DEVELOPMENT OF THE FOUR-TIER DIAGNOSTIC TEST TO IDENTIFY STUDENT MISCONCEPTIONS IN THE STATIC FLUIDS CHAPTER

Himawan Putranta✉
Fahdah Afifah

Department of Physics Education, Faculty of Tarbiyah and Teacher Training, Universitas Islam Negeri Sunan Kalijaga, Indonesia

✉ himawan.putranta@uin-suka.ac.id

ABSTRACT

Misconceptions about static fluid concepts in physics are common among students, making it essential for teachers to identify and address them. This research aims to develop and evaluate the quality of a four-tier diagnostic test instrument and identify student misconceptions in the static fluid chapter. The sample for this research comprised 91 grade 11 students from the State Madrasah Aliyah in Bantul, Indonesia, selected using a purposive sampling technique. The findings of this research indicate that the four-tier diagnostic test instrument is suitable for identifying student misconceptions. The test validation results showed 17 valid test items and 1 invalid test item. This diagnostic test is reliable, with a person reliability coefficient of 0.73, an item reliability coefficient of 0.96, and a Cronbach's alpha coefficient of 0.72. The test items include two very difficult, seven difficult, and nine moderate items. The discrimination power of the 17 test items is good, except for one, which is poor. This instrument also found that the most common student misconceptions were in the hydrostatic pressure sub-chapter (71% of students) and in surface tension (66% of students).

KEYWORDS

Conceptual understanding, four-tier diagnostic test, misconceptions, static fluids, Winstep software

HOW TO CITE

Putranta H., Afifah F. (2025) 'Development of the Four-Tier Diagnostic Test to Identify Student Misconceptions in the Static Fluids Chapter', *Journal on Efficiency and Responsibility in Education and Science*, vol. 18, no. 4, pp. 268–281. http://dx.doi.org/10.7160/eriesj.2025.180403

Article history
Received
May 3, 2024
Received in revised form
May 7, 2025
Accepted
December 2, 2025
Available on-line
December 31, 2025

Highlights

- *A Rasch-validated four-tier diagnostic test effectively detects students' misconceptions in static fluid concepts.*
- *Students exhibit high-confidence misconceptions predominantly in hydrostatic pressure and surface tension topics.*
- *Competency-based diagnostic mapping provides actionable evidence for targeted and efficient instructional remediation.*

INTRODUCTION

Physics is a field of knowledge that focuses on natural phenomena (Kösem and Özdemir, 2014). Before being formally taught in schools, students already have an initial understanding of basic physics concepts through their experiences with natural phenomena in everyday life. Understanding physics concepts is key to success in physics learning (Cai et al., 2021). Students' conceptual understanding includes their comprehension of a concept, especially in the context of physics, which enables them to express it, depict it in various forms of representation, provide examples related to the concept, and apply it to solve everyday problems (Ozkan and Selcuk, 2015). Concepts in physics have been clearly articulated and accepted by scientists. If someone has personal interpretations or understanding of these concepts, it is referred to as conceptions.

Concepts are crucial components for students. Students can develop conceptual understanding through school experiences or daily activities (Maknun, 2020). Students' diverse experiences can influence whether they understand concepts through scientific conceptions (Mason and Just, 2016). There is a correlation between conceptual understanding and misconceptions. Misconceptions are concepts in students' minds that may not align with expert concepts, potentially misleading students' learning (Dack, 2019). Misconceptions arise when students' prior knowledge deviates from scientifically accepted concepts (Hill and Chin, 2018). Misconceptions can manifest as errors in initial understanding, errors in linking different concepts, and inappropriate ideas. Students' misconceptions can also stem from various factors, such as internal factors, teacher negligence, deficiencies in textbooks, contextual issues, and mismatches in the teaching methods used by teachers during the learning process (Suprapto, 2020). Teachers need to address these misconceptions, as they can hinder students' ability to receive and assimilate new knowledge, ultimately

affecting students' success in the learning process (Assem et al., 2023). Neglecting students' conceptions in learning activities can make the learning process difficult for students, thus potentially lowering their learning achievements by up to 80% below the minimum completion (Neito et al., 2025). The danger of misconceptions becomes more serious if left unaddressed, as they can affect students' understanding of future concepts (Hasanah, 2020). Teachers must understand and detect students' misconceptions to help them overcome them effectively.

However, a few teachers, around fewer than 50%, still pay attention to methods of identifying and resolving student misconceptions (Robbins et al., 2025). This indication is also evident from initial interviews conducted at one of the MAN (Islamic Senior High School) in the Bantul region. The results of interviews with physics teachers indicate that student learning outcomes in static fluids remain low, with only 40% of students completing the material. Additionally, physics teachers at the MAN had never used the four-tier diagnostic test to identify student misconceptions, and these misconceptions usually surfaced when students asked questions or expressed their understanding. This was found by teachers accidentally. However, this method is ineffective because most students feel embarrassed and reluctant to ask the teacher further questions about concepts they do not understand.

Several methods can be used to evaluate students' understanding and misconceptions, including concept mapping, concept-related interviews, and diagnostic test instruments. One common approach to identifying student misconceptions is the use of diagnostic tests, according to Taslidere (2016), which are evaluations designed to identify students' weaknesses and facilitate appropriate improvement. Diagnostic tests also aim to evaluate indications such as conceptual understanding, misconceptions, and a lack of understanding of concepts. Some variants of multiple-choice diagnostic tests include one-tier, two-tier, three-tier, and four-tier diagnostic tests (Türkoguz, 2020). The four-tier diagnostic test is one diagnostic test that can identify misconceptions. The four-tier diagnostic test was developed from the three-tier multiple-choice diagnostic test (Istiyono et al., 2023). The development lies in adding students' confidence in choosing answers and reasons. The first tier consists of multiple-choice questions with three distractors and one correct answer, which students must select. The second tier assesses students' confidence levels in the chosen answer. The third tier evaluates students' reasons for answering questions, with four provided options. The fourth tier assesses students' confidence levels in choosing those reasons (Caleon and Subramaniam, 2010).

In addition, this study adopted a four-level diagnostic test instrument because it has several advantages. The four-tier diagnostic test instrument has several advantages, including its ability to differentiate more deeply between students' confidence levels in their answers and the reasons they chose (Madina et al., 2022). Providing a more comprehensive diagnosis related to student misconceptions. Identifying more accurately the parts of the material that require additional attention and assisting in planning more effective learning to address student misconceptions (Wahyuni et al., 2021). Meanwhile, this study aims to construct and test the feasibility of a four-tier diagnostic test instrument. To identify misconceptions of grade XI students regarding static fluid using a four-level diagnostic test instrument.

Problem statement

Physics learning in schools often faces challenges, including students' misconceptions. These misconceptions arise because students initially understand physics concepts through everyday experiences before receiving formal learning. This understanding does not always align with the scientific concepts accepted by experts, which can hinder students' learning process. According to Caleon and Subramaniam (2010), misconceptions that are not properly addressed can interfere with the understanding of advanced concepts and affect student learning outcomes. In line with this statement, one of the physics materials that often experiences misconceptions is static fluid. Research by Putra et al. (2020) shows that students have difficulty understanding the concept of hydrostatic pressure and Pascal's law. This is due to the wrong initial understanding and ineffective learning methods for identifying and overcoming these misconceptions. In addition, teachers are often unaware of misconceptions because the evaluation tools available to them do not adequately assess students' understanding in depth. Thus, to overcome this problem, an evaluation instrument is needed that accurately identifies students' misconceptions. The four-tier diagnostic test is an effective tool in revealing students' misconceptions. This test not only assesses students' answers but also their reasons and confidence level in those answers. According to research by Çelikkanlı and Kızılcık (2022), the use of a four-tier diagnostic test can help teachers design more targeted learning strategies to address students' misconceptions. However, the implementation of this test remains limited in educational settings, so further efforts are needed to develop and apply it more widely in physics learning.

METHOD

Research design

This study adopts a research and development (R&D) approach to design and validate an innovative diagnostic tool addressing student misconceptions about the static fluid concept in physics. Meanwhile, the research model used is the development study model, which aims to address educational problems, particularly student misconceptions, by developing an evaluation tool (Van den Akker et al., 2006). The first stage of this development study is the preliminary research phase, which includes a literature review and field surveys. The second stage is the prototype development stage, which includes creating, validating, and revising the four-tier instrument. The third stage is the summative evaluation, which encompasses both limited and extensive testing. During the testing stage of the four-tier diagnostic test, a pre-experimental single-group post-test design was used (Setiawan et al., 2019). This design was chosen to measure students' misconceptions about the static fluids chapter using a four-tier diagnostic test after students received instruction on the physics

topic. Meanwhile, the fourth or final stage is the systematic reflection and documentation stage, which involves reporting the research.

Participants

The test subjects in this study were selected using a purposive sampling method. The decision to use purposive sampling was made because all test subjects had studied the static fluid chapter (Maison et al., 2020). Class XI MIPA 1 was chosen as the subject for limited testing, while XI MIPA 2 and XI MIPA 3 were chosen for extensive testing. They selected class XI because they had studied the static fluid chapter. Additionally, based on interviews with physics teachers at one of the MANs in Bantul, it was found that students' learning outcomes in the static fluid chapter remained low, with student completion rates of only 40%. This indicates students' misconceptions about static fluid chapters. Information regarding the number of students can be found in Table 1.

Class	Total Students
XI MIPA 1	26
XI MIPA 2	32
XI MIPA 3	33
Total	91

Table 1: Subject profile of research on identifying static fluid misconceptions

Instrument and procedures

Students' misconceptions about static fluids are measured using a four-tier diagnostic test. Researchers and participants followed ethical procedures throughout data collection. The initial stage of this research was to develop a four-tier diagnostic test instrument to measure student misconceptions in the chapter on static fluids, which is valid and reliable. The researchers developed a new instrument to ensure that the four-tier diagnostic test instrument was relevant to the research objective of measuring students' misconceptions in the static fluid chapter.

A literature review on static fluids was conducted before designing the four-tier diagnostic test instrument. In the static fluid chapter, six sub-chapters serve as the basis for developing four-tier diagnostic test instruments: density, hydrostatic pressure, Archimedes' law, surface tension, capillarity, and viscosity (Asrizal et al., 2022; Kundu et al., 2015). Furthermore, a four-tier diagnostic test instrument was developed to measure students' misconceptions about static fluids, comprising 18 items. The test instruments developed are then validated by five expert validators before being distributed to students. In general, the procedures carried out in this research can be illustrated in Figure 1.

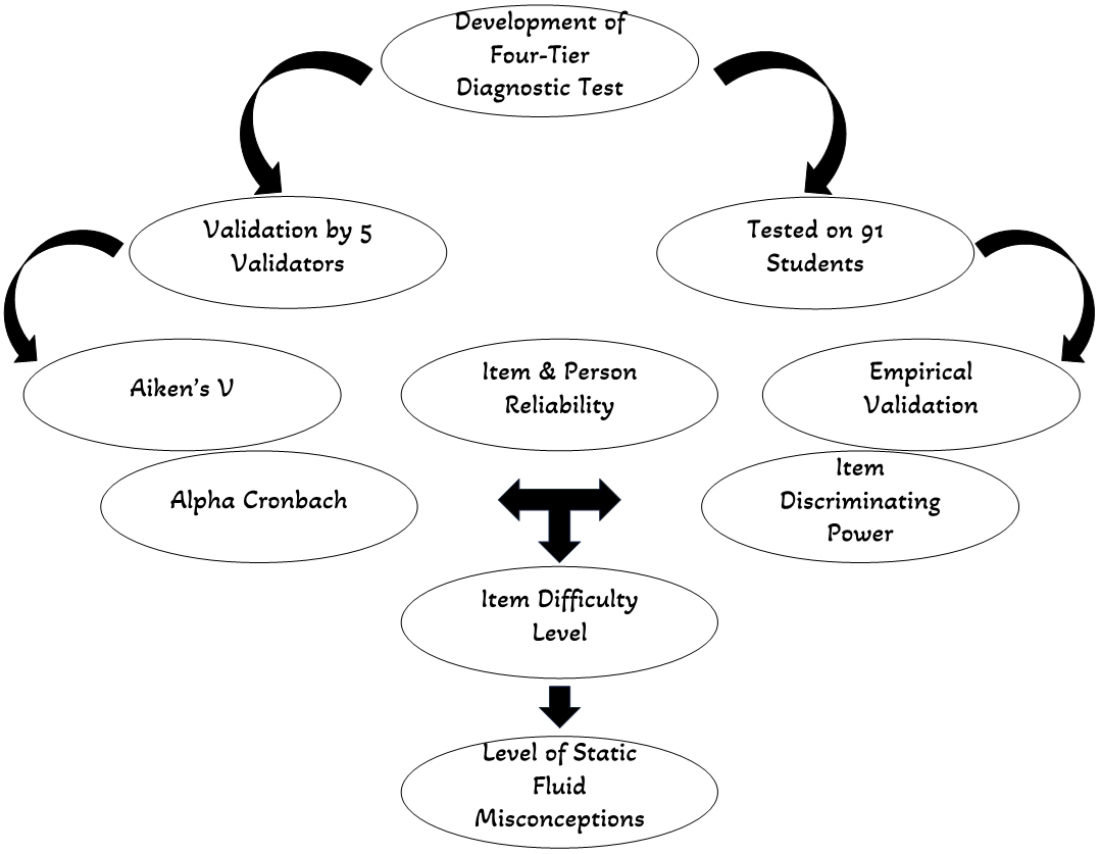

Figure 1: the process of developing an instrument to identify static fluid misconceptions

Based on Figure 1, which outlines the flow of this research process, in analyzing student misconceptions, a combination of the four-tier diagnostic test answer categories presented in Table 2 can be used (Yuberti et al., 2020).

No.	Category	Combination Answers				Score
		Answers	Confidence Level	Reason	Confidence Level	
1.	Understanding of Concept (UC)	Correct	Sure	Correct	Sure	4
2.	Partially Understand (PU)	Correct	Sure	Correct	Unsure	3
3.	Partially Understand (PU)	Correct	Unsure	Correct	Sure	3
4.	Partially Understand (PU)	Correct	Unsure	Correct	Unsure	3
5.	Partially Understand (PU)	Correct	Sure	Wrong	Sure	3
6.	Partially Understand (PU)	Correct	Sure	Wrong	Unsure	3
7.	Partially Understand (PU)	Correct	Unsure	Wrong	Sure	3
8.	Partially Understand (PU)	Correct	Unsure	Wrong	Unsure	3
9.	Partially Understand (PU)	Wrong	Sure	Correct	Sure	3
10.	Partially Understand (PU)	Wrong	Sure	Correct	Unsure	3
11.	Partially Understand (PU)	Wrong	Unsure	Correct	Sure	3
12.	Partially Understand (PU)	Wrong	Unsure	Correct	Unsure	3
13.	Misconceptions (M)	Wrong	Sure	Wrong	Sure	2
14.	Don't Understand of Concept (DUC)	Wrong	Sure	Wrong	Unsure	1
15.	Don't Understand of Concept (DUC)	Wrong	Unsure	Wrong	Sure	1
16.	Don't Understand of Concept (DUC)	Wrong	Unsure	Wrong	Unsure	1
17.	Cannot Be Encoded (CBE)	Some tiers are not answered, or more than one option is available	Some tiers are not answered, or more than one option is available	Some tiers are not answered, or more than one option is available	Some tiers are not answered, or more than one option is available	0

Table 2: Categories of combination answers in the four-tier diagnostic test

Based on the four-tier diagnostic test instrument developed, it is more closely aligned with the ability to solve physics problems. However, students not only choose the correct answer and reason, but also believe in them.

Validity and reliability of the four-tier diagnostic test

Five expert validators, including two lecturers, two physics teachers, and one colleague (a physics education practitioner), first tested the four-tier diagnostic test instrument for validity and reliability. The validity test of this four-tier diagnostic test instrument includes a content validity test and an empirical validity test. Content validity is analyzed using Aiken's V equation. An item in the content validity test is considered valid if the calculated Aiken's V coefficient value is greater than or equal to the minimum value listed in Aiken's table (Aiken, 1985). The results of the content validity analysis of the four-tier diagnostic test items can be shown in Table 3.

Based on Table 3, all four-tier diagnostic test items were declared valid with revisions by the five validators. The items can be tested on students after being revised based on the validator's suggestions. Based on the results of limited trials, an empirical validity analysis can be conducted to assess the suitability of each item with the fit model (Van Laar et al., 2018). The empirical validity of each item can be assessed by reviewing the WINSTEPS program output, including MNSQ, ZTSD, and Pt Measure Correlation values. Outfit Mean Square (MNSQ) is a component of empirical validity that measures how much the data deviate from the Rasch model, especially in cases of extreme misfit (e.g., answers that do not match the questions, or questions that are too easy or too difficult). Outfit Z-standard (ZTSD) transforms the Outfit MNSQ value into a standard z score, which follows a normal distribution. Meanwhile, Point Measure Correlation (Pt Measure Correlation) is the correlation between responses to specific items and participants' total ability, as defined by the Rasch model (Lee et al., 2020). Meanwhile, empirical validity can be assessed using the results of limited trials, as shown in Table 4. Based on Table 4, 17 items are empirically valid, but 1 item, item 5, is declared invalid. Questions with invalid criteria cannot be used and are discarded in extensive trials. Meanwhile, the reliability of the four-tier diagnostic test was assessed using the item separation index (item estimation) and the person separation index (case estimation) in the WINSTEPS program (Laliyo et al., 2021). The greater the item separation index value, the greater the accuracy of the test items using the PCM model. In addition, the higher the person-separation index value, the greater the consistency of each item in measuring people's abilities (Lightfoot et al., 2021). The reliability of the four-tier diagnostic test was also analyzed using Cronbach's Alpha equation. Cronbach's Alpha measures the interaction between the subjects and the items (Vaske et al., 2017). Meanwhile, the reliability analysis results for the four-tier diagnostic test items are shown in Table 5.

Item	Aiken's V Coefficient	Criteria
1	0.83	Valid
2	0.80	Valid
3	0.85	Valid
4	0.80	Valid
5	0.80	Valid
6	0.81	Valid
7	0.81	Valid
8	0.82	Valid
9	0.82	Valid
10	0.80	Valid
11	0.85	Valid
12	0.84	Valid
13	0.83	Valid
14	0.83	Valid
15	0.83	Valid
16	0.81	Valid
17	0.85	Valid
18	0.84	Valid

Table 3: Content validity results of four-tier diagnostic test items

Item	MNSQ	ZTSD	Pt Measure Correlation	Criteria
1	0.92	-0.2	0.33	Valid
2	1.39	1.1	0.40	Valid
3	1.59	2.0	-0.20	Valid
4	0.73	-0.7	0.42	Valid
5	**0.42**	**-2.9**	**0.27**	**Invalid**
6	1.34	1.3	-0.27	Valid
7	0.98	0.1	0.69	Valid
8	1.04	0.3	0.30	Valid
9	1.02	0.2	0.60	Valid
10	1.08	0.4	0.38	Valid
11	0.74	-1.1	0.38	Valid
12	0.93	-0.2	0.80	Valid
13	1.11	0.5	0.59	Valid
14	1.13	0.5	0.62	Valid
15	0.82	-0.7	0.26	Valid
16	0.73	-1.1	0.47	Valid
17	0.77	-0.8	0.29	Valid
18	0.67	-1.0	0.80	Valid

Table 4: Empirical validity results of four-tier diagnostic test items

Limited Trial			
Reliability Type	**Reliability Coefficient**	**Information**	**Criteria**
Person reliability	0.72	Reliable	Moderate
Alpha Cronbach	0.74	Reliable	Good
Item Reliability	0.90	Reliable	Good
Extensive Trial			
Reliability Type	**Reliability Coefficient**	**Information**	**Criteria**
Person reliability	0.73	Reliable	Moderate
Alpha Cronbach	0.72	Reliable	Good
Item Reliability	0.96	Reliable	Excellent

Table 5: Reliability results of four-tier diagnostic test items

Based on Table 5, the reliability values, including person reliability, item reliability, and Cronbach's alpha, are all reliable. This means that the developed four-tier instrument can be used to identify students' misconceptions about static fluid chapters.

Difficulty level and differentiating power of the four-tier diagnostic test

Test items are effective when their difficulty is balanced, neither too easy nor too difficult. A moderate difficulty level is optimal because it encourages students to apply more effort in solving problems (Hong et al., 2021). Test items that are too easy tend to make students less likely to improve their understanding, while those that are too difficult cause students to lose enthusiasm and feel inadequate (Lourdusamy and Magendiran, 2021). The difficulty level of the four-tier diagnostic test in this study was analyzed using the WINSTEPS program. The item difficulty index is obtained from item statistical output, which includes the item score or the number of students who answered correctly, and the logit measure for each item (Saidi and Siew, 2019). The criteria for the logit measure values of the test items are shown in Table 6.

Logit Measure Value	Criteria
> +1.37 SD	Very Difficult
0.00 to +1.37 SD	Difficult
-1.37 SD to 0.00	Moderate
< -1.37 SD	Easy

Table 6: Conditions for logit measure values

The item difficulty level test was conducted to evaluate the quality of the four-tier diagnostic test items. Test items are considered good if the difficulty level is adequate or moderate (Bichi and Talib, 2018). Meanwhile, analysis of the test's differentiating power assesses its ability to separate students with high abilities from those with low abilities (Machts et al., 2016). If students with both high and low ability can answer a test item correctly, then the test item is considered ineffective because it lacks differentiating power. On the other hand, if all students in the low and high groups give incorrect answers, the test items lack differentiating power (Rabin et al., 2021). The differentiating ability of each test item was analyzed using the WINSTEPS program. Analysis of the test items' differentiating power is reflected in the DIF output, with a focus on the resulting probability values. An item is considered to have good discriminating power if the resulting probability value is greater than 0.05 (Iqbal and Malzahn, 2017). An item can be relied upon as a data collection tool if it has sufficient item-distinguishing power.

Data analysis

The analysis used to identify student misconceptions in the static fluid chapter measured students' conceptual understanding levels as percentages. Student misconceptions are identified by paying attention to student answers, which can be grouped based on Table 2 (Saricayir et al., 2016). Meanwhile, the percentage of students' understanding of concepts in the static fluid chapter can be calculated using the equation presented in Table 7.

Category	Analytical Equations	Information
Understanding of Concept (*UC*)	$UC = \frac{n_{UC}}{N} \times 100\%$	n_{UC} is the number of students experiencing UC
Partially Understand (*PU*)	$PU = \frac{n_{PU}}{N} \times 100\%$	n_{PU} is the number of students experiencing PU
Misconceptions (*M*)	$M = \frac{n_{M}}{N} \times 100\%$	n_{M} is the number of students experiencing M
Don't Understand of Concept (*DUC*)	$DUC = \frac{n_{DUC}}{N} \times 100\%$	n_{DUC} is the number of students experiencing DUC

Note: N is the total number of students

Table 7: Equation of concept understanding level analysis

RESULTS

The results of this research include the construct results, the quality of the four-tier diagnostic test instrument, and the identification of class XI students' misconceptions regarding static fluids. The results obtained from this research are a development of the four-tier diagnostic test instrument that has been tested.

Four-tier diagnostic test instrument construction

The construct results found that the four-tier diagnostic test instrument developed consisted of 17 questions grouped into six indicators of competency achievement. The four-tier diagnostic test instrument includes a test question grid, work instructions, an answer key, an answer sheet, and scoring guidelines. The grouping of four-tier diagnostic test instrument

questions into six indicators of competency achievement is shown in Table 8.

Based on Table 8, the most common four-tier diagnostic test items are on hydrostatic pressure. This is because the concept is still a physics concept that students find difficult to understand in the static fluid chapter. Furthermore, the preliminary study results also show that students still experience errors when asked to analyze the application of hydrostatic pressure in everyday life. Density, surface tension, and viscosity are static fluid concepts that are difficult to understand under hydrostatic pressure. So, the four-tier diagnostic test items for these three concepts are arranged into three items each. Meanwhile, two four-tier diagnostic tests are formulated based on Archimedes' law and capillarity. This is because the concept of static fluids is easiest for students to understand.

No.	Competence Achievement Indicators (CAI)	Items
1.	Density	5, 13,15
2.	Hydrostatic Pressure	1, 4, 12, 14
3.	Archimedes' Law	3, 9
4.	Surface Tension	7, 10, 11
5.	Capillarity	6, 8
6.	Viscosity	2, 16, 17

Table 8: Grouping of four-tier diagnostic test instrument items

Results of the difficulty level and the differentiating power of the four-tier diagnostic test

Analysis of the difficulty level of test items aims to evaluate their quality as a data-collection tool, with the criterion that they fall into the category of good test questions. A test item is considered good if it has a balanced difficulty level, is neither too easy nor too difficult, or can be categorized as moderate. This study analyzed the difficulty level of four-tier diagnostic test items using the WINSTEPS application, shown in Table 9.

Items	Score 1	Score 2	Score 3	Score 4	Count	Measure	Criteria
11	3	16	6	1	26	1.72	Very Difficult
6	-	20	4	2	26	1.51	Very Difficult
8	4	7	14	1	26	1.24	Difficult
15	-	11	12	3	26	0.84	Difficult
10	2	7	12	5	26	0.70	Difficult
5	-	3	22	1	26	0.42	Difficult
12	3	5	8	10	26	0.34	Difficult
16	1	4	15	6	26	0.27	Difficult
17	1	1	18	6	26	0.04	Difficult
3	2	1	14	9	26	-0.04	Moderate
13	1	5	7	13	26	-0.20	Moderate
14	1	4	5	16	26	-0.56	Moderate
7	1	3	5	17	26	-0.76	Moderate
1	-	1	11	14	26	-0.87	Moderate
18	1	1	7	17	26	-0.98	Moderate
2	1	1	6	18	26	-1.10	Moderate
4	-	-	10	16	26	-1.23	Moderate
9	1	2	2	21	26	-1.36	Moderate

Table 9: Difficulty level of four-tier diagnostic test items

Table 9 shows that the difficulty level evaluation includes two items in the very difficult category: items 6 and 11. There are seven items with a difficult level of difficulty and nine with a medium level of difficulty. Furthermore, the discriminating power of test items reflects a test's ability to differentiate between students with high and low ability. A question is considered to have good discriminating power if the resulting probability value exceeds 0.05. This study analyzed the discriminating power of four-tier diagnostic test items using the WINSTEPS application, presented in Table 10.

Based on the analysis of the differentiating power of the questions in Table 10, 18 questions have logit values above 0.05, indicating that the test items exhibit good quality. Meanwhile, 1 question 14 has a logit value below 0.05. Questions with a logit value below 0.05 are considered to have poor discriminating power. Therefore, revisions are needed to point number 14 so that it can serve as a reliable tool for assessing students' misconceptions in static-fluid chapters.

Item	Aiken's V Coefficient	Criteria
1	0.710	Good
2	0.610	Good
3	0.341	Good
4	0.153	Good
5	0.846	Good
6	0.473	Good
7	0.199	Good
8	0.569	Good
9	0.323	Good
10	0.838	Good
11	0.342	Good
12	0.319	Good
13	0.412	Good
14	0.046	Bad
15	0.739	Good
16	0.669	Good
17	0.576	Good
18	0.691	Good

Table 10: Discriminating power of four-tier diagnostic test items

Results of student misconceptions identification on the static fluid chapter

The results of the analysis of student misconceptions in the static fluid chapter provide information on the number of students and the percentages who understand the concept (UC), partially understand it (PU), hold a misconception (M), and do not understand it (DUC). Table 11 below contains details of the misconceptions of 65 students regarding the chapter on static fluids from grades XI MIPA 2 and XI MIPA 3. This is because grade XI MIPA 1 students were used as participants to assess the instrument's empirical validity.

Competence Achievement Indicators (CAI)	Item	UC		PU		M		DUC	
		Total	%	Total	%	Total	%	Total	%
Density	5	41	63	17	26	7	11	-	-
Density	13	26	40	19	29	11	17	9	14
Density	15	17	26	38	58	5	8	5	8
Hydrostatic Pressure	1	43	66	17	26	3	5	2	3
Hydrostatic Pressure	4	17	26	44	68	2	3	2	3
Hydrostatic Pressure	12	4	6	8	12	46	71	7	11
Hydrostatic Pressure	14	6	9	16	25	37	57	6	9
Archimedes' Law	3	22	34	36	55	3	5	4	6
Archimedes' Law	9	6	9	39	60	16	25	4	6
Surface Tension	7	4	6	28	43	16	25	17	26
Surface Tension	10	9	14	10	15	43	66	3	5
Surface Tension	11	4	6	16	25	42	65	3	5
Capillarity	6	41	63	10	15	12	18	2	3
Capillarity	8	33	51	16	25	14	22	2	3
Viscosity	2	48	74	14	22	2	3	1	2
Viscosity	16	37	57	13	20	13	20	2	3
Viscosity	17	44	68	17	26	3	5	1	2

Table 11: Results of student misconceptions identification on the static fluid chapter

Based on Table 11, a separate graph can be prepared to further specify the level of student misconceptions. The graph of student misconceptions regarding the static fluid chapter can be shown in Figure 2.

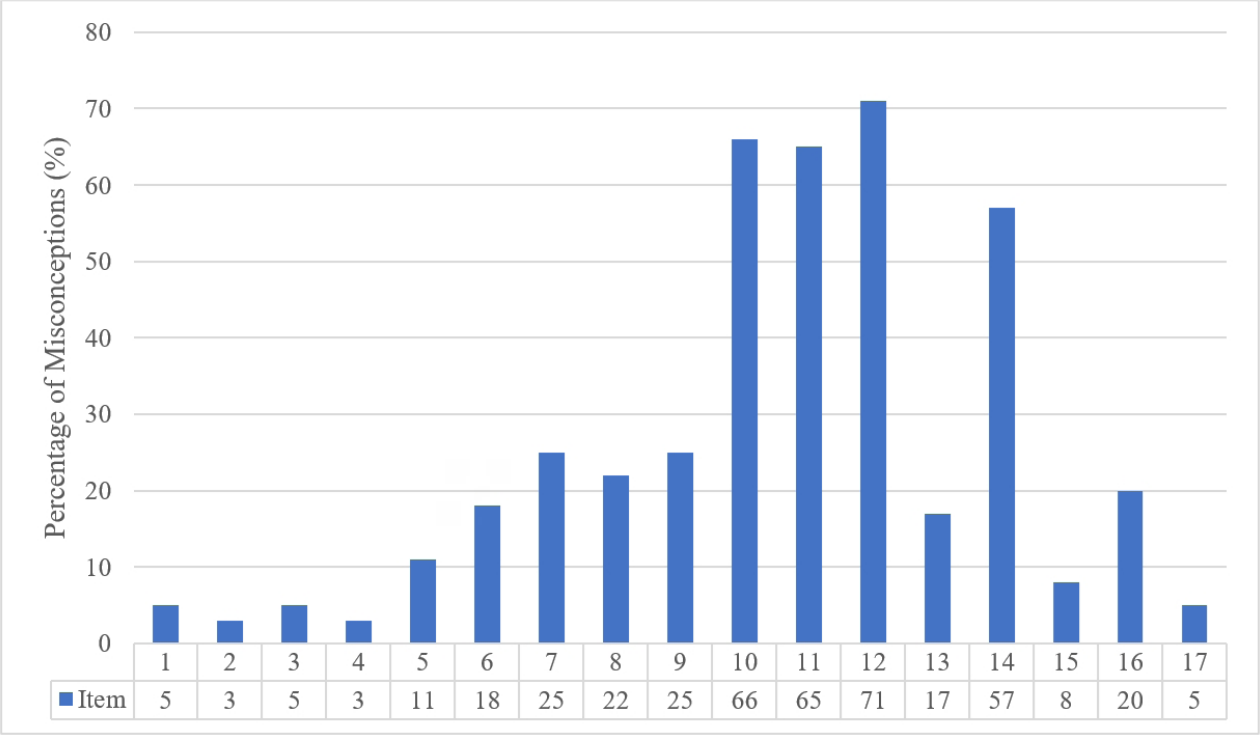

Figure 2: Percentage of student misconceptions on static fluids

Figure 2 shows that students had the most misconceptions on item 12, with 71% (46 students) exhibiting misconceptions. Question number 12 is an item related to the hydrostatic pressure sub-chapter. Apart from question 12, students had the most misconceptions in questions 10 and 11. These two questions are related to the sub-chapter on surface tension. Most students understand the concept in number 2, sub-chapter 'Viscosity', with 48 students (74% of the class). This means that most students' answers are correct and receive full points on question 2. Meanwhile, most students do not understand the concept in sub-chapter 7, surface tension: 17 students (26%). This means that 17 students who worked on number 7 scored 1.

DISCUSSION

Quality of four-tier diagnostic test instruments

The findings of this study demonstrate that the development and validation of the four-tier diagnostic test instrument, conducted through expert judgment and empirical testing, have successfully produced an assessment tool that generally meets the standards of diagnostic measurement. Expert validation ensured conceptual accuracy, while the subsequent Rasch-based analysis provided empirical evidence of item functioning. However, the detection of several items requiring revision or elimination indicates that diagnostic efficiency has not yet reached optimal levels. Inefficient items, such as item 5, which failed to meet the outfit MNSQ, ZTSD, and Pt Measure Correlation criteria simultaneously, burden the assessment process by generating noise rather than meaningful diagnostic information. This finding is consistent with Van Laar et al. (2018), who highlight that misfitting items reduce the efficiency of inferential decision-making and compromise researchers' responsibility to provide accurate diagnostic interpretations. The present study demonstrates that empirical validation using the Rasch model is not merely an analytical choice but an ethical responsibility to ensure that each item is fit for purpose in identifying misconceptions with precision and fairness.

In terms of reliability, both the limited and extensive trials produced person reliability, item reliability, and Cronbach's alpha values exceeding the minimum threshold of 0.60 recommended by Çebi and Reisoğlu (2023) and Sürücü and Maslakçı (2020), indicating that the instrument is sufficiently stable for classroom use. Reliability consistency indicates that the instrument can efficiently capture students' conceptual profiles without introducing unnecessary measurement error. However, the exceptionally high item reliability in the extensive trial (0.96) raises critical implications. Although high item reliability suggests strong internal coherence, it also indicates that the instrument may lack heterogeneity in item difficulty, thereby limiting the test's diagnostic responsibility to represent diverse conceptual challenges. This limitation suggests that future development must incorporate items targeting a broader spectrum of cognitive complexity to ensure that the assessment remains not only efficient in operation but also responsibly inclusive in representing the diversity of student reasoning patterns.

Regarding item difficulty and discriminating power, the Rasch analysis reveals that most items performed adequately. However, two items showed excessively great difficulty, and item 14 displayed weak discriminating power. Items that are too difficult reduce assessment efficiency because they fail to differentiate students at varying ability levels, thereby

producing data that are uninformative for instructional decision-making. This aligns with Saidi and Siew (2019) and Kim et al. (2016), who warn that poorly discriminating items jeopardize educators' diagnostic responsibility to accurately identify where and why misconceptions occur. The extremely great difficulty of items 11 and 6 may stem from indicators or stimuli misaligned with students' prior knowledge, while item 14's low discrimination suggests problems with wording clarity or unintended cues. These problematic items must therefore be revised to ensure that the instrument fulfils its responsibility to generate fair, transparent, and actionable diagnostic information that teachers can use in targeted instructional planning.

When positioned within the broader landscape of existing literature, the results of this study confirm the alignment of the developed instrument with modern Rasch-based measurement principles and its feasibility as a diagnostic tool in physics education, particularly for the static fluid chapter. Nevertheless, the strengths reported do not absolve researchers from the ongoing responsibility of iterative refinement. Continuous revision cycles comprising expert review, limited trials, Rasch calibration, and targeted item reconstruction are essential for maintaining diagnostic efficiency and accountability. Such cycles ensure that the instrument not only identifies misconceptions accurately but also supports teachers in designing interventions that are both instructionally effective and ethically responsible in addressing student learning needs. In practical classroom contexts, this diagnostic instrument can serve as a crucial evaluative tool for identifying misconceptions among grade XI students. Through its four-tier structure comprising content response, confidence rating, reasoning selection, and reasoning confidence, teachers can pinpoint misconceptions at a more granular level. This enables them to act responsibly by providing more in-depth, targeted instructional explanations. Teachers are encouraged to integrate four-tier diagnostic tests into formative assessment practices to increase assessment efficiency and improve instructional responsiveness. Future research should also investigate instructional designs tailored to specific misconception profiles, ensuring that responses to diagnostic findings are pedagogically responsible and evidence-based.

Additionally, comparative studies examining the diagnostic efficiency of four-tier tests versus traditional one-, two-, or three-tier tests are needed to determine the extent to which the added complexity of the four-tier structure truly enhances diagnostic accuracy. Such comparative analyses are essential for determining whether the increased effort required in developing and administering four-tier instruments is justified by correspondingly improved diagnostic outcomes, an important consideration in balancing educational efficiency and assessment responsibility. The Rasch analysis conducted in this study used the Winsteps application with a sample drawn from a single MAN in Bantul, which may limit generalizability. Furthermore, the instrument currently focuses solely on the static fluid chapter. Thus, future studies should expand the instrument to other physics domains to evaluate its broader applicability and to ensure responsible advancement of diagnostic tools across the physics curriculum.

Identify student misconceptions in the static fluids chapter

The findings of this study reveal a multidimensional pattern of students' conceptual mastery, partial comprehension, and deeply rooted misconceptions across key static fluid subtopics: hydrostatic pressure, density, surface tension, viscosity, and capillarity. These patterns reaffirm longstanding evidence in physics education that students frequently employ intuitive, experience-based reasoning rather than mechanistic scientific principles when interpreting fluid phenomena. However, the present findings extend earlier scholarship by demonstrating that students' confidence levels do not necessarily correlate with conceptual accuracy, raising critical questions about pedagogical responsibility in interpreting student confidence as a proxy for understanding. For example, the exceptionally high misconception rate in Question 12, where 71% of students incorrectly attributed the function of a hanging infusion bottle to capillary action, parallels the conceptual conflation documented by Li et al. (2020) and Liu and Li (2019). Yet this study adds nuance: many students expressed *high confidence in incorrect answers*, a signal that their reasoning frameworks are both robust and systematically flawed. This misalignment highlights a responsibility issue: teachers cannot rely on correctness or confidence alone as indicators of learning. Instead, instructional design must incorporate explicit mechanisms to probe underlying reasoning and prevent false certainty from propagating through subsequent learning experiences.

From an efficiency perspective, diagnostic instruments must therefore capture not only answer accuracy but also reasoning quality, enabling teachers to allocate instructional time more strategically to high-risk misconceptions. The correct scientific explanation of fluid flow driven by a pressure difference, as stated by Deiters and Kraska (2023) and Levy et al. (2021), further underscores the need for learning sequences that contrast similar phenomena while foregrounding the governing variables and mathematical relationships, thereby improving conceptual differentiation without increasing instructional load. A similar pattern emerges in Question 4, where students demonstrated partial understanding of hydrostatic pressure in real-world diving contexts. Consistent with Domínguez et al. (2022) and Raissi et al. (2020), students recalled formulas but failed to integrate density, depth, and gravitational acceleration into coherent physical reasoning. This indicates a *systemic pedagogical gap,* the insufficient bridging of theoretical representations and applied contexts, which compromises the efficiency of instruction, as time spent teaching formulas yields limited conceptual payoff. To increase instructional efficiency and uphold pedagogical accountability, educators must adopt simulation-based visualizations or hands-on experiments, enabling students to concretely experience pressure gradients rather than relying on symbolic abstractions.

Misconceptions about density (Question 13) continue this theme. With 28% of students demonstrating either partial or no understanding, the findings corroborate difficulties reported by Kiray and Simsek (2021) and Zenger and Bitzenbauer (2022). Errors such as predicting that low-

density objects would sink in high-density fluids reflect a failure to coordinate relative densities, indicating that rule-memorization approaches are pedagogically inefficient and ethically questionable when they produce superficial recall rather than conceptual insight. The strength of this diagnostic pattern lies in its ability to pinpoint specific misconceptions, allowing teachers to take targeted, responsible action. However, the instrument alone cannot reveal the cognitive origins of these misconceptions, necessitating qualitative follow-up research such as think-aloud protocols. Such approaches would enhance the efficiency and accountability of future diagnostic cycles by clarifying the reasoning errors that assessments cannot directly capture. Surface tension misconceptions (Questions 7 and 10) reveal similar instructional vulnerabilities. Students' beliefs that hot or soapy water has higher surface tension contradict molecular-level explanations documented by Nguyen et al. (2020), Tang et al. (2020), and Tsompou and Kocherbitov (2022).

These misconceptions highlight the lack of molecular macroscopic linkages in current teaching practices, a form of *curricular inefficiency* where students struggle to generalize concepts because instruction does not sufficiently develop explanatory depth. Integrating molecular dynamics simulations or inquiry-based experimentation is therefore both an efficient and responsible pedagogical alternative, providing high-yield conceptual clarity through experiential engagement rather than additional lecturing time. The findings for Question 9 on Archimedes' Principle echo patterns reported by Naylor and Tsai (2022) and Noxaïc and Fadel (2022): students memorize formulas but fail to internalize that the buoyant force equals the weight of the displaced fluid. Low confidence and fragmented conceptual frameworks indicate unstable knowledge structures that undermine learning progression. Addressing this issue requires repeated exposure to diverse problem variations and structured opportunities for students to articulate their reasoning, an efficient strategy for reinforcing conceptual stability across contexts while ensuring responsible instructional practice.

In contrast, the high conceptual performance in viscosity (Question 2) and capillarity (Question 6) offers essential insights into instructional efficiency. Consistent with findings by Kaptay (2021), Shafiei et al. (2023), and Gupta et al. (2023), these topics benefit from highly observable and manipulable phenomena, demonstrating that conceptual understanding improves when teaching practices integrate physical intuition, observation, and inquiry. These results reinforce the need for physics instruction that prioritizes experiential engagement over purely representational approaches, thereby supporting both efficient learning progression and equitable access to conceptual understanding for diverse learners. Taken together, the findings suggest that students understand fluid concepts more effectively when instruction is grounded in concrete, observable, and intuitive experiences. However, they struggle when concepts require abstraction or differentiation among superficially similar principles. This observation carries significant implications for curricular efficiency and teacher responsibility: physics education must integrate conceptual, experimental, and representational domains more coherently so that students can internalize distinctions among related fluid concepts without additional cognitive burden.

Methodologically, while the four-tier diagnostic test provides invaluable information for identifying misconceptions, its multiple-choice format limits insight into students' reasoning pathways. To enhance *assessment responsibility*, future research should complement diagnostic tests with interviews, classroom observations, or open-ended tasks that more explicitly reveal cognitive processes. Additional studies comparing the efficiency and diagnostic accuracy of four-tier tests against one-, two-, and three-tier formats would also clarify whether the extra layers of confidence judgment truly improve conceptual detection or increase assessment load without proportional benefit. Finally, the study's context is limited to a single MAN in Bantul, and the materials are restricted to static fluids, which limits generalizability. Responsible future research should expand sample diversity and extend the diagnostic instrument to additional physics topics to ensure broader applicability. Nevertheless, the current findings already serve as a powerful evaluative tool for teachers, enabling targeted intervention and more efficient allocation of instructional resources based on empirically verified patterns of student reasoning.

CONCLUSIONS

Three key conclusions can be drawn from the research results and discussion presented in this study. The first conclusion is that the four-tier diagnostic test instrument includes a grid of test questions, work instructions, an answer key, an answer sheet, and scoring guidelines. The test question structure consists of four levels, including questions with one answer key and three distractors, level of confidence in the answer, reason options, and level of confidence in the reason. The second conclusion is that the four-tier diagnostic test instrument developed has met validity standards, allowing it to be tested on research subjects. The results of the empirical validity analysis of trials limited to 18 items indicate that 17 items have validity, while one item falls into an invalid category. The reliability analysis results on the limited and extensive trials produced the same person reliability coefficient and Cronbach's alpha coefficients. However, the item reliability coefficient from the extensive trial was greater than that from the limited trial because one empirically invalid question item had been removed. Additionally, an analysis of the questions' difficulty levels revealed that two were very difficult, nine were difficult, and nine were moderate. Evaluation of the discriminating power of the questions revealed that 19 questions had good discriminating power, while 1 had poor discriminating power. The third conclusion, based on a four-tier diagnostic test instrument, is that students' misconceptions most often occur in the hydrostatic pressure sub-chapter (71%) and in surface tension (66%). The four-tier diagnostic test instrument has proven effective in categorizing students into four categories: understanding the concept, partially understanding, misconceptions, and not understanding the concept. This is confirmed by extensive trials, which show different values for each indicator across the six categories.

REFERENCES

Aiken, L. R. (1985) 'Three coefficients for analyzing the reliability and validity of ratings', *Educational and Psychological Measurement,* Vol. 45, No. 1, pp. 131–142. https://doi.org/10.1177/0013164485451012

Asrizal, A., Zan, A. M., Mardian, V. and Festiyed, F. (2022) 'The impact of static fluid e-module by integrating STEM on learning outcomes of students', *Journal of Education Technology*, Vol. 6, No. 1, pp. 110–118. https://doi.org/10.23887/jet.v6i1.42458

Assem, H. D., Nartey, L., Appiah, E., and Aidoo, J. K. (2023) 'A review of students' academic performance in physics: attitude, instructional methods, misconceptions, and teachers' qualification', *European Journal of Education and Pedagogy*, Vol. 4, No. 1, pp. 84–92. https://doi.org/10.24018/ejedu.2023.4.1.551

Bichi, A. A. and Talib, R. (2018) 'Item response theory: An introduction to latent trait models to test and item development', *International Journal of Evaluation and Research in Education*, Vol. 7, No. 2, pp. 142–151. https://doi.org/10.11591/ijere.v7i2.12900

Cai, S., Liu, C., Wang, T., Liu, E. and Liang, J. C. (2021) 'Effects of learning physics using Augmented Reality on students' self-efficacy and conceptions of learning', *British Journal of Educational Technology*, Vol. 52, No. 1, pp. 235–251. https://doi.org/10.1111/bjet.13020

Caleon, I. S. and Subramaniam, R. (2010) 'Do students know what they know and don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions', *Research in Science Education*, Vol. 40, No. 3, pp. 313–337. https://doi.org/10.1007/s11165-009-9122-4

Çebi, A. and Reisoğlu, İ. (2023) 'Adaptation of self-assessment instrument for educators' digital competence into Turkish culture: A study on reliability and validity', *Technology, Knowledge and Learning*, Vol. 28, No. 2, pp. 569–583. https://doi.org/10.1007/s10758-021-09589-0

Çelikkanlı, N. Ö. and Kızılcık, H. Ş. (2022) 'A review of studies about four-tier diagnostic tests in physics education', *Journal of Turkish Science Education*, Vol. 19, No. 4, pp. 1291–1311. https://doi.org/10.36681/tused.2022.175

Dack, H. (2019) 'Understanding teacher candidate misconceptions and concerns about differentiated instruction', *The Teacher Educator*, Vol. 54, No. 1, pp. 22–45. https://doi.org/10.1080/08878730.2018.1485802

Deiters, U. K. and Kraska, T. (2023) *High-Pressure Fluid Phase Equilibria: Phenomenology and Computation*, Amsterdam: Elsevier.

Domínguez, J. M., Fourtakas, G., Altomare, C., Canelas, R. B., Tafuni, A., García-Feal, O. et al. (2022) 'DualSPHysics: From fluid dynamics to multiphysics problems', *Computational Particle Mechanics*, Vol. 9, No. 5, pp. 867–895. https://doi.org/10.1007/s40571-021-00404-2

Gupta, S., Chatterjee, S. and Chanda, A. (2023) 'Influence of vertically treaded outsoles on interfacial fluid pressure, mass flow rate, and shoe-floor traction during slips', *Fluids,* Vol. 8, No. 3, pp. 82–90. https://doi.org/10.3390/fluids8030082

Hasanah, U. (2020) 'The effectiveness of STEM education for overcoming students' misconceptions in high school physics: Engineering viewpoint', *Science Education International*, Vol. 31, No. 1, pp. 5–13. https://doi.org/10.33828/sei.v31.i1.1

Hill, H. C. and Chin, M. (2018) 'Connections between teachers' knowledge of students, instruction, and achievement outcomes', *American Educational Research Journal*, Vol. 55, No. 5, pp. 1076–1112. https://doi.org/10.3102/0002831218769614

Hong, Y. S., Han, C. P., and Cho, S. S. (2021) 'Level-based learning algorithm based on the difficulty level of the test problem', *Applied Sciences*, Vol. 11, No. 10, pp. 4380–4393. https://doi.org/10.3390/app11104380

Iqbal, Q. and Malzahn, D. (2017) 'Evaluating discriminating power of single-criteria and multi-criteria models towards inventory classification', *Computers & Industrial Engineering,* Vol. 104, No. 1, pp. 219–223. https://doi.org/10.1016/j.cie.2016.12.007

Istiyono, E., Dwandaru, W. S. B., Fenditasari, K., Ayub, M. R. S. S. N., and Saepuzaman, D. (2023) 'The development of a four-tier diagnostic test based on modern test theory in physics education', *European Journal of Educational Research*, Vol. 12, No. 1, pp. 371–385. https://doi.org/10.12973/eu-jer.12.1.371

Kaptay, G. (2021) 'A unified equation for the viscosity of pure liquid metals', *International Journal of Materials Research*, Vol. 96, No. 1, pp. 24–31. https://doi.org/10.3139/146.018080

Kim, Y. J., Almond, R. G. and Shute, V. J. (2016) 'Applying evidence-centered design for the development of game-based assessments in physics playground', *International Journal of Testing*, Vol. 16, No. 2, pp. 142–163. https://doi.org/10.1080/15305058.2015.1108322

Kiray, S. A. and Simsek, S. (2021) 'Determination and evaluation of the science teacher candidates' misconceptions about density by using the four-tier diagnostic test', *International Journal of Science and Mathematics Education*, Vol. 19, no. 5, pp. 935–955. https://doi.org/10.1007/s10763-020-10087-5

Kösem, Ş. D. and Özdemir, Ö. F. (2014) 'The nature and role of thought experiments in solving conceptual physics problems', *Science & Education*, Vol. 23, No. 1, pp. 865–895. https://doi.org/10.1007/s11191-013-9635-0

Kundu, P. K., Cohen, I. M. and Dowling, D. R. (2015) *Fluid Mechanics*, 6th ed., San Diego, CA: Academic Press.

Laliyo, L. A. R., Hamdi, S., Pikoli, M., Abdullah, R. and Panigoro, C. (2021) 'Implementation of four-tier multiple-choice instruments based on the partial credit model in evaluating students' learning progress', *European Journal of Educational Research*, Vol. 10, No. 2, pp. 825–840. https://doi.org/10.12973/eu-jer.10.2.825

Lee, J. E., Recker, M. and Yuan, M. (2020) 'The validity and instructional value of a rubric for evaluating online course quality: An empirical study', *Online Learning*, Vol. 24, No. 1, pp. 245–256. https://doi.org/10.24059/olj.v24i1.1949

Levy, R., Okun, Z. and Shpigelman, A. (2021) 'High-pressure homogenization: Principles and applications beyond microbial inactivation', *Food Engineering Reviews*, Vol. 13, pp. 490–508. https://doi.org/10.1007/s12393-020-09239-8

Li, Y., Luo, H., Li, H., Chen, S., Jiang, X. and Li, J. (2020) 'Dynamic capillarity during displacement process in fractured tight reservoirs with multiple fluid viscosities', *Energy Science & Engineering*, Vol. 8, No. 2, pp. 300–311. https://doi.org/10.1002/ese3.558

Lightfoot, C. J., Wilkinson, T. J., Memory, K. E., Palmer, J., and Smith, A. C. (2021) 'Reliability and validity of the patient activation measure in kidney disease: results of Rasch analysis', *Clinical Journal of the American Society of Nephrology*, Vol. 16, No. 6, pp. 880–888. https://doi.org/10.2215/CJN.19611220

Liu, J. and Li, S. (2019) 'Capillarity-driven migration of small objects: A critical review', *The European Physical Journal E*, Vol. 42, No. 1, pp. 1–23. https://doi.org/10.1140/epje/i2019-11759-1

Lourdusamy, R. and Magendiran, P. (2021) 'A systematic analysis of difficulty level of the question paper using students' marks: a case study', *International Journal of Information Technology*, Vol. 13, No. 3, pp. 1127–1143. https://doi.org/10.1007/s41870-020-00599-2

Machts, N., Kaiser, J., Schmidt, F. T. and Moeller, J. (2016) 'Accuracy of teachers' judgments of students' cognitive abilities: A meta-analysis', *Educational Research Review*, Vol. 19, pp. 85–103. https://doi.org/10.1016/j.edurev.2016.06.003

Madina, I. S., Ratnawulan, R., Sari, S. Y., Sundari, P. D., and Mayasari, T. (2022) 'Identification of students' misconceptions about straight motion materials using a four-tier diagnostic test and certainty of response index (CRI)', *Pillar of Physics Education*, Vol. 15, No. 1, pp. 21–30. https://doi.org/10.24036/12658171074

Maison, D., Astalini, D. A. K. and Sumaryanti, R. P. (2020) 'Supporting assessment in education: E-assessment interest in physics', *Universal Journal of Educational Research*, Vol. 8, No. 1, pp. 89–97. https://doi.org/10.13189/ujer.2020.080110

Maknun, J. (2020) 'Implementation of a guided inquiry learning model to improve understanding of physics concepts and critical thinking skills of vocational high school students', *International Education Studies*, Vol. 13, No. 6, pp. 117–130. https://doi.org/10.5539/ies.v13n6p117

Mason, R. A. and Just, M. A. (2016) 'Neural representations of physics concepts', *Psychological Science*, Vol. 27, No. 6, pp. 904–913. https://doi.org/10.1177/0956797616641941

Naylor, D. and Tsai, S. S. (2022) 'Archimedes' principle with surface tension effects in undergraduate fluid mechanics', *International Journal of Mechanical Engineering Education*, Vol. 50, No. 3, pp. 749–763. https://doi.org/10.1177/03064190211055431

Neito, R., Vilhunen, E., Lavonen, J. and Reivelt, K. (2025) 'Students' situational interest and perceived relevance during designed coherent physics learning modules', *International Journal of Science Education*, Vol. 2, No. 1, pp. 1–24. https://doi.org/10.1080/09500693.2025.2488400

Nguyen, H. N. G., Zhao, C. F., Millet, O. and Gagneux, G. (2020) 'An original method for measuring liquid surface tension from capillary bridges between two equal-sized spherical particles', *Powder Technology,* Vol. 363, No. 1, pp. 349–359. https://doi.org/10.1016/j.powtec.2019.12.049

Noxaïc, A. L. and Fadel, K. (2022) 'How to use the Archimedes Paradox for educational purposes', *The Physics Teacher*, Vol. 60, No. 2, pp. 137–139. https://doi.org/10.1119/10.0009424

Ozkan, G. and Selcuk, G. S. (2015) 'The effectiveness of conceptual change texts and context-based learning on students' conceptual achievement', *Journal of Baltic Science Education*, Vol. 14, No. 6, 753–764. https://doi.org/10.33225/jbse/15.14.753

Putra, A. S. U., Hamidah, I. and Nahadi. (2020) 'The development of five-tier diagnostic test to identify misconceptions and causes of students' misconceptions in waves and optics materials', *Journal of Physics: Conference Series*, Vol. 1521, No. 2, pp. 1–9. https://doi.org/10.1088/1742-6596/1521/2/022020

Rabin, L. A., Krishnan, A., Bergdoll, R. and Fogel, J. (2021) 'Correlates of exam performance in an introductory statistics course: Basic math skills along with self-reported psychological/behavioral and demographic variables', *Statistics Education Research Journal,* Vol. 20, No. 1. https://doi.org/10.52041/serj.v20i1.97

Raissi, M., Yazdani, A. and Karniadakis, G. E. (2020) 'Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations', *Science*, Vol. 367, No. 6481, pp. 1026–1030. https://doi.org/10.1126/science.aaw4741

Robbins, M. E., DiQuattro, G. J. and Burkholder, E. W. (2025) 'Assessment of expert decisions in graduate quantum mechanics', *Physical Review Physics Education Research*, Vol. 21, No. 1, pp. 1–10. https://doi.org/10.1103/PhysRevPhysEducRes.21.010125

Saidi, S. S. and Siew, N. M. (2019) 'Reliability and validity analysis of statistical reasoning test survey instrument using the Rasch measurement model', *International Electronic Journal of Mathematics Education,* Vol. 14, No. 3, pp. 535–546. https://doi.org/10.29333/iejme/5755

Saricayir, H., Ay, S., Comek, A., Cansiz, G. and Uce, M. (2016) 'Determining students' conceptual understanding level of thermodynamics', *Journal of Education and Training Studies*, Vol. 4, No. 6, pp. 69–79. https://doi.org/10.11114/jets.v4i6.1421

Setiawan, A., Mardapi, D. and Andrian, D. (2019) 'The development of an instrument for assessing students' affective domain using self-and peer-assessment models', *International Journal of Instruction,* Vol. 12, No. 3, pp. 425–438. https://doi.org/10.29333/iji.2019.12326a

Shafiei, M., Kazemzadeh, Y., Martyushev, D. A., Dai, Z. and Riazi, M. (2023) 'Effect of chemicals on the phase and viscosity behavior of water in oil emulsions', *Scientific Reports*, Vol. 13, No. 1, pp. 1–14. https://doi.org/10.1038/s41598-023-31379-0

Suprapto, N. (2020) 'Do we experience misconceptions?: An ontological review of misconceptions in Science', *Studies in Philosophy of Science and Education*, Vol. 1, No. 2, pp. 50–55. https://doi.org/10.46627/sipose.v1i2.24

Sürücü, L. and Maslakçi, A. (2020) 'Validity and reliability in quantitative research', *Business & Management Studies: An International Journal*, Vol. 8, No. 3, pp. 2694–2726. https://doi.org/10.15295/bmij.v8i3.1540

Tang, Z., Fang, K., Bukhari, M. N., Song, Y. and Zhang, K. (2020) 'Effects of viscosity and surface tension of a reactive dye ink on droplet formation', *Langmuir*, Vol. 36, No. 32, pp. 9481–9488. https://doi.org/10.1021/acs.langmuir.0c01392

Taslidere, E. (2016) 'Development and use of a three-tier diagnostic test to assess high school students' misconceptions about the photoelectric effect', *Research in Science & Technological Education,* Vol. 34, No. 2, pp. 164–186. https://doi.org/10.1080/02635143.2015.1124409

Tsompou, A. and Kocherbitov, V. (2022) 'The effects of water purity on the removal of hydrophobic substances from solid surfaces without surfactants', *Journal of Colloid and Interface Science*, Vol. 608, No. 1, pp. 1929–1941. https://doi.org/10.1016/j.jcis.2021.10.040

Türkoguz, S. (2020) 'Comparison of threshold values of three-tier diagnostic and multiple-choice tests based on response time', *Anatolian Journal of Education*, Vol. 5, No. 2, pp. 19–36. https://doi.org/10.29333/aje.2020.522a

Van den Akker, J., Gravemeijer, K., McKenney, S. and Nieveen, N. (2006) *Educational Design Research*, London: Routledge.

Van Laar, E., van Deursen, A. J., van Dijk, J. A. and de Haan, J. (2018) '21st-century digital skills instrument aimed at working professionals: Conceptual development and empirical validation', *Telematics and Informatics*, Vol. 35, No. 8, pp. 2184–2200. https://doi.org/10.1016/j.tele.2018.08.006

Vaske, J. J., Beaman, J. and Sponarski, C. C. (2017) 'Rethinking internal consistency in Cronbach's Alpha', *Leisure Sciences*, Vol. 39, No. 2, pp. 163–173. https://doi.org/10.1080/01490400.2015.1127189

Wahyuni, N., Bhakti, Y. B., Mutakin, T. Z. and Astuti, I. A. D. (2021) 'The development of a four-tier diagnostic test instrument to identify the learners' misconception of circular motions', *Impulse: Journal of Research and Innovation in Physics Education,* Vol. 1, No. 1, pp. 24–31. https://doi.org/10.14421/impulse.2021.11-03

Yuberti, Y., Suryani, Y. and Kurniawati, I. (2020) 'Four-tier diagnostic test with the certainty of response index to identify misconceptions in physics', *Indonesian Journal of Science and Mathematics Education*, Vol. 3, No. 2, pp. 245–253. https://doi.org/10.24042/ijsme.v3i2.6061

Zenger, T. and Bitzenbauer, P. (2022) 'Exploring German secondary school students' conceptual knowledge of density', *Science Education International*, Vol. 33, No. 1, pp. 86–92. https://doi.org/10.33828/sei.v33.i1.9