 0.19 logits, which is below the 0.5-logit threshold. According to Bond & Fox, an SEM < 0.5 logits signifies that the item parameter estimates are accurate and stable. Therefore, it can be concluded that the items in this instrument not only exhibit proportional difficulty levels but also possess high measurement precision (Bond & Fox, 2015).

In addition, the difficulty of items can be interpreted in the context of the curriculum and learning objectives in fluid dynamics. Item S14, the most difficult, asks students to compare pressures at multiple points in a pipe with varying cross-sectional areas. To answer this correctly, students must apply Bernoulli's principle, understand the relationship between velocity and pressure, and carefully analyze a diagram. This requires multi-step reasoning and integration of concepts, representing higher-order cognitive skills such as application and analysis according to Bloom's taxonomy. In contrast, item S1, the easiest, asks students to classify flow as laminar or turbulent based on simple observation of flow patterns. This involves recalling definitions and making a straightforward comparison between observed behaviors, requiring lower-order cognitive skills such as understanding and comprehension.

Thus, the Rasch-based difficulty estimates align with the conceptual complexity of the items:

statistically harder items demand deeper reasoning and integration of fluid dynamics principles, whereas easier items focus on basic observation and recall. This demonstrates that the instrument not only measures difficulty quantitatively but also effectively differentiates cognitive complexity across the curriculum content.

Item and Person Characteristics

Beyond establishing psychometric adequacy, it is equally important to examine how the instrument operates in practice by analyzing the interaction between item difficulty and student ability. In Rasch analysis, effective measurement depends not only on reliable items but also on appropriate targeting and consistent response behavior. Therefore, exploring person-fit patterns, response consistency, and the alignment between item difficulty and ability distribution provides deeper insight into how well the instrument captures variation in students' critical thinking skills.

Person Fit Level

In Rasch analysis, the person fit level denotes the degree to which a respondent's response pattern conforms to the measurement model's expected outcomes. The fit is assessed using the Outfit and Infit Mean Square (MNSQ) indices, with optimal values between 0.5 and 1.5 (Wright, B. D. & Linacre, 1994). Z-standardized values (ZSTD) should ideally range from -2.0 to +2.0; however, they may be overlooked in large sample sizes due to their sensitivity (Boone, Staver, & Yale, 2014; Linacre, 2016). The Point-Measure Correlation (PTMEA CORR) serves as another indicator, demonstrating the consistency of participants' responses regarding item difficulty; negative values indicate abnormalities, whilst positive values above 0.3 are considered acceptable (Linacre, 2002). Table 9 presents the Rasch person-fit indices (Outfit MNSQ, ZSTD, and PTMEA CORR) for respondents identified as misfitting according to the established fit criteria.

Table 9. Person fit statistics for misfitting respondents

Person	Outfit MNSQ	Outfit ZSTD	PTMEA CORR
129P	3.30	1.6	-0.32
87P	0.73	-1.1	0.57
167L	1.28	0.9	-0.01
098L	2.25	1.5	-0.21
040P	2.20	2.1	-0.15
196L	1.49	1.3	-0.06
194P	2.10	2.3	-0.35
095P	2.09	1.6	-0.35
189P	2.08	2.3	-0.16
110L	1.99	1.5	-0.31
197L	1.47	1.0	-0.10
159P	1.31	0.6	-0.01
089L	1.27	0.6	-0.02
033L	1.99	1.2	-0.22
170P	1.95	1.0	-0.14
162P	1.88	1.6	-0.35
190L	1.86	3.3	-0.47
023P	1.45	1.0	-0.02
181L	1.45	1.2	-0.15
173L	1.85	1.9	-0.09

108P	1.78	0.9	-0.11
076L	1.64	1.3	-0.08
177L	1.59	0.9	0.03
054P	1.44	2.3	-0.11
009P	1.57	2.0	-0.09
109L	1.24	0.6	-0.01
050P	0.24	-0.6	0.59
123P	0.24	-0.6	0.59
148P	0.24	-0.6	0.59
174L	1.43	2.1	-0.06
002L	1.41	0.8	-0.10
149P	0.24	-0.6	0.59
070L	0.39	-0.8	0.64
048L	0.37	-0.8	0.67
073P	0.37	-0.8	0.67
107L	0.45	-1.0	0.71
053P	0.42	-1.1	0.74
066P	0.42	-1.1	0.74
128P	1.36	1.4	-0.16
005L	1.35	0.7	-0.04
090L	1.31	0.6	-0.01

Table 9 indicates that 41 respondents were identified as misfitting according to the established Rasch criteria. Of these, 27 students (13.5%) did not meet the combined Outfit MNSQ and ZSTD thresholds, while 14 students (7%) failed to satisfy the Outfit MNSQ and PTMEA CORR criteria. These findings suggest response patterns that deviate from model expectations and may undermine the stability of individual ability estimates. As Bond & Fox state, negative or very low item correlations may suggest that high-ability students failed to answer relatively easy items, leading to distorted score interpretations (Bond & Fox, 2015). Therefore, identifying misfitting students is an essential step to ensure measurement quality in Rasch-based research.

To further explore whether misfit responses were associated with particular institutional contexts, the distribution of misfitting respondents was examined across participating schools. Based on the distribution across schools, misfitting respondents were identified in all participating institutions. A proportional examination revealed noticeable variation, with Schools B (32.5%) and C (32.0%) exhibiting substantially higher proportions of misfit students compared to Schools A (13.6%), D (14.3%), and F (13.3%). In comparison, School E showed a moderate proportion (18.2%). Although School B had the highest number of misfit students (n = 13), the distribution across institutions indicates that misfit responses were not restricted to a single context. This pattern suggests that inconsistent response behavior may stem from individual-level factors, such as guessing, inattentiveness, or varying levels of conceptual understanding. It may also reflect contextual influences, such as instructional approaches or assessment conditions.

An examination of the ability distribution among the misfitting respondents indicated that these individuals were dispersed across low, moderate, and high ability levels. Misfit cases were not concentrated in any specific proficiency group, as both higher- and lower-ability students were represented among the identified misfit respondents. This finding suggests that

inconsistent response patterns cannot be attributed solely to ability level but may instead reflect variations in individual response behavior across the ability spectrum.

In the context of physics education, particularly within the topic of dynamic fluids, critical thinking assessment requires students to integrate conceptual reasoning, mathematical relationships, and contextual interpretation. Students may demonstrate procedural competence in applying formulas yet experience conceptual inconsistencies when interpreting fluid phenomena. Such partial or fragmented understanding may result in irregular response patterns, even among students with moderate or high estimated ability levels. Conversely, lower-ability students may exhibit misfit due to guessing or limited comprehension of item demands. Therefore, the presence of misfitting respondents across schools and ability levels may reflect the complex cognitive demands inherent in assessing critical thinking skills in physics rather than merely deficiencies in overall proficiency.

Scalogram

A scalogram visually depicts participants' answer patterns to items in an instrument, correlating item difficulty with respondents' competence. Within the framework of Rasch analysis, the scalogram is used to assess response consistency and evaluate the data's alignment with the Rasch model. A careful examination of the scalogram provides diagnostic insights into student difficulties. Systematic patterns of incorrect responses highlight specific areas in fluid dynamics, such as multi-step conceptual reasoning involving pressure and flow relationships, where students commonly struggle. These insights can guide targeted instructional interventions to strengthen understanding in challenging topics. The rows in Figures 2, 3, 4, and 5 denote persons (respondents), whilst the columns signify the test items. The numeral "1" signifies a valid response, whereas "0" denotes an improper response.

```
TTMAN SCALOGRAM OF RESPONSES:
rson |Item
     |   1111  111 112   1
     |1725196450286370893 4
     |--------------------
  20 +1111111111111011111  020L
  50 +1111111111111111110  050P
  90 +1111111110111111111  090L
  97 +1111111111111110111  097P
  99 +1111111111111111011  099L
```

Figure 2. Consistent pattern

The scalogram analysis in Figure 2 reveals a highly consistent response pattern, with nearly all items answered correctly, indicating a high level of critical thinking ability among these students.

```
  54 +0011010111000110010 1  054P
  55 +0111001001100111001 0  055P
 115 +0110011111011001000 0  115L
 138 +1110001000110101000 1  138L
 158 +1110110101000000100 1  158P
 174 +0101111000000110100 1  174L
 182 +0111100010010110000 1  182P
```

Figure 3. Misfit pattern

Based on Figure 3, there are indications of misfit students-those who were able to answer difficult items correctly but failed to answer easier ones. For example, students 174L and 182P answered the easiest item (item 1) incorrectly but answered the most difficult item (item 14) correctly. This suggests a mismatch between measured ability and item difficulty for these individuals, which may reflect lapses in attention, misunderstanding of specific items, or non-systematic guessing rather than true mastery of the content.

```
 127 +1111111111111110011  127P
 145 +1111111111111101110  145P
 150 +1111111111111101110  150L
 151 +1111111111110101111  151L
 154 +1111111111011101111  154L
```

Figure 4. Identical pattern

Figure 4 shows identical response patterns among several students, such as 145P and 150L. While at first glance this might suggest potential anomalies, it is important to interpret such patterns cautiously. Identical response sequences can result from a range of factors, including similar instructional experiences, coincidental response patterns, or common misconceptions, rather than necessarily indicating cheating. Therefore, these observations should be considered as signals for further review rather than definitive evidence of academic dishonesty.

```
160 +00110000000000010000  160P
198 +00000000001110000000  198P

 33 +00000000000010000100  033L
130 +00000011000000000000  130P
177 +00100000000000001000  177L
```

Figure 5. Random pattern

The response pattern in Figure 5 reveals numerous errors, even on easy items, reflecting very low critical-thinking skills and suggesting that students were simply guessing the correct answers. Overall, the distribution of incorrect responses across items demonstrates the diagnostic value of scalograms: educators can identify which topics in fluid dynamics consistently challenge students, allowing for more focused remediation or reinforcement of key concepts.

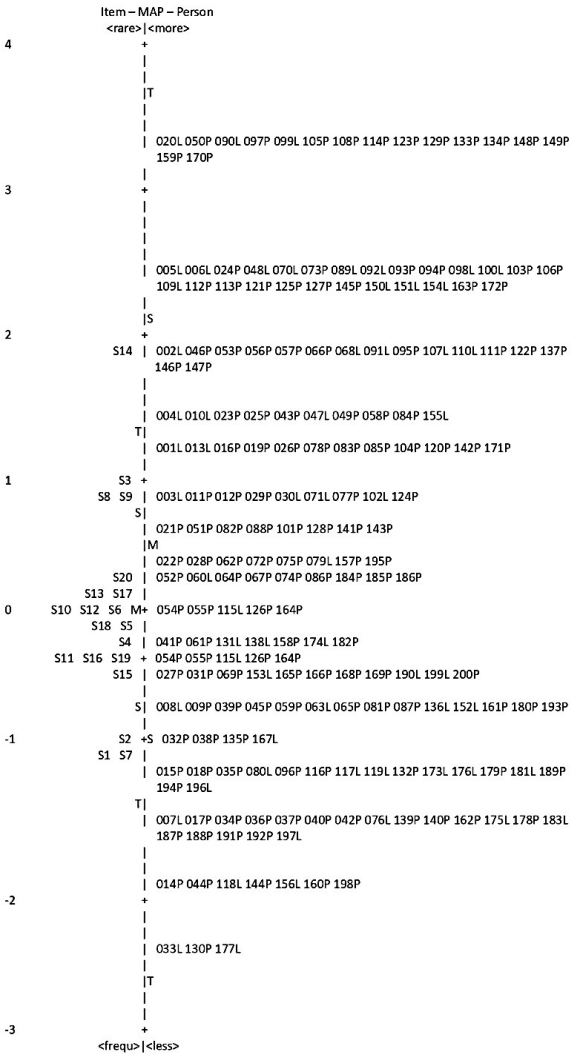

Figure 6. Wright map analysis of student ability and item difficulty

Wright Map

The Wright Map is a fundamental analytical output in Rasch measurement that locates student ability and item difficulty on a common logit scale. Figure 6 presents the Wright Map obtained from the Rasch analysis. The left side of this map illustrates the distribution of student skills, whereas the right side depicts the distribution of item difficulty.

In this study, Wright Map analysis was conducted to evaluate the instrument's targeting, specifically the alignment between students' critical thinking abilities and item difficulty levels within the dynamic fluid domain. The person's ability estimates ranged from "3 to +4 logits, while item difficulty estimates ranged from "1 to +2 logits. The distributions of person ability and item difficulty are summarized in Tables 10 and 11.

The distribution of student abilities shows a fairly wide range, with the highest ability reaching a logit value of +4, as seen in participants 020L, 050P, and 090L. Conversely, students with the lowest ability are positioned at logit -3, such as participants 033L, 130P, and 177L. This indicates

a significant variation in ability levels among test takers. To provide a clearer and more structured presentation of these findings, the distribution of student ability levels is summarized in Table 10.

Table 10. Summary of student ability distribution based on the wright map

Ability Category	Logit Range	Example Participants	Interpretation
Very high ability	+3 to +4	020L, 050P, 090L	Correctly answered most items, including difficult ones
Moderate ability	-1 to +2	majority of students	Ability aligned with most item difficulties
Very low ability	-3 to -2	033L, 130P, 177L	Experienced difficulty even on easier items

The item difficulties are evenly distributed, with the most difficult item, logit +2 item S14, which could only be answered correctly by high-ability students. On the other hand, the easiest items, such as S1 and S7, are located at logit -1 and were answered correctly by nearly all students. This distribution reflects a reasonably good range of item difficulty, although there may be some mismatch at the extremes of student ability. Table 11 presents a summary of the item difficulty levels.

An examination of the targeting between person ability and item difficulty indicates that most students are clustered within the moderate ability

Table 11. Summary of item difficulty distribution

Difficulty Category	Logit Value	Item Code	Interpretation
Most Difficult	+2	S14	Answered correctly only by high-ability students
Moderate	0 to +1	majority of items	Suitable for average student ability
Easiest	-1	S1, S7	Answered correctly by nearly all students

range, whereas items are concentrated between "1 and +2 logits. This suggests that the instrument is generally well targeted to students with average critical-thinking ability. However, the greater spread in person ability relative to item difficulty indicates limited measurement precision at the extreme ends of the ability continuum. The existing items may not sufficiently challenge students at the very high end (+3 to +4 logits). In contrast, students at the very low end ("3 logits) may require additional, easier items to improve measurement sensitivity. Therefore, it is recommended to include items at both higher and lower difficulty levels to broaden the measurement continuum and improve targeting accuracy. Expanding the difficulty range would enhance the instrument's capacity to capture subtle differences in critical thinking performance across diverse ability levels, thereby strengthening its overall measurement precision.

Gender Fairness Analysis

Fairness in educational assessment is critical to ensuring that measurement outcomes are not systematically biased toward particular groups. In the context of critical thinking assessment, gender-related bias may compromise interpretability and equity. Therefore, Differential Item Functioning (DIF) analysis was conducted to determine whether the instrument operates equivalently across male and female students.

Differential Item Functioning (DIF)

In Rasch Model analysis, Differential Item Functioning (DIF) is a critical component used to determine whether an item operates differently across respondent groups (e.g., by gender, grade level, or background). In this context, comparisons were conducted between male and female groups, with Differential Item Functioning (DIF) assessed using probability (PROB) values derived from the chi-square test, as shown in Table 12.

Table 12. DIF results based on chi-square probability values

Item	PROB.
S1	0.0028
S2	0.6223
S3	0.0030
S4	1.0000
S5	0.3743
S6	0.7919
S7	0.6538
S8	0.0954
S9	0.8821
S10	0.6632
S11	0.2834
S12	0.1092
S13	0.9378
S14	0.1282
S15	0.8195
S16	0.1766
S17	0.0057
S18	0.4733
S19	0.9237
S20	0.3565

Bond & Fox assert that an item exhibits no significant differential item functioning (DIF) if the p-value exceeds 0.05, indicating statistical non-significance (Bond & Fox, 2015). Based on the analysis results, out of the 20 items tested, three items exhibited significant DIF indications: S1 (PROB = 0.0028), S3 (PROB = 0.0030), and S17 (PROB = 0.0057). These items have probability values below the threshold, indicating differences in response patterns between male and female students on these items. Such differences may arise from factors such as item context, the use of specific terminology, or cultural interpretations that may advantage one group over the other. Conversely, the other items have PROB values above 0.05, indicating no substantial disparity in responses between the two respondent groups. The statistical analysis is corroborated by the Person DIF Plot, which graphically depicts performance disparities between groups using the DIF Measure (diff.) for each item, as shown in Figure 7.

The DIF curve depicted above shows the male group (L) in blue, the female group (P) in red, and the combined average of both responder groups in green. The graph illustrates a significant disparity between the lines representing male and female groups for items S1, S3, and S17, suggesting that these items assess critical thinking abilities differently across the categories. These items tend to pose unequal challenges for the two groups, indicating significant Differential Item Functioning (DIF). This conclusion corresponds to the prior statistical study, which indicated that these three items have p-values within the significance level. In contrast, for items such as S2, S4, and S9, the patterns of both group lines appear close or overlapping, reflecting that these items function consistently and fairly in measuring critical thinking skills without favoring either group.

To interpret these differences substantively, a qualitative review of the three flagged items was conducted, focusing on cognitive demands, abstraction level, and potential triggers for misconceptions. Item S1 requires students to classify the flow of faucet water as laminar or turbulent based on observable characteristics. Although the context is everyday and neutral, the task demands translating visual phenomena into formal fluid-dynamics terminology. Distractors combine partially correct intuitive reasoning with

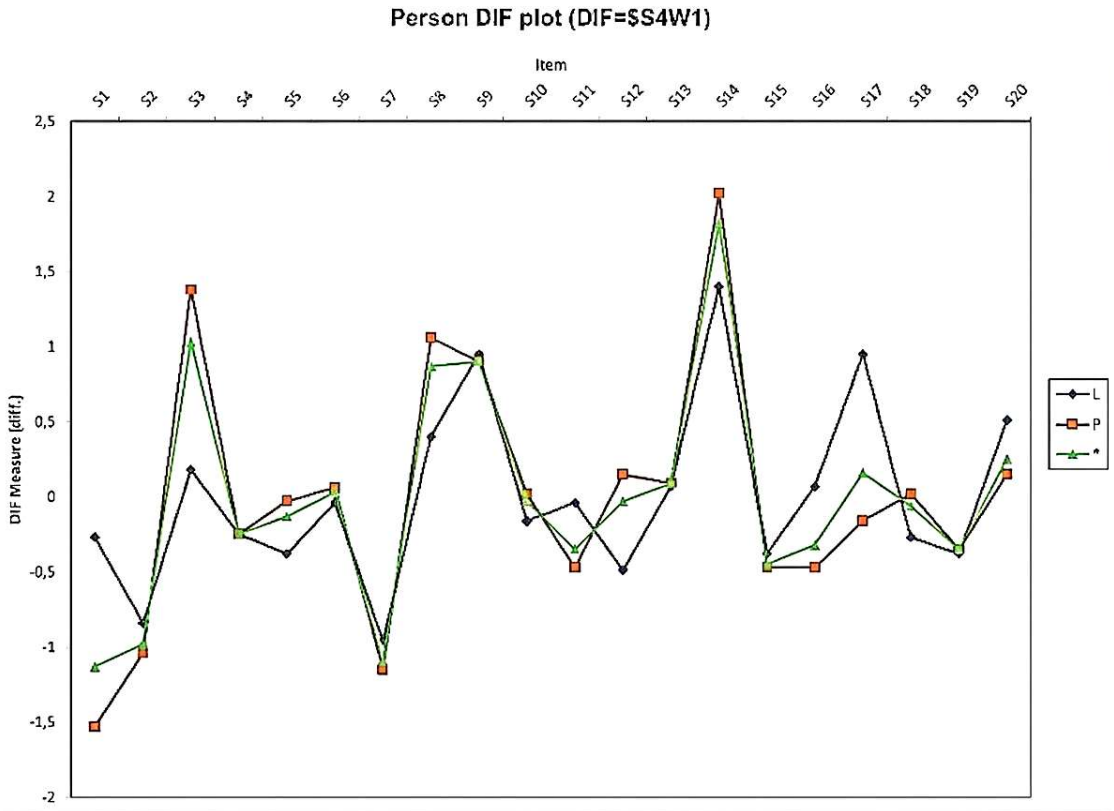

Figure 7. DIF plot

conceptual inaccuracies. For example, Option A states that faster flow must be laminar because it appears more forceful. In contrast, Option C correctly explains that increasing velocity can cause the flow to transition from laminar to turbulent. This contrast requires students to move beyond perceptual judgment and apply theoretical classification criteria rather than relying on surface appearance. Students who depend primarily on intuitive visual cues may therefore be more likely to select the incorrect option. Thus, the DIF in S1 likely reflects differences in abstraction and conceptual linkage rather than contextual bias.

Item S3 involves a multi-step systems problem concerning water distribution in a three-story building. Students must integrate hydrostatic pressure, gravitational height differences, and solution evaluation. Several distractors employ correct terminology but flawed reasoning, potentially activating common misconceptions about pressure, diameter, or continuity. For instance, Option A proposes increasing the pipe diameter on the upper floor based on the continuity principle, and Option E suggests reducing friction losses by using smoother materials and larger-diameter pipes. Although both options use technically accurate terminology, they do not directly resolve the vertical pressure imbalance caused by gravitational height differences. In contrast, Option D proposes installing a flow-control valve on the pipeline serving the lower floors to reduce the flow rate and increase the pressure supplied to the upper floors. This option directly targets the central issue, namely the vertical pressure imbalance resulting from gravitational height differences. It demonstrates a systemic understanding of how flow regulation can influence pressure distribution across different levels of the building. Students

must therefore distinguish between superficially plausible engineering modifications and conceptually relevant pressure-augmentation mechanisms, thereby increasing cognitive load and requiring integrated system-level reasoning. The item, therefore, differentiates students based on their ability to construct a coherent physical model rather than to apply isolated formulas. The observed DIF may stem from variation in strategic reasoning approaches and in the complexity of systems thinking.

Item S17 assesses understanding of Bernoulli's principle in a pesticide sprayer system. The task requires reasoning about invisible pressure changes due to increased velocity. Distractors explicitly reflect well-documented misconceptions, such as assuming that higher velocity increases pressure or that gravity pulls the liquid upward. Specifically, Option C states that increased air velocity raises pressure inside the nozzle, directly contradicting Bernoulli's principle, while Option E attributes the upward motion of liquid to gravitational force. In contrast, Option A correctly explains that higher air velocity lowers pressure, creating a pressure differential that causes the liquid to rise. This conceptual reversal between intuitive belief and formal theory represents the core reasoning challenge of the item. The pronounced DIF peak suggests that this item strongly differentiates students based on the depth of their conceptual understanding and their resistance to intuitive reasoning errors.

Overall, the DIF observed in S1, S3, and S17 appears to be associated with differences in abstraction demands, multi-step reasoning complexity, and susceptibility to misconceptions in fluid mechanics rather than with overt gender-based content bias. While the contexts are functionally neutral, variations in cognitive processing strategies or instructional emphasis may contribute to differential response patterns. These findings highlight the importance of refining distractor clarity and ensuring that items measure critical thinking in fluid dynamics without unintended construct-irrelevant variance.

CONCLUSION

This study evaluated the efficacy of a critical-thinking skills instrument for dynamic fluids using the Rasch Model. The results demonstrate that the instrument exhibits good validity and reliability, as evidenced by a Cronbach's Alpha of 0.89 and an item reliability of 0.93, indicating robust internal consistency and measurement precision. The Rasch analysis revealed a proportional distribution of item difficulties (ranging from "1.13 to 1.83 logits), which accommodates variations in students' abilities while also identifying specific challenges, such as item S14 being the most difficult. DIF analysis showed that most items function fairly across genders, except for items S1, S3, and S17, which require further review. This suggests that the instrument is generally equitable, although potential bias exists in some items. The Wright Map highlighted the need to add items that can assess very high or very low ability levels, guiding future instrument development. This study validates the efficacy of the Rasch Model in assessing critical thinking abilities. It provides a reliable instrument for educators to evaluate and enhance students' higher-order thinking in dynamic fluid topics. Items identified as highly difficult highlight content areas that require greater instructional emphasis. The item-level results directly inform which concepts should be reinforced, clarified, or improved in classroom instruction.

DECLARATION OF GENERATIVE AI USAGE IN THE WRITING PROCESS

Statement: During the writing of this manuscript, the authors employed ChatGPT (OpenAI) to assist with language refinement. The authors have reviewed and edited the content generated by this tool and assume full

responsibility for the content of the published article.

REFERENCES

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., & Wade, C. A. (2015). Strategies for teaching students to think critically/ : a meta-analysis. *Review of Educational Research*, *85*(2), 275–314. https://doi.org/10.3102/0034654314551063

Akhvlediani, M., Abdaladze, L., & Lataria, K. (2023). The Challenges of the XXI century–development of critical thinking among the students. *KREBSI*, *15*, 275–285.

Alyaumusyifa, N., Rahma, I., & Kaniawati, I. (2024). Characterization of An Instrument Test to Measure Students ' Creative Thinking Skills (CTS) of Static Fluid Topic Based on Rasch Model Analysis. *Sustainability Education*, *1*(1), 159–171.

Arifin, Z., Sukarmin, Saputro, S., & Kamari, A. (2025). The effect of inquiry-based learning on students ' critical thinking skills in science education/ : A systematic review and meta-analysis. *EURASIA Journal of Mathematics, Science and Technology Education*, *21*(3). https://doi.org/https://doi.org/10.29333/ejmste/15988

Arslan, S. (2010). Traditional instruction of differential equations and conceptual learning. *Teaching Mathematics and Its Applications*, *29*, 94–107. https://doi.org/10.1093/teamat/hrq001

Badmus, O. T., & Jita, L. C. (2024). Physics difficulty and problem-solving: Exploring the role of mathematics and mathematical symbols. *Interdisciplinary Journal of Education Research*, *6*, 1–14. https://doi.org/10.38140/ijer-2024.vol6.08

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences, 3rd ed.* Routledge/Taylor & Francis Group. https://psycnet.apa.org/record/2015-33472-000

Bondarev, V. N., Bezverkhii, P. P., & Kosenko, S. I. (2013). Analysis of experimental data within the statistical theory of critical phenomena. *Russian Journal of Physical Chemistry A*, *87*(11), 1838–1844.

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

Chen, X. (2021). Quantitative descriptive epidemiology. In *Quantitative Epidemiology*. Springer. https://doi.org/10.1007/978-3-030-83852-2_3

Christensen, W. M., & Thompson, J. R. (2010). Investigating student understanding of physics concepts and the underlying calculus concepts in thermodynamics. *Proceedings of the 13th Annual Conference on Research in Undergraduate Mathematics Education March 2010*, 1–24. http://adsabs.harvard.edu/abs/2010APS..MARH42004T

Ciesla, J. R., & Yao, P. (2011). Validation of a targeted peer relations scale for adolescents treated for substance use disorder/ : an application of rasch modeling. *Substance Abuse/ : Research and Treatment*, 35–44. https://doi.org/10.4137/SART.S7367

Cottrell, S. (2017). Critical thinking skills/ : effective analysis, argument, and reflection. *Bloomsbury Publishing*, *100*.

Dwyer, C. P., Hogan, M. J., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. *Thinking Skills and Creativity*, *12*, 43–52. https://doi.org/https://doi.org/10.1016/j.tsc.2013.12.004

Etkina, E., & Planinšiė, G. (2015). Defining and developing "critical thinking" through devising and testing multiple explanations

of the same phenomenon. *The Physics Teacher*, *53*(7), 432–437. https://doi.org/10.1119/1.4931014

Facione, P. A. (2015). Permission to reprint for non-commercial uses critical thinking: what it is and why it counts. *Insight Assessment*, *5*(1), 1–30. www.insightassessment.com

Farzad, M., Layeghi, F., Hosseini, A., Whiteneck, G., & Asgari, A. (2017). Using the rasch model to develop a measure of participation capturing the full range of participation characteristics for the patients with hand injuries. *Journal of Hand and Microsurgery*, *09*(02), 084–091. https://doi.org/10.1055/s-0037-1604060

Fisher, W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, 1095.

García-Carmona, A. (2023). Scientific thinking and critical thinking in science education: two distinct but symbiotically related intellectual processes. *Science and Education*, *34*(1), 227–245. https://doi.org/10.1007/s11191-023-00460-5

Halimatun Sa, L., Siahaan, P., Suhendi, E., Samsudin, A., Hadiana Aminudin, A., Rais, A., Sari, I., & Rachmadtullah, R. (2020). Critical thinking instrument test (ctit): developing and analyzing sundanese students' critical thinking skills on physics concepts using rasch analysis. *International Journal of Psychosocial Rehabilitation*, *24*(January), 2020. https://doi.org/10.37200/IJPR/V24I8/PR281423

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development: An NCME instructional module. *Educational Measurement: Issues and Practice*, *12*(3), 253–262.

Hurrell, D. P. (2021). Conceptual knowledge OR Procedural knowledge OR Conceptual knowledge AND Procedural knowledge/ : Why the conjunction is important for teachers. Conceptual Knowledge, Procedural Knowledge, or Conjunction is Important to Teachers. *Australian Journal of Teacher Education*, *46*(2).

Jong, T. De. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science*, 105–134. https://doi.org/10.1007/s11251-009-9110-0

Kassiavera, S., Suparmi, A., & Cari, C. (2024). ISSN 1648-3898 ISSN 2538-7138 Application of rasch model in two-tier test for assessing critical thinking in physics education. *Journal of Baltic Science Education*, *23*(6), 1227–1242. https://doi.org/https://doi.org/10.33225/jbse/24.23.1227

Krange, I., & Ludvigsen, S. (2008). What does it mean/ ? Students ' procedural and conceptual problem solving in a CSCL environment designed within the field of science education. *Computer-Supported Collaborative Learning*, 25–51. https://doi.org/10.1007/s11412-007-9030-4

Lestari, A. S., & Samsudin, A. (2020). Using rasch model analysis to analyze students' scientific literacy on heat and temperature. *Proceedings of the 7th Mathematics, Science, and Computer Science Education International Seminar, MSCEIS 2019*, 1–8. https://doi.org/10.4108/eai.12-10-2019.2296483

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 878.

Linacre, J. M. (2016). *A User's Guide to WINSTEPS: Rasch Model Computer Programs*. Mesa Press.

Liu, Z., Guo, H., Zhou, Z., Ma, F., & Zeng, Y. (2025). How creative self-efficacy influences problem- solving skills in

engineering education/ : the dual mediating role of critical thinking and metacognition. *BMC Psychology*. https://doi.org/https://doi.org/10.1186/s40359-025-03630-y

Morris, B. J., Croker, S., Zimmerman, C., Gill, D., & Romig, C. (2013). Gaming science/ : the " gamification " of scientific thinking. science education and its role in the twenty-first century. *Frontiers in Psychology*, *4*(September), 1–16. https://doi.org/10.3389/fpsyg.2013.00607

Nunez, W. A., Scott, M. C. B., & Morgan, D. E. (2025). struggle with higher-order cognitive skills in multiple course formats. *Journal of Microbiology & Biology Education*, *26*(2).

Nurdini, N., Suhandi, A., Ramalis, T., & Samsudin, A. (2020). Developing multitier instrument of fluids concepts (mifo) to measure student ' s conception/ : a rasch analysis approach. *Journal of Advanced Research in Dynamical and Control Systems ·*, *October*. https://doi.org/10.5373/JARDCS/V12I6/S20201273

Nwabuko, O. C., Iwu, L. O., Njoku, P. U., & Nwamoh, U. N. (2024). An overview of research study designs in quantitative research methodology. *American Journal of Medical and Clinical Research & Reviews*, *3*(5), 1–6. https://doi.org/10.58372/2835-6276.1169

Omolade, O. K. (2025). Wright map analysis to determine nurses and midwives ' knowledge of treatment of primary postpartum haemorrhage in Nigeria. *International Medical Education*, *4*(6), 1–13. https://doi.org/https://doi.org/10.3390/ime4020006

Paas, F., & Van Merriënboer, J. J. G. (2020). Cognitive-Load theory/ : methods to manage working memory load in the learning of complex tasks. *Current Directions in Psychological Science*, *29*(4), 394–398. https://doi.org/10.1177/0963721420922183

Pasquinelli, E., & Richard, O. (2023). Critical thinking as the ability to sort and qualify the information available, to form one's own judgement. *European Journal of Education*, *58*(3), 422–433. https://doi.org/https://doi.org/10.1111/ejed.12565

Price, O., & Lovell, K. (2018). Quantitative research design. In *A research handbook for patient and public involvement researchers* (pp. 40–50). Manchester University Press.

Razak, A. A. (2022). Improving critical thinking skills in teaching through problem-based learning for students/ : a scoping review. *International Journal of Learning, Teaching and Educational Research*, *21*(2), 342–362.

Rivas, S. F., Saiz, C., & Ossa, C. (2022). Metacognitive strategies and development of critical thinking in higher education. *Frontiers in Psychology*, *13*, 913219.

Rusmin, L., Misrahayu, Y., Pongpalilu, F., Radiansyah, & Dwiyanto. (2024). Social critical thinking and problem-solving skills in the 21st century open access. *Join: Journal of Social Science*, *1*(5), 144–162.

Samsudin, A., Aminudin, A. H., Novia, H., Suhandi, A., Fratiwi, N. J., Yusup, M., Supriyatman, S., Masrifah, M., Adimayuda, R., Prahani, B. K., Wibowo, F. C., Faizin, M. N., & Costu, B. (2023). Identifying javanese students' conceptions on fluid pressure with wright map analysis of rasch. *Journal of Natural Science and Integration*, *6*(2), 173. https://doi.org/10.24014/jnsi.v6i2.21822

Sanjaya, A. P., Tanjung, Y. I., & Mihardi, S. (2024). Development of physics test instruments to measure problem solving

skills on dynamic fluid materials. *Journal of Innovative Physics Teaching*, *2*(2), 89–101.

Santos, L. F. (2017). The role of critical thinking in science education. *Journal of Education and Practice*, *8*(20), 159–173.

Slater, P., & Hasson, F. (2024). Quantitative research designs, hierarchy of evidence and validity. *Journal of Psychiatric and Mental Health Nursing*, 656–660. https://doi.org/10.1111/jpm.13135

Stéphan Vincent-Lancrin. (2024). Critical thinking. In *Elgar Encyclopedia of Interdisciplinarity and Transdisciplinarity* (pp. 124–128). https://doi.org/10.4337/9781035317967.ch27

Sujatmika, S., Sutarno, Masykuri, M., & Prayitno, B. A. (2025). Applying the Rasch model to measure students' critical thinking skills on the science topic of the human circulatory system. *Eurasia Journal of Mathematics, Science and Technology Education*, *21*(4). https://doi.org/10.29333/ejmste/16221

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

Sya'Bandari, Y., Firman, H., & Rusyati, L. (2018). The validation of science virtual test to assess 7th grade students' critical thinking on matter and heat topic (SVT-MH). *Journal of Science Learning*, *1013*(1). https://doi.org/10.1088/1742-6596/1013/1/012067

Thornhill-Miller, B., Camarda, A., Mercier, M., Burkhardt, J.-M., Morisseau, T., Bourgeois-Bougrine, S., & Lubart, T. (2023). Creativity, critical thinking, communication, and collaboration: Assessment, certification, and promotion of 21st century skills for the future of work and education. *Journal of Intelligence*, *11*(3), 54.

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*(4), 663–682. https://doi.org/10.1007/s10763-016-9723-0

Walker, A. A., & Wind, S. A. (2020). Identifying misfitting achievement estimates in performance assessments: an illustration using rasch and mokken scale analyses. *International Journal of Testing*, *20*(3), 231–251. https://doi.org/10.1080/15305058.2019.1673758

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 370–371.

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

Wright, B. D. (1999). Rasch measurement model. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment*. Pergamon (Elsevier Science).

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Mesa Press.

Young, H. D., & Freedman, R. A. (2012). *Sears and Zemansky's University Physics with Modern Physics* (13th ed.). Addison-Wesley.

Žakelj, A., & Štemberger, T. (2025). An empirical study on basic and conceptual knowledge, procedural knowledge and problem solving among primary school students. *International Journal of Instruction*, *18*(4), 627–650.

Zuhaida, A., Zuhri, M. K., & Yusuf Al Ayyubi, S. H. (2022). Analysis of students' critical thinking skills through science, technology, engineering and mathematics (STEM) approach. *Nucleation and Atmospheric Aerosols*, *2600*(1). https://doi.org/10.1063/5.0112996