e-ISSN: 2355-620X
DOI: 10.12928/jrkpf.v13i1.1948

Static fluid concept understanding test instrument: Development and validation through Rasch analysis

Delilah Nur Maisyaroh, Sutopo*, and Parno

Physics Department, University of Malang, Malang, Indonesia

Email: sutopo.fisika@um.ac.id

Abstract
Students often experience persistent conceptual difficulties in static fluid topics due to fragmented understanding and reliance on intuitive reasoning. Therefore, the development of valid and reliable diagnostic instruments is essential to accurately assess students' conceptual understanding. This study aims to develop and validate a Static Fluid Concept Understanding Test covering hydrostatics, buoyancy, Pascal's law, and surface tension. The research employed an instrument development and validation design using Rasch modeling to evaluate item functioning, dimensionality, and measurement quality. Data were collected from 54 secondary students and supported by expert validation. The results indicate high item reliability (0.93) with stable item calibration, although person reliability is relatively low, likely due to limited variability in student ability. Several items exhibit near-misfit patterns, reflecting variation in students' reasoning in conceptually demanding contexts. Dimensionality analysis suggests essential unidimensionality, while Differential Item Functioning (DIF) analysis shows no clear systematic bias across gender. In conclusion, the instrument demonstrates strong item-level measurement quality and potential as a diagnostic tool. In practice, it can help physics teachers identify students' misconceptions and design targeted instructional strategies. However, further validation with larger, more diverse samples is recommended to enhance its precision and generalizability.

Keywords: Conceptual assessment, Naïve understanding, Rasch model, Students' reasoning

Article submitted 2025-12-27. Revision uploaded 2026-04-06.
Accepted for publication 2026-04-13.
Available online on 2026-04-30.
https://doi.org/10.12928/jrkpf.v13i1.1948

I. Introduction

In contemporary physics education, conceptual understanding stands as the most consequential marker of substantive learning, shaping how instructional effectiveness is interpreted and evaluated [1], [2]. Yet, the coherent integration and application of scientific conceptual structures to interpret, explain, and predict physical phenomena is persistently difficult for many learners [3]. In the context of static fluids, such understanding is not merely a collection of isolated ideas but rather a coherent structure of interrelated principles [4]. Core concepts such as hydrostatic pressure, Pascal's law, and Archimedes' principle are inherently interconnected. Specifically, understanding pressure as a function of depth and fluid density underpins reasoning about pressure transmission in hydraulic systems, which in turn relates to the interpretation of buoyant force as a consequence of fluid displacement [5]–[8]. Consequently, meaningful

* *Corresponding author*

learning in this domain requires not only conceptual accuracy but also the integration of these principles into a coherent cognitive framework.

Despite this structural nature of knowledge, extensive research in physics education demonstrates that students' conceptual structures are often fragmented, unstable, and dynamically constructed rather than scientifically coherent [3]. Learners tend to rely on intuitive reasoning derived from everyday experiences, which leads to persistent and systematic misconceptions [2], [3], [9]. Current research converges on three particularly persistent naïve understandings in static fluids: many students (i) attribute hydrostatic pressure to container shape, cross-sectional area, or total fluid volume; (ii) misinterpret Pascal's law by conflating pressure with force, assuming that transmitted pressure or force scales directly with piston area or that forces at different pistons must share the same direction; and (iii) treat buoyant force as determined primarily by object-related properties instead of by the weight of the displaced fluid, leading to systematic errors in predicting floating and sinking [10]–[12]. These patterns of reasoning are not random errors but reflect systematic deviations from the underlying conceptual structure of static fluids, indicating the presence of stable naïve understandings that guide student responses. Such findings highlight the need for assessment approaches that can capture not only correctness but also the quality and coherence of students' reasoning.

Various assessment instruments have been developed to investigate students' understanding of static fluid concepts using different methodological approaches. Traditional assessments based on Classical Test Theory (CTT) primarily rely on total scores and item difficulty indices, which are inherently dependent on sample characteristics and therefore limited in their ability to provide objective measurement [13]–[16]. To address these limitations, more recent developments include diagnostic instruments such as the Static Fluid Concept Inventory (SFCI) [17] and multitier assessments (e.g., MIFO and SFTTI) designed to capture students' reasoning processes and alternative conceptions more explicitly [18], [19]. These instruments have made important contributions to the field by providing richer insights into students' conceptual difficulties, particularly in identifying patterns of misconceptions and reasoning strategies.

At the same time, each approach reflects different analytical emphases. For example, the Static Fluid Concept Inventory (SFCI) focuses on classifying students' conceptions through a structured multitier format, while instruments such as MIFO and SFTTI extend this approach by incorporating additional reasoning layers to capture students' thought processes in greater detail [17]–[19]. These developments have significantly improved the diagnostic capacity of assessments compared to traditional methods. However, the primary focus of many of these instruments remains on identifying misconceptions or comparing group performance outcomes. As a result, the extent to which they capture the internal organization and coherence of conceptual understanding, particularly how students coordinate multiple concepts within a unified structure, remains an area for further exploration.

In parallel, Rasch modeling has been increasingly adopted in physics education research as a means of achieving more objective measurement by calibrating item difficulty and student ability on a common interval scale [14], [16]. Rasch analysis provides a range of analytical tools, including item fit statistics, Wright mapping, and dimensionality analysis, which enable researchers to examine both measurement properties and response patterns in greater detail [20]. These features offer the potential to move beyond score-based evaluation toward a more nuanced understanding of how students interact with assessment items and how their responses reflect underlying conceptual structures [14]–[16], [21]. Despite this potential, the application of Rasch modeling in the context of static fluid assessment has often focused on establishing measurement quality, such as reliability and item functioning, or on conducting group-level comparisons [22], [23]. While these contributions are valuable, there remains an opportunity to explore further how Rasch-based indicators, particularly model-data misfit and response variability, can be interpreted as indicators of unexpected response patterns that may reflect inconsistencies in students' conceptual understanding [15], [24], [25].

From this perspective, the present study does not seek to replace existing diagnostic instruments, but rather to extend prior work by exploring the use of Rasch-based analysis as a complementary approach for examining conceptual understanding. Specifically, this study develops a Static Fluid Concept Understanding Test comprising 15 items designed to represent key conceptual relationships among hydrostatic pressure, Pascal's law, Archimedes' principle, and surface tension. Accordingly, this study aims to develop and validate a diagnostic instrument for assessing students' conceptual understanding of static fluid concepts using Rasch modeling. The focus is placed on examining item quality, exploring the hierarchical distribution of item difficulty, and interpreting response patterns that may reflect variations in students' conceptual understanding. Through this approach, the study contributes an alternative perspective to existing research by highlighting how Rasch-based analysis can be used not only for measurement purposes but also for gaining insight into the

structure of students' reasoning. Therefore, there is a need to develop and validate robust diagnostic instruments using modern measurement approaches to assess students' understanding of static fluid concepts accurately. This study addresses this gap by developing and validating the Static Fluid Concept Understanding Test based on Rasch modeling, thereby contributing to the development of more valid and reliable assessments in physics education.

II. Methods

This research employed an instrument development and validation design anchored in psychometric principles and Rasch Measurement Theory. Instrument development followed established standards for construct-aligned assessment design, content validation, and empirical calibration [1], [15], [16], [26], while validation procedures were guided by methodological criteria for measurement stability, diagnostic sensitivity, and structural coherence [15]. Within this framework, Rasch analysis was adopted to evaluate item functioning, model–data fit, dimensionality, and instrument fairness, reflecting contemporary recommendations for test development in physics education and conceptual diagnostics [1], [16], [27].

The instrument development process involved adapting and modifying items from previous studies [28]–[35] and aligning them with the targeted conceptual indicators. An initial pool of 20 items was constructed and subjected to expert validation to ensure content relevance, conceptual accuracy, and linguistic clarity. The validation process used a Likert scale from 0 to 3, and the instrument yielded an average score of 2.7 out of 3, indicating that the items were highly appropriate for measuring students' conceptual understanding [36]. The validated items were then tested in a pilot study to evaluate their empirical performance. Based on the initial reliability analysis, four items were removed due to inadequate measurement consistency. A subsequent validation stage was conducted with a different group of students, resulting in improved reliability. The instrument demonstrated strong internal consistency, with a KR-20 coefficient of 0.91, indicating high reliability for measurement purposes. From the remaining 16 items, one item exhibiting extreme misfit was excluded, yielding a final set of 15 items for Rasch calibration.

The final instrument structure and item distribution are presented in Table 1, which links static fluid subconcepts to conceptual indicators. Although surface tension is ultimately represented by a single item in the final instrument, this outcome resulted from the item refinement process, during which other items within this subtopic were excluded based on empirical evaluation. This allocation, therefore, reflects both the refinement procedure and the relatively limited emphasis of surface tension within the targeted curriculum scope. Future instrument refinement may consider expanding item representation for this subconstruct.

Table 1. Test blueprint for the static fluid concept understanding instrument

Static Fluid Subconcept	Conceptual Indicator	Sample Item Focus	Item Count
Hydrostatic Pressure	Understands the dependence of hydrostatic pressure on depth and fluid density; compares pressure at different positions and conditions.	Comparing pressure at equal depths in different container shapes, pressure variation with depth, and fluid type.	5
Pascal's Principle	Applies pressure transmission principles in hydraulic systems; analyzes the relation between force, area, and pressure.	Predicting force and pressure outcomes in piston systems with unequal cross-sectional area.	4
Archimedes' Principle	Explains buoyant force as the weight of displaced fluid; predicts floating/sinking behavior based on density differences.	Determining buoyant force and object behavior in fluids; relating volume displacement to buoyancy.	5
Surface Tension	Recognizes intermolecular forces at fluid surfaces and their effects in phenomena such as capillarity and floating objects.	Explaining capillary rise and surface tension effects in fluid interfaces.	1
Total			15

Data were collected using the finalized 15-item diagnostic instrument administered to 54 eleventh-grade science students from a senior high school in Malang. In addition to selecting answers, students were required to provide written explanations for their choices. These responses were used to support the interpretation of response patterns and provide additional context for students' reasoning. Participants were selected using purposive sampling, with the criterion that they had completed instruction on static fluid concepts. Permission for data collection was obtained from the school through formal administrative procedures in accordance with

institutional ethical guidelines, and participants were informed of the study's purpose prior to participation. Participation was voluntary and conducted with the awareness and approval of both the school and the students.

Following data collection, the adequacy of the sample size was considered in relation to Rasch measurement requirements [37]. Rasch measurement allows for instrument calibration with relatively small samples while maintaining estimation stability. According to Rasch measurement guidelines, a sample size of approximately 50 participants can yield stable item estimates within ±1 logit at a 99% confidence level [37], [38]. Therefore, the sample size used in this study meets the adequacy criteria for Rasch calibration. While larger sample sizes generally improve the stability and precision of Rasch estimates, the sample size used in this study is sufficient for preliminary calibration.

The collected data were analyzed using the Rasch model to evaluate the instrument's psychometric quality. Rasch analysis was conducted using Ministep version 5.10.4, a Rasch modeling software developed based on the Winsteps framework [39], [40]. Item responses were analyzed using a dichotomous scoring approach. The analysis included: (a) reliability and separation indices; (b) item fit statistics (INFIT and OUTFIT), evaluated using Mean Square (MNSQ) and standardized Z values (ZSTD), with acceptable criteria of 0.5–1.5 for MNSQ and −2.0 to +2.0 for ZSTD [20], [37]; (c) Principal Component Analysis of Residuals (PCAR) to assess dimensionality; (d) Wright mapping to examine the alignment between item difficulty and student ability; and (e) Differential Item Functioning (DIF) to evaluate item fairness across subgroups. However, DIF results were interpreted cautiously due to the limited subgroup sample sizes in line with established Rasch measurement guidelines.

III. Results and discussion

The results of the Rasch analysis are presented to evaluate the psychometric quality and diagnostic capability of the developed instrument. The analysis focuses on reliability and separation indices, item fit statistics, dimensionality testing, and hierarchical mapping of item difficulty, each serving as an evidentiary basis for assessing construct alignment, measurement precision, and the instrument's sensitivity to naïve understanding linked response patterns. The presentation of results follows a results discussion integration format, with each empirical finding immediately followed by its theoretical interpretation and comparison with prior research to ensure analytical coherence and contextual relevance.

Reliability and Measurement Precision

The Rasch reliability and separation results, as shown in Table 2, indicate contrasting measurement characteristics between items and persons. The high item reliability (0.93) and item separation index (3.83) suggest that the items are well-calibrated and distributed across varying levels of difficulty, supporting stable estimation of item parameters and supporting the robustness of the instrument at the item level. In contrast, the very low person reliability (0.13) and person separation (0.22) indicate limited variability in participants' abilities, which limits the instrument's capacity to distinguish among different levels of learner proficiency. This pattern suggests that the sample may be relatively homogeneous, resulting in restricted response variance, which may be associated with limited variability in participant ability rather than necessarily indicating deficiencies in item quality [16], [41]. In Rasch measurement, person reliability is highly sensitive to the spread of ability distribution; therefore, when respondents exhibit similar ability levels, reliability estimates tend to decrease despite well-functioning items [42].

Table 2. Rasch reliability and separation summary

Parameter	Value	Interpretation
Item Reliability	0.93	High items are stably calibrated
Item Separation	3.83	Good; suggests adequate spread of item difficulty
Person Reliability	0.13	Low; reflects limited variability in participant ability
Person Separation	0.22	Low; indicates limited differentiation across learner proficiency
KR-20 (Internal Consistency)	0.13	Low; likely influenced by restricted response variance

This interpretation is also consistent with the discrepancy between the internal consistency obtained during pilot testing (KR-20 = 0.91) and that observed in the main data collection (KR-20 = 0.13). The high reliability in the pilot phase indicates that the instrument demonstrates potential for consistent measurement under conditions with sufficient variability in participant ability, while the substantially lower reliability in the

main study suggests that the observed limitation is more likely attributable to sample characteristics, particularly limited variability in student responses, rather than inherent flaws in the instrument design.

Comparable studies on static fluid instruments using Rasch modeling [17], [22] have reported similar patterns when sampling was limited to a single school context, indicating that low person reliability may arise from sampling constraints rather than instrument malfunction. Related Rasch-based instrument validations in other physics education domains also demonstrate that high item stability can coexist with limited person discrimination under conditions of restricted sample heterogeneity [1], [41]. Therefore, while the instrument demonstrates strong measurement properties at the item level, its person-level diagnostic precision remains sensitive to sample composition. This suggests that broader and more heterogeneous samples may be required to enhance person separation and improve the instrument's discriminative capability [37], [38]. Nevertheless, the current findings provide a preliminary basis for diagnostic application, particularly for analyzing item functioning and naïve understanding–linked response patterns.

For future applications, it is recommended that the instrument be administered to larger and more heterogeneous samples to improve person separation and measurement precision. Involving students from multiple schools or diverse academic backgrounds may help achieve a broader distribution of ability levels. Based on Rasch measurement guidelines, sample sizes of approximately 30–50 participants are generally sufficient for stable item calibration within ±1 logit at a 95–99% confidence level. Larger samples (e.g., 100–150) are generally recommended for higher precision (±0.5 logit) and improved generalizability of the measurement results [37], [38].

Item Fit and Naïve Understanding Patterns

Item fit was evaluated using both Mean Square (MNSQ) and standardized Z values (ZSTD) based on established Rasch criteria. Acceptable fit was defined as 0.5–1.5 for MNSQ and −2.0 to +2.0 for ZSTD. Values within 1.50–2.00 were considered indicative of potential misfit while still retaining diagnostic relevance, as such deviations may reflect meaningful variations in response patterns rather than random error [24], [43]. The analysis identified several items approaching the upper boundary of acceptable fit. In particular, items S7 (Infit MNSQ = 1.17; ZSTD = 1.59) and S11 (Infit MNSQ = 1.13; ZSTD = 1.40) exhibited localized deviations from model expectations. As shown in Table 3, both items have statistically acceptable response patterns but suggest possible inconsistencies in students' conceptual understanding. Although still within acceptable limits, they are considered borderline misfits and treated as near-misfit items for further diagnostic analysis [43].

Table 3. Summary of items approaching the misfit threshold and diagnostic interpretation

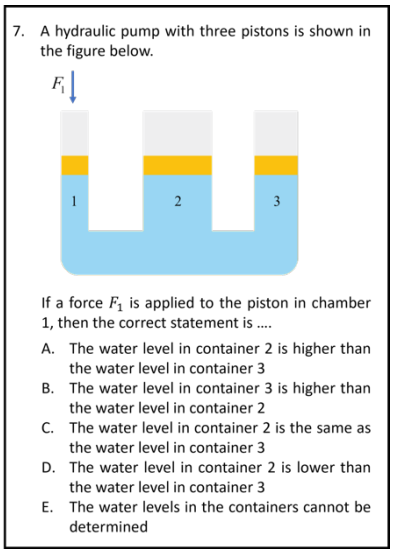
Item	Concept Focus	Infit Outfit		Classification	Naïve Understanding Triggered	Diagnostic Interpretation
		MNSQ	ZSTD			
S7	Hydraulic pressure distribution	1.37 / 1.22	1.59 / 1.59	Near-misfit	Context-dependent reasoning; difficulty coordinating pressure transmission across connected systems	May indicate potential difficulty in integrating spatial reasoning with Pascal's principle; potential conceptual fragmentation
S11	Buoyancy & density inference	1.36 / 1.12	1.40 / 1.00	Near-misfit	Formula-based (plug-and-chug) reasoning; reliance on surface-level proportional heuristics	May reflect a tendency toward procedural reasoning rather than conceptual integration of density and displacement relationships
S14	Buoyancy (daily context)	1.21 / 1.66	0.92 / 1.83	Borderline / overfit tendency	Over-regular response pattern; consistent intuitive schema	May reflect stable but non-scientific reasoning; requires cautious interpretation
S3	Pressure comparison (depth)	1.17 / 1.48	1.60 / 1.48	Near-misfit (upper boundary)	Partial neglect of depth or contextual inconsistency	May indicate sensitivity to representational context; may require refinement
S13	Archimedes (daily life transfer)	1.31 / 1.10	0.16 / 0.23	Well-fit (monitoring)	Experience-based reasoning	Conceptually acceptable but relatively superficial understanding
S2	U-tube pressure distribution	1.01 / 1.33	0.10 / 0.33	Well-fit (monitoring)	Intuitive balancing reasoning	Early-stage conceptual structuring retains diagnostic value

In Rasch measurement, borderline misfit reflects deviations from model-expected response patterns and is often interpreted as an indication of unexpected or inconsistent responses rather than definitive measurement error [16], [43]. Such deviations may arise in contexts where items require the coordination of multiple conceptual elements or involve cognitively demanding reasoning processes. In the present study, qualitative inspection of students' written explanations provides additional context for interpreting these patterns. For example, responses to item S11 suggest that some students relied on formula-based (plug and chug) or procedural strategies (e.g., direct application of density relationships) without fully integrating the underlying conceptual relationships between density, volume, and buoyant force. Similarly, responses to item S7 indicate that some students treated pressure as a localized effect of applied force, rather than as a uniformly transmitted quantity within a connected fluid system. These interpretations should be understood as indicative rather than definitive, as the analysis of written responses was limited in scope. Nevertheless, such patterns are consistent with findings in physics education research showing that students often rely on context-dependent or surface-level reasoning when dealing with conceptually demanding topics such as pressure transmission and buoyancy [12], [44].

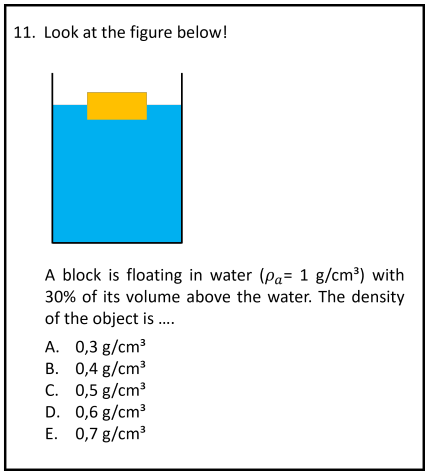
These patterns suggest that the observed misfit may be associated with inconsistencies in how learners coordinate formal scientific principles with intuitive reasoning, particularly in conceptually complex domains. This interpretation aligns with prior work in physics education research, indicating that student reasoning often emerges from the interaction between intuitive and formal knowledge structures, especially in situations involving conceptual conflict or transitional understanding [2], [20], [45].

Qualitative inspection of item content and students' written explanations provides additional context for interpreting these results. Item S7 (Figure 1a), which presents a hydraulic system involving differential piston areas, requires learners to coordinate Pascal's principle with spatial reasoning about fluid continuity. Response patterns suggest that some students may treat pressure as a localized consequence of applied force, predicting differential surface elevations across connected containers as if pressure propagation diminishes with distance. This tendency is consistent with prior findings indicating that students often conceptualize pressure transmission as directionally constrained rather than uniformly distributed [12], [46], [47].

Similarly, Item S11 (Figure 1b), which requires inference of density from buoyant equilibrium, elicits reasoning based on proportional surface heuristics. Analysis of students' explanations suggests that some rely on formula-based or plug-and-chug strategies (e.g., 30% volume above water, so density 0.3 g/cm³) without fully integrating the relational structure of Archimedes' principle, a pattern commonly associated with surface-level or heuristic-based reasoning in physics problem solving [44]. Such patterns have been widely reported in buoyancy contexts, where learners apply surface-level reasoning or computational shortcuts that are not fully grounded in conceptual understanding [45]. Taken together, these patterns suggest that the observed near-misfit may be associated with inconsistencies in how learners integrate formal scientific principles with intuitive reasoning, particularly in conceptually demanding domains such as pressure transmission and buoyancy.



(a)



(b)

Figure 1. (a) Test instrument S7, (b) Test instrument S11

Comparable patterns have been reported in Rasch-based diagnostics of conceptual physics assessments, where elevated misfit values tend to occur in items designed to elicit model competition, that is, when intuitive reasoning interacts with formal scientific constructs [14]. For example, Purwanto *et al.* [17] reported increased residuals in items targeting hydrostatic pressure differences due to students relying on shape-dependent interpretations of pressure. At the same time, another study [22] identified similar tendencies in buoyancy tasks involving proportional volume reasoning. In the present study, the observed near-misfit appears to be localized in items associated with conceptually demanding subdomains, particularly those requiring integration of multiple principles. This pattern suggests that the observed deviations may not be randomly distributed across the instrument, but may instead be associated with specific areas of conceptual difficulty.

However, this interpretation should be approached with caution, as additional analyses, such as residual correlation or distractor pattern analysis, were not conducted in this study. Therefore, while the observed deviations are consistent with theoretically expected areas of conceptual difficulty, further investigation is required to determine whether these patterns reflect stable cognitive structures or context-dependent response variability.

Within these limitations, the findings indicate that items such as S7 and S11 may provide useful diagnostic information by revealing response patterns associated with naïve or transitional understanding. Their borderline misfit, therefore, is retained not solely as evidence of measurement error, but as a potential indicator of underlying variability in students' conceptual understanding [41]. In this regard, the present instrument contributes to the perspective that Rasch analysis, when interpreted alongside learning-related evidence, can support both psychometric evaluation and exploratory insights into student reasoning in static fluid contexts.

Dimensionality & Construct Coherence

The Principal Components Analysis of Residuals (PCAR) indicated that the Rasch measures accounted for 32.2% of the total variance, while the unexplained variance in the first contrast was 2.96 eigenvalue units (13.4%). Subsequent contrasts decreased to 2.49 (11.3%) and 2.09 (9.1%), respectively. The essential unidimensionality index reached 41.0%, suggesting that a substantial proportion of item responses align with a single latent construct of static fluid conceptual understanding.

Although the first contrast slightly exceeds the commonly referenced 2.0 eigenvalue threshold used to flag potential multidimensionality, such results should be interpreted cautiously. In applied educational contexts, eigenvalues marginally above this threshold do not necessarily indicate the presence of a distinct secondary dimension, particularly when supported by theoretical coherence and item design considerations [15], [40]. The residual contrasts appear to correspond to groupings of conceptually demanding items, particularly those involving hydrostatic pressure propagation (S7), buoyancy reasoning (S14), and density–volume relationships (S11). This pattern suggests that the residual structure may reflect localized variations in item difficulty or response patterns rather than a clearly defined unintended secondary dimension [44].

Comparable findings have been reported in Rasch-based evaluations of fluid mechanics and conceptual physics instruments [17], [19], where residual contrasts tend to emerge in items that require coordination of multiple concepts or involve cognitively demanding reasoning. In such contexts, residual variance may be associated with differences in how students engage with complex conceptual tasks, rather than solely indicating violations of the unidimensional measurement assumption. In the present study, the concentration of residual variance in specific conceptually demanding items suggests that the observed multidimensional signals may not be randomly distributed, but instead linked to identifiable areas of conceptual challenge [15], [41]. From this perspective, PCAR results may provide complementary information about item functioning and response variability [24], [40], in addition to their role in evaluating dimensionality.

However, this interpretation should be approached with caution. The eigenvalue of the first contrast exceeds the conventional threshold, and additional analyses, such as subscale modeling or validation with more heterogeneous samples, would be necessary to more definitively evaluate the dimensional structure of the instrument. Overall, the findings suggest that the instrument demonstrates evidence of essential unidimensionality, while also exhibiting localized residual variation associated with conceptually complex subdomains [14], [15]. Rather than indicating a fundamental measurement flaw, these patterns may reflect the interaction between item demands and variations in students' conceptual reasoning, highlighting areas for further refinement and investigation in future studies.

Wright Map: Hierarchical Alignment of Item Difficulty and Learner Ability

The Wright Map, as presented in Figure 2, reveals a clustering of learner abilities around the logit mean, with very limited spread across the continuum (approximately -1.5 to +1.5 logits). In contrast, item difficulty spans a broader hierarchical range, from relatively accessible items (S10, S8, S2 at approximately –1 to –2 logits) to substantially more demanding items (S12, S13 at approximately +2.0 logits). This asymmetry confirms earlier findings: the instrument contains sufficient difficulty gradation, yet the participant sample demonstrates restricted variability in conceptual mastery.

This distribution indicates that the test is more effective for mapping item difficulty than for precisely differentiating learner ability in a homogeneous population. High-difficulty items (S14, S12, S13), located above most students on the logit scale, align with contexts where naïve understandings are activated, such as spontaneous reasoning about buoyancy forces (S14) or implicit geometric reasoning in hydraulic systems (S12, S13). These patterns are consistent with documented findings that students default to perceptual cues (e.g., height, shape, and visible force direction) rather than applying invariant principles such as Pascal's Law or Archimedes' principle [48], [49]. Conversely, items situated below the learner's mean (S10, S8, S2) capture phenomena where intuitive reasoning aligns more closely with normative conceptions, typically in cases without geometric manipulation or competing perceptual cues.

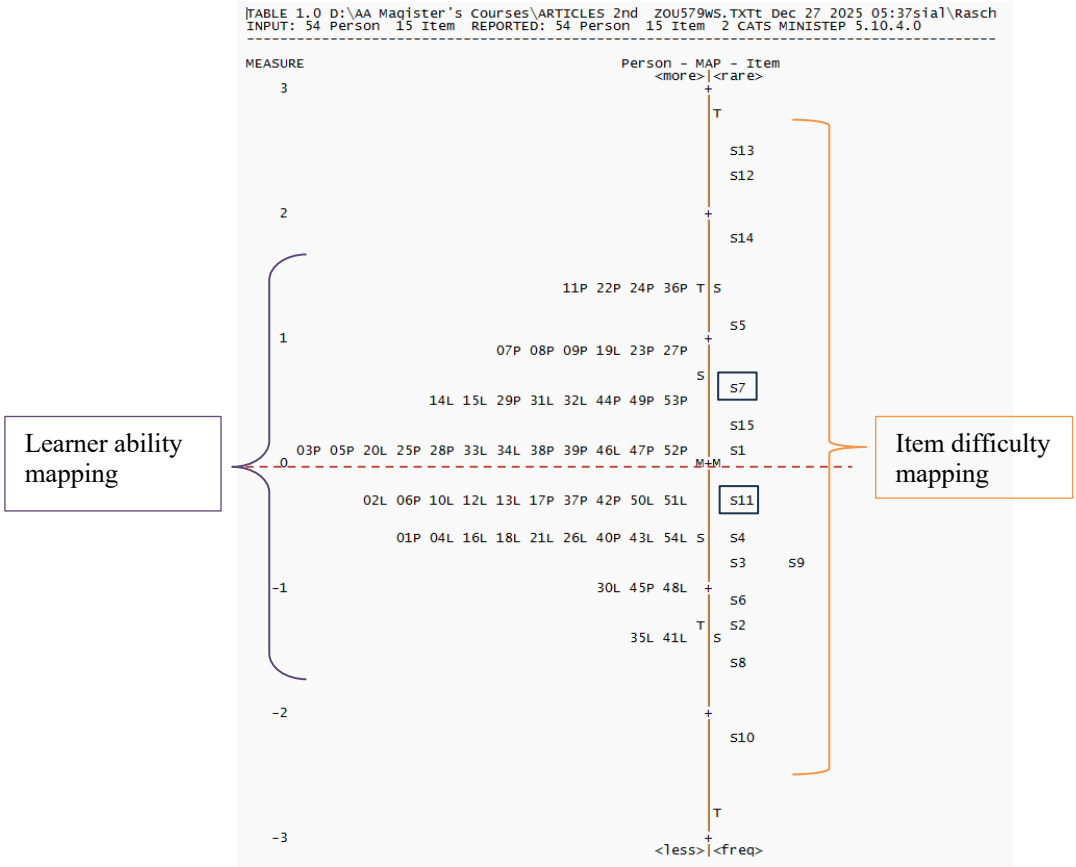


Figure 2. Wright Map

Wright's map shows that item S11 (0.00 logits) occupies the conceptual center of the scale, indicating that it functions as an anchor point for assessing baseline understanding of density buoyancy relations. S11 (Pascal context with floating block, 0 logits) indicates confusion between mass, weight, and density, buoyancy relations, mirroring findings reported [10], [12], [44], [50], that floating phenomena trigger default substance-based reasoning (lighter objects float because they weigh less), rather than equilibrium of forces. Conversely, S7 (+1.0 logits) is positioned at the upper boundary of the item hierarchy, reflecting its role as a high-leverage indicator of advanced conceptual coordination. The item's elevated difficulty suggests that recognizing pressure invariance across non-uniform containers requires restructuring of hybrid mental models that treat pressure as shape-dependent, a phenomenon widely documented as form-dependent pressure reasoning [49]. Together, these two items establish a conceptual gradient within the scale: S11 as a core conceptual anchor

and S7 as a threshold indicator, enabling the instrument to differentiate between surface-level correctness and structurally coherent understanding.

Comparable studies using Rasch modeling on static fluid inventories [17], [22] also report difficulty alignment concentrated in buoyancy and hydraulic contexts, yet the present results diverge in that the top-end items (S12–S14) create a clearer difficulty tier. This suggests that while the item set is not superior in scope, it contributes diagnostic granularity by clarifying where conceptual instability concentrates. Rather than implying greater instrument advancement, these differences offer an alternative mapping topology that may complement existing inventories. Collectively, the Wright Map confirms that the instrument's structural hierarchy is sensitive to the activation of naïve understandings in high-cognitive-load contexts, yet its capacity to classify learners remains limited by sample homogeneity. Thus, while the hierarchy supports diagnostic interpretation, broader sampling is required to appraise the instrument's full measurement potential and to validate whether its difficulty structure generalizes across populations.

Differential Item Functioning (DIF) and Fairness Across Gender

Based on the finalized DIF output for the L (male) and P (female) groups, none of the 15 items exceeded the combined criteria of statistical significance and substantive DIF magnitude (|logit| ≈ 0.5–0.7). While several items, such as S5 ($\chi^2 = 5.44$, $p = .0302$) and S9 ($\chi^2 = 6.52$, $p = .0172$), showed statistically significant contrasts at the nominal level, these values remain within a borderline range and should be interpreted cautiously rather than as definitive evidence of DIF. All other items align within expected measurement error ranges. Inspection of the DIF plot, illustrated in Figure 3, shows fluctuating profiles across items with no consistent advantage trend for either gender, supporting the absence of a clear systematic bias.

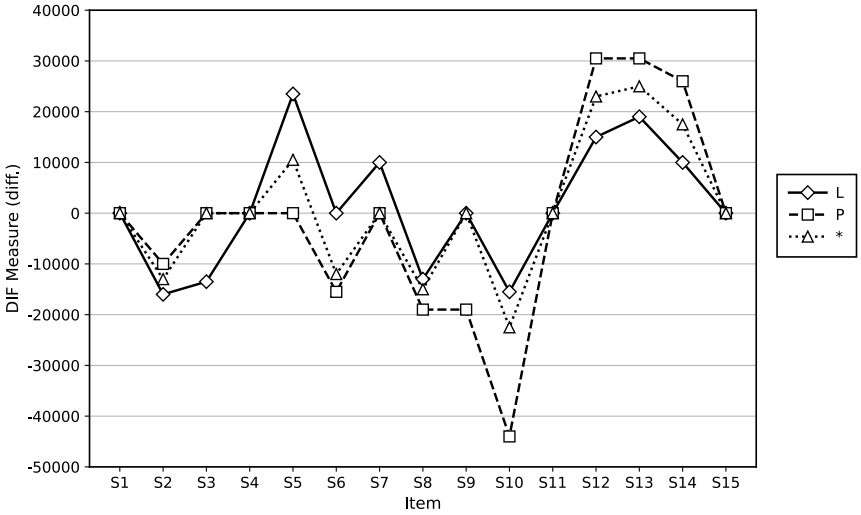


Figure 3. Person DIF Plot

These findings suggest that item functioning appears largely invariant across gender within the present sample, indicating no clear evidence of systematic bias. However, this interpretation should be treated with caution due to the relatively small subgroup sample sizes, which may limit the statistical power of DIF detection and increase the uncertainty of subgroup comparisons.

Minor contrasts in S5 (hydrostatic paradox context) and S9 (pressure depth comparison in asymmetric vessels) suggest that the observed contrasts may be associated with context-related response variation rather than construct-irrelevant bias. For instance, differences in familiarity with representational formats or prior exposure to specific problem contexts may contribute to localized variation in response patterns [12], [21], [44]. Such localized variation is consistent with previous Rasch-based studies, which report that small DIF contrasts often reflect contextual or instructional factors rather than systematic bias [15], [16], [41].

Comparable findings have also been reported in gender-based analyses of physics concept inventories, where borderline DIF is frequently linked to representational familiarity or contextual framing effects rather than necessarily reflecting inherent bias [51]. In fluid mechanics contexts, variations in visual–spatial representation (e.g., U-shaped tubes or buoyancy scenarios) have been shown to momentarily advantage certain groups depending on prior experience [49]. In line with these findings, the present results indicate that

DIF patterns are localized and item-specific rather than systematically favoring one group [50]. This supports the interpretation of psychometric fairness in the majority of items.

Nevertheless, given the limited sample size and subgroup distribution, further validation using larger and more balanced samples would be beneficial to more robustly evaluate measurement invariance. Future studies may also consider incorporating additional DIF detection approaches or cross-validation across different populations. Within these limitations, the current findings provide preliminary support for the instrument's use in evaluating conceptual understanding across mixed-gender populations, while highlighting specific items (S5 and S9) for further monitoring and potential refinement.

IV. Conclusions

This study reports the development and initial validation of a Static Fluid Concept Understanding Test, grounded in Rasch measurement principles, to examine students' conceptual understanding of hydrostatics, Pascal's law, buoyancy, and surface tension. The results indicate that the instrument demonstrates strong measurement properties at the item level, including high item reliability and stable item calibration across varying levels of difficulty. At the same time, several findings should be interpreted with appropriate caution. The observed near-misfit patterns in selected items (e.g., S7 and S11) may be associated with variations in students' reasoning, particularly in conceptually demanding contexts, rather than definitive measurement error. Similarly, the dimensionality analysis suggests evidence of essential unidimensionality, while also indicating localized residual variation that may reflect differences in how students engage with specific conceptual subdomains.

The analysis of Differential Item Functioning (DIF) further suggests that item performance appears broadly invariant across gender within the present sample, although the relatively small subgroup sizes limit the strength of this conclusion. In addition, the low person reliability observed in the main data collection highlights the influence of sample characteristics, particularly restricted variability in student ability, on measurement precision. Within these limitations, the instrument has potential as a diagnostic tool for exploring students' conceptual understanding and response patterns in static fluid contexts. Rather than serving as a definitive assessment framework, the instrument provides a preliminary basis for identifying areas of conceptual difficulty and informing further investigation.

Future research is needed to strengthen the instrument's validity and generalizability through larger, more heterogeneous samples, cross-institutional validation, and extended analyses of response patterns. Further refinement of specific items, particularly those exhibiting borderline misfit or context sensitivity, may also enhance the diagnostic resolution and interpretability of the instrument.

References

[1] S. I. Hofer, R. Schumacher, and H. Rubin, "The test of basic Mechanics Conceptual Understanding (bMCU): using Rasch analysis to develop and evaluate an efficient multiple choice test on Newton's mechanics," *Int. J. STEM Educ.*, vol. 4, no. 1, p. 18, Dec. 2017, doi: 10.1186/s40594-017-0080-5.

[2] J. L. Docktor and J. P. Mestre, "Synthesis of discipline-based education research in physics," *Phys. Rev. Spec. Top. - Phys. Educ. Res.*, vol. 10, no. 2, p. 020119, Sep. 2014, doi: 10.1103/PhysRevSTPER.10.020119.

[3] D. E. Brown, "Students' Conceptions as Dynamically Emergent Structures," *Sci. Educ.*, vol. 23, no. 7, pp. 1463–1483, Jul. 2014, doi: 10.1007/s11191-013-9655-9.

[4] E. Kuo, M. M. Hull, A. Elby, and A. Gupta, "Assessing mathematical sensemaking in physics through calculation-concept crossover," *Phys. Rev. Phys. Educ. Res.*, vol. 16, no. 2, p. 020109, Jul. 2020, doi: 10.1103/PhysRevPhysEducRes.16.020109.

[5] D. Naylor and S. S. H. Tsai, "Archimedes' principle with surface tension effects in undergraduate fluid mechanics," *Int. J. Mech. Eng. Educ.*, vol. 50, no. 3, pp. 749–763, Jul. 2022, doi: 10.1177/03064190211055431.

[6] G. C. Nihous, "Notes on hydrostatic pressure," *J. Ocean Eng. Mar. Energy*, vol. 2, no. 1, pp. 105–109, Feb. 2016, doi: 10.1007/s40722-015-0035-1.

[7] G. T. Pickett, "Volume Integral of the Pressure Gradient and Archimedes' Principle," Jul. 2014, [Online]. Available: http://arxiv.org/abs/1407.7562

[8] H. Saputro, L. Yuliati, P. Parno, S. Sunaryono, P. H. Winingsih, and R. Sebastian, "An Analysis of Senior High School Students' Problem-Solving Ability in Static Fluid Physics," in *Proceedings of International Conference on Teacher Profession Education*, Yogyakarta: Universitas Sarjana Wiyata Tamansiswa, 2025, pp. 32–40. [Online]. Available: https://seminar.ustjogja.ac.id/index.php/ICoTPE/article/view/3542

[9] A. A. diSessa, "Toward an Epistemology of Physics," *Cogn. Instr.*, vol. 10, no. 2–3, pp. 105–225, Apr. 1993, doi: 10.1080/07370008.1985.9649008.

[10] M. E. Loverude, C. H. Kautz, and P. R. L. Heron, "Helping students develop an understanding of Archimedes' principle. I. Research on student understanding," *Am. J. Phys.*, vol. 71, no. 11, pp. 1178–1187, Nov. 2003, doi: 10.1119/1.1607335.
[11] M. Goszewski, A. Moyer, Z. Bazan, and D. J. Wagner, "Exploring student difficulties with pressure in a fluid," 2013, pp. 154–157. doi: 10.1063/1.4789675.
[12] U. Besson, "Students' conceptions of fluids," *Int. J. Sci. Educ.*, vol. 26, no. 14, pp. 1683–1714, Nov. 2004, doi: 10.1080/0950069042000243745.
[13] Estianinur, Parno, E. Latifah, and M. Ali, "Exploration of students' conceptual understanding in static fluid through experiential learning integrated STEM with formative assessment," 2021, p. 050010. doi: 10.1063/5.0043129.
[14] M. Planinic, W. J. Boone, A. Susac, and L. Ivanjek, "Rasch analysis in physics education research: Why measurement matters," *Phys. Rev. Phys. Educ. Res.*, vol. 15, no. 2, p. 020111, Jul. 2019, doi: 10.1103/PhysRevPhysEducRes.15.020111.
[15] T. G. Bond and C. M. Fox, *Applying the Rasch Model*. Psychology Press, 2013. doi: 10.4324/9781410614575.
[16] W. J. Boone, "Rasch Analysis for Instrument Development: Why, When, and How?," *CBE—Life Sci. Educ.*, vol. 15, no. 4, p. rm4, Dec. 2016, doi: 10.1187/cbe.16-04-0148.
[17] M. G. Purwanto, A. Suhandi, B. Coştu, A. Samsudin, and M. Nurtanto, "Static Fluid Concept Inventory (SFCI): A Gender Gap Analysis using Rasch Model to Promote a Diagnostic Test Instrument on Students' Conception," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 6, pp. 3798–3812, 2020.
[18] T. Septiantini *et al.*, "Static Fluid Four-Tier Instrument (SFFTI): Develop and Identify K-11 Brebes-Scholars' Alternative Conception with Rasch Analysis," vol. 29, pp. 3190–3199, May 2020.
[19] N. Nurdini, A. Suhandi, T. Ramlan Ramalis, A. Samsudin, N. J. Fratiwi, and B. Coştu, "Developing Multitier Instrument of Fluids Concepts (MIFO) to Measure Student's Conception: A Rasch Analysis Approach," *J. Adv. Res. Dyn. Control Syst.*, vol. 12, no. 6, pp. 3069–3083, 2020, doi: 10.5373/jardcs/v12i6/s20201273.
[20] W. J. Boone, J. R. Staver, and M. S. Yale, *Rasch Analysis in the Human Sciences*. Dordrecht: Springer Netherlands, 2014. doi: 10.1007/978-94-007-6857-4.
[21] L. Bao and E. F. Redish, "Model analysis: Representing and assessing the dynamics of student learning," *Phys. Rev. Spec. Top. - Phys. Educ. Res.*, vol. 2, no. 1, p. 010103, Feb. 2006, doi: 10.1103/PhysRevSTPER.2.010103.
[22] I. D. A. Irawan, R. M. Indraloka, N. A. Basri, U. Salmah, and P. Parno, "Analysis of Concept Understanding Test Items on Static Fluid Material Using Rasch Model," *J. Pendidik. Fis.*, vol. 13, no. 1, pp. 1–13, Jan. 2025, doi: 10.26618/jpf.v13i1.15687.
[23] K. Khusaini, N. Azizah, P. Suwasono, C. I. Yogihati, and A. D. Andriani, "Rasch Model Application: Identification of Students' Conceptual Understanding of Static Fluid," *J. Pendidik. Fis. Indones.*, vol. 21, no. 1, pp. 54–68, Jun. 2025, doi: 10.15294/jpfi.v21i1.15023.
[24] M. Müller, "Item fit statistics for Rasch analysis: can we trust them?," *J. Stat. Distrib. Appl.*, vol. 7, no. 1, p. 5, Dec. 2020, doi: 10.1186/s40488-020-00108-7.
[25] F. Shaw, "Descriptive IRT vs. Prescriptive Rasch," Rasch Measurement Transactions.
[26] R. F. DeVellis and C. T. Thorpe, *Scale Development: Theory and Applications*. Sage Publications, Inc., 2021.
[27] W. Boone and J. Rogan, "Rigour in quantitative analysis: The promise of Rasch analysis techniques," *African J. Res. Math. Sci. Technol. Educ.*, vol. 9, no. 1, pp. 25–38, Jan. 2005, doi: 10.1080/10288457.2005.10740574.
[28] T. N. Diyana, S. Sutopo, and D. Haryoto, "The analysis of college students difficulty in acquiring static fluid concept," *Momentum Phys. Educ. J.*, pp. 11–18, May 2020, doi: 10.21067/mpej.v4i1.4113.
[29] V. R. Mustikasari, M. Annisa, and M. Munzil, "Identifikasi Miskonsepsi Konsep Tekanan Zat Siswa Kelas VIII-C SMPN 1 Karangploso Semester Genap Tahun Pelajaran 2017-2018," *J. Pembelajaran Sains*, vol. 1, no. 2, pp. 39–50, 2017, doi: 10.17977/um033v1i2p39-50.
[30] F. N. Anjelin, F. Ailiyah, B. R. Kurniawan, and M. N. Kholifah, "Identifikasi Penguasaan Konsep Materi Hukum Archimedes dan Hukum Pascal Berbantuan Quizizz," *Exp. J. Sci. Educ.*, vol. 1, no. 1, pp. 19–27, Dec. 2020, doi: 10.18860/experiment.v1i1.11114.
[31] Sri Rahmadani Pulu and Abd. Haji Amahoru, "Analisis Miskonsepsi Mahasiswa pada Pembelajaran IPA menggunakan Tes Diagnostik Multiple Choice Berbantuan CRI (Certainty of Response Index)," *J. Pendidik. MIPA*, vol. 13, no. 2, pp. 478–486, Jun. 2023, doi: 10.37630/jpm.v13i2.1039.
[32] I. G. M. Raga, "Identifikasi pemahaman konsep mahasiswa pada pokok bahasan fluida statis ditinjau dari teori resource," Universitas Negeri Malang, 2022.
[33] A. B. Rizkiyati, B. Supriadi, and M. Maryani, "Tingkat pemahaman konsep siswa SMKN 5 Jember pada pokok bahasan fluida statis menggunakan tes diagnostik four tier test," in *Prosiding Seminar Nasional Pendidikan Fisika*, Jember: Universitas Jember, 2018, pp. 197–202. [Online]. Available: https://garuda.kemdiktisaintek.go.id/documents/detail/878159
[34] V. M. Salma, S. E. Nugroho, and I. Akhlis, "Pengembangan E-Diagnostic Test Untuk Mengidentifikasi Pemahaman Konsep Fisika Siswa SMA Pada Pokok Bahasan Fluida Statis," *Unnes Phys. Educ. J.*, vol. 5, no. 1, 2016, doi: 10.15294/upej.v5i1.12701.
[35] I. N. A. Dewi, S. Kusairi, and L. Yuliati, "Miskonsepsi Siswa SMA pada Materi Hukum Archimedes'," in *Prosiding*

Seminar Nasional Tahun 2016, Surabaya: Universitas Negeri Surabaya, 2016, pp. 339–343. [Online]. Available: https://lib.um.ac.id/wp-content/uploads/2017/09/Miskonsepsi-Siswa-SMA-Pada-Materi-Hukum-Archimedes.pdf
[36] G. Pescaroli, O. Velazquez, I. Alcántara-Ayala, C. Galasso, P. Kostkova, and D. Alexander, "A Likert Scale-Based Model for Benchmarking Operational Capacity, Organizational Resilience, and Disaster Risk Reduction," *Int. J. Disaster Risk Sci.*, vol. 11, no. 3, pp. 404–409, Jun. 2020, doi: 10.1007/s13753-020-00276-9.
[37] B. Sumintono and W. Widhiarso, *Aplikasi Pemodelan Rasch: pada Assessment Pendidikan*. Penerbit Trim Komunikata, 2015.
[38] J. M. Linacre, A. W. Heinemann, B. D. Wright, C. V. Granger, and B. B. Hamilton, "The structure and stability of the functional independence measure," *Arch. Phys. Med. Rehabil.*, vol. 75, no. 2, pp. 127–132, Feb. 1994, doi: 10.1016/0003-9993(94)90384-0.
[39] J. M. Linacre, "A User's Guide to Winsteps: Rasch-Model Computer Programs," Winsteps.com. Accessed: Apr. 29, 2026. [Online]. Available: https://www.winsteps.com/a/Winsteps-Manual.pdf
[40] J. M. Linacre, "Data variance explained by Rasch measures," Rasch Measurement Transactions. Accessed: Apr. 29, 2026. [Online]. Available: https://www.rasch.org/rmt/rmt201a.htm
[41] T. Bond, Z. Yan, and M. Heene, *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. New York: Routledge, 2020. doi: 10.4324/9780429030499.
[42] Y. Yang *et al.*, "Psychometric evaluation of the academic involution scale for college students in China: An application of Rasch analysis," *Front. Psychol.*, vol. 14, Feb. 2023, doi: 10.3389/fpsyg.2023.1135658.
[43] Y. Jin, C. A. Rodriguez, L. Shah, and G. T. Rushton, "Examining the Psychometric Properties of the Redox Concept Inventory: A Rasch Approach," *J. Chem. Educ.*, vol. 97, no. 12, pp. 4235–4244, Dec. 2020, doi: 10.1021/acs.jchemed.0c00479.
[44] C. R. Gette, M. Kryjevskaia, M. R. Stetzer, and P. R. L. Heron, "Probing student reasoning approaches through the lens of dual-process theories: A case study in buoyancy," *Phys. Rev. Phys. Educ. Res.*, vol. 14, no. 1, p. 010113, Mar. 2018, doi: 10.1103/PhysRevPhysEducRes.14.010113.
[45] R. Duit and D. F. Treagust, "Conceptual change: A powerful framework for improving science teaching and learning," *Int. J. Sci. Educ.*, vol. 25, no. 6, pp. 671–688, Jun. 2003, doi: 10.1080/09500690305016.
[46] C. Cari, S. N. Pratiwi, H. Affandy, and D. A. Nugraha, "Investigation of undergraduate student concept understanding on Hydrostatic Pressure using two-tier test," *J. Phys. Conf. Ser.*, vol. 1511, no. 1, p. 012085, Apr. 2020, doi: 10.1088/1742-6596/1511/1/012085.
[47] N. Bessas, E. Tzanaki, D. Vavougios, and V. P. Plagianakos, "Diagnosing students' misconception in Hydrostatic Pressure through a 4-tier test," *Heliyon*, vol. 10, no. 23, p. e40425, Dec. 2024, doi: 10.1016/j.heliyon.2024.e40425.
[48] G. Ozkan and G. S. Selcuk, "Facilitating conceptual change in students' understanding of concepts related to pressure," *Eur. J. Phys.*, vol. 37, no. 5, p. 055702, Sep. 2016, doi: 10.1088/0143-0807/37/5/055702.
[49] D. Andrich, *Rasch Models for Measurement*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc., 1988. doi: 10.4135/9781412985598.
[50] D. Kaltakci Gurel, A. Eryilmaz, and L. C. McDermott, "A Review and Comparison of Diagnostic Instruments to Identify Students' Misconceptions in Science," *EURASIA J. Math. Sci. Technol. Educ.*, vol. 11, no. 5, Oct. 2015, doi: 10.12973/eurasia.2015.1369a.
[51] E. Valencia, "Gender-biased evaluation or actual differences? Fairness in the evaluation of faculty teaching," *High. Educ.*, vol. 83, no. 6, pp. 1315–1333, Jun. 2022, doi: 10.1007/s10734-021-00744-1.