Pegem Journal of Education and Instruction, Vol. 14, No. 3 , 2024 (pp. 361-373 )

RESEARCH ARTICLE WWW.PEGEGOG.NET

# Development of an Online Two-Tier Test to Explore Students' Conceptions on Objects in Static Fluid

**Sarintan N. Kaharu[1*]. Yusdin Gagaramusu[2], Azizah Azizah[3], Nurul Kamisani[4], Nurgan Tadeko[5], Jusman Mansyur[6]**

[1-6]Tadulako University, Indonesia

## Abstract

This study aims to develop an online test that can be used to characterize students' conception of objects in static fluid. The test developed by carried out several stages, namely: modified previous essay test and formed the Tier-1 of multiple-choice questions with options taken from the findings previous studies; combined the reasons for choosing options in Tryout-1 and the findings of previous studies for forming Tier-2; expert validation process; refinement process, readability test tryout with revisions; data collection involving 318 pre-service elementary school teachers to characterize the online test based on item response theory; and revised some items. Rasch analysis was implemented to characterized the test and subjects. The developed test consists of 23 items in two-tier covering the context of floating, sinking, and suspending objects. The results of validation showed that in the development process, the test has very good for item validity. Rasch analysis showed that, overall, the test functioned well with its item reliability is very good category. Based on the findings and diificulty level, the test can be used for middle school students up to college level with some notion. Some limitations of the test are presented in this paper.

**Keywords:** mental model, online two-tier test, representation, static fluid.

## Introduction

Understanding floating, suspending, and sinking is the starting point for developing insights into very basic static fluid phenomena, such as buoyancy and its applications covered by Archimedes' law. Without a grasp of the concepts underlying these basic phenomena, students may face obstacles when attempting to understand advanced fluid concepts.

However, research results demonstrate that grasping basic concepts in physics is difficult and challenging not only for elementary schools but also for high school and physics students (Mansyur, Werdhiana, Darsikin, Kaharu & Tadeko, 2022a; 2022b). This difficulty can be attributed to the fact that children develop initial theories about the physical world based on their everyday experiences in an early age (Shen, Liu, & Chang, 2017; Minogue & Borland, 2016). For instance, they often assume that small or light objects float and large or heavy objects sink, as observed in their daily lives (Yin, Tomita, & Shavelson, 2008). However, turning these initially naive ideas into scientifically accurate explanations is a slow and challenging process, leading to the formation of misconceptions during the learning process (Gette, Kryjevskaia, Stetzer, & Heron, 2018). As a result, students may have a hybrid understanding that combines elements of their naive initial ideas with mental models in the scientific category (Vosniadou & Brewer, 1992). This highlights the importance of identifying and addressing misconceptions in physics education, as they can hinder the learning and understanding of more advanced physics concepts.

The topic of floating and sinking is a subject that has been extensively studied by researchers. For example, Chien, Hsiung, & Chen (2009) applied Vygotsky's theory to investigate the development of thinking behavior in 5-year-old children related to the phenomena of floating and sinking objects. Teo, Yan, & Ong (2017) explored the conceptions and ideas of 6-year-old children about floating and sinking. They found that there were ideas of children that were not found in previous studies, namely that objects sink because water is soft. Hsin & Wu (2011) applied the scaffolding approach

**Corresponding Author e-mail:** sarintankaharu@untad.ac.id
**https://orcid.org/0000-0002-2069-0921**

**How to cite this article:** Kaharu SN. Gagaramusu Y, Azizah A, Kamisani N, Tadeko N, Mansyur J (2024). Development of an Online Two-Tier Test to Explore Students' Conceptions on Objects in Static Fluid. Pegem Journal of Education and Instruction, Vol. 14, No. 3, 2024, 361-373

**Source of support:** Nil.

**Conflict of interest:** Nil.

**DOI:** 10.47750/pegegog.14.03.34

**Accepted:** 22.10.2023 **Received:** 26.08.2023

**Publised :** 01.07.2024

in learning and studied its effect on students' conceptions of floating and sinking.

Minogue, Borland, Russo, Chen, & Grady (2015) investigated pre-service elementary school teachers' conceptions regarding the integration of buoyancy sub-concepts and emphasized the relationship between objects and the surrounding fluid. They found that students who received haptic feedback were more likely to use "haptically-based" terms such as mass, gravity, buoyant force, and pushing. The research recommends the need to build a local theory of haptic cognition mediated by language. Castillo, Waltzer, & Kloos (2017) investigated a series of hands-on experiences by students related to factors that cause objects to sink faster. They found that the errors that occurred in the activity were temporary compared to students who used static images. Gette et al. (2018) designed instructional activities to reduce intuitive reasoning and improve student understanding. This research found that instructional modifications designed to eliminate strong intuitive attraction resulted in significantly better performance on the concept of buoyant force. Djudin's research (2021) applied the 3-2-1 reading strategy integrated with refutation text to reduce misconceptions about buoyancy. The research concluded that significant conceptual change occurred in a number of students as a result of the treatment.

In addition to research that focuses on exploring conceptions, misconceptions, and instructional design in physics education, there are also studies that specifically aim to develop assessment instruments to better understand students' understanding of floating and sinking. For instance, Yin et al. (2008) developed a test to probe students' misconceptions related to floating and sinking. Viyanti, Cari, Sunarno, & Prastyo (2017) designed a rubric to evaluate high school students' ability to construct evidence-based arguments about the nature of floating and sinking. Kafiyani, Samsudin, & Saepuzaman (2019) similarly developed a four-tier diagnostic test to identify and categorize high school students' mental models of static fluids. These studies provide valuable insights into the effectiveness of different assessment tools and can inform the development of more targeted instructional strategies to enhance students' understanding of buoyancy and other physics concepts.

The above description illustrates that previous researches have predominantly focused on the context of floating and sinking phenomena, including exploring conceptions, misconceptions, instructional design, and assessment development. An extensive search for research articles specifically on the suspending phenomenon yielded no results. It is possible that the suspending phenomenon is not a focus for researchers because it is rare to find in everyday life. The concept of suspending is generally briefly discussed in textbooks (Kaharu & Mansyur, 2021; Mansyur et al., 2022a; 2022b) by examining a similarity of the density of the object and the density of water. Usually, the discussion is supported by diagrams of the position of the suspending object accompanied by force diagrams at work. We never find in textbooks where the position of the object is depicted as not in the middle of the depth. Perhaps the author's intention was to differentiate the position for the other two conditions, namely floating and sinking. This habit of representation is suspected to have an impact on the conceptions built by learners that a suspending object 'must be' in the middle of the depth.

These assumptions prompted the research that began with developing a test to explore learners' mental models and representation patterns about suspending objects (Kaharu & Mansyur, 2021). The research developed an essay-type test with 30 items. The suspending phenomenon is the dominant aspect that can be explored through this test, while still considering the floating and sinking. The test has been applied in Mansyur et al.'s (2022a) research, and mental models of elementary, junior high, high school students, and pre-service physics teachers regarding the floating, suspending, and sinking phenomena were identified. The test was also used in Mansyur et al.'s (2022b) research for exploring external representation pattern of students. The research found a consistent representation pattern associated with learners' mental models of the position of floating objects, namely that they are located in the middle of the height/depth of the water. The research team encountered difficulties in using the test in both studies. Despite requiring short answers, an essay-type test with many items based on paper and pencil made it difficult for the researchers to tabulate and analyze data because everything was still done manually. Based on these experiences, this study overcomes the shortcomings of existing test, especially in terms of ease and practicality, by developing a test that can be used online by adapting a previous test into an online two-tier test (O2Tt).

## Literature Review

A complete understanding of the structure that causes objects to float or sink requires nontrivial knowledge, which includes analyzing the relationship between buoyant force and gravitational force. Textbook authors and teachers often hide behind the concept of 'relative density' when explaining this phenomenon. Similarly, the approach commonly used is limited to demonstrating buoyant force as presented in textbooks (Karmilof-Smith, 1990).

The research findings indicate that students enter the classroom with preconceived ideas about science influenced

by their previous experiences, textbooks, teacher explanations, or everyday language (Chen, Bao, Fritchman, & Ma, 2021). According to the constructivist view, students commonly generate their own understanding and hypotheses regarding the mechanics of the natural world. As a consequence, their process of knowledge or theory formation occasionally contradicts the established theories comprehended by scientists (Hooda & Devi, 2018). For instance, when examining children's interpretations regarding the phenomena of objects either floating or sinking when immersed in water, insights gleaned from interviews with students in elementary and secondary school settings (Havu-Nuutinen, 2005) indicated that students' interpretations frequently centered on a singular dimension. They would make reference to attributes such as the object's mass (light objects would float), its size (larger objects would sink), or its configuration (objects with holes would sink). Moreover, in their explanations, air was perceived as an active force exerting an upward pull on the object, whereas water was viewed as a force drawing it downward. These explanations pertaining to buoyancy and sinking cannot be equated to scientific explanations (which are rooted in concepts of density and buoyant force) because instead of considering the interplay between the object and the encompassing fluid, students fixate on the object's individual property.

In general, the process of reshaping initial understandings to establish accurate scientific concepts has been delineated as conceptual change, as described by Karmiloff-Smith (1990). To illustrate, when applying the notion of density in the context of objects floating or sinking, further shifts in conceptualization are essential. To effectively grasp concepts like comparing the density of both the object and the fluid, as well as comparing the forces of gravity and buoyancy, simultaneous consideration and integration of these ideas are required. Frequently, the subjects of density and buoyancy are introduced primarily during middle school, with the rationale that students should possess the capability to comprehend the formal aspects of pertinent formulas, such as proportions involving quantity comparisons. However, even among middle school students, there are those who retain the same intuitive conceptions, as highlighted by Havu-Nuutinen (2005). Research conducted by Tao et al. (2011) demonstrated parallels between preschoolers and adults in terms of their inadequate strategies for evaluating the buoyancy of objects. The majority of curricula tailored for middle school concentrate on instigating conceptual change concerning the disparities in students' conceptualizations of materials. As an illustration, the utilization of a box matrix containing dots to visually depict the density of diverse materials within instructional unit aids students in distinguishing between weight and density, facilitating an integrated comprehension of density, as noted by Havu-Nuutinen (2005).

Research in cognitive science proves that when using a single representation format (verbal, symbolic or graphical) that requires quantitative problem-solving, students tend to think in terms of pattern matching with equation forms rather than using a qualitative approach that provides deeper interpretation and understanding (Canlas, 2019). Similarly, when deciphering a depiction (whether graphical or purely symbolic), their tendency is to emphasize surface-level characteristics rather than extracting physical information in an optimal manner. It is conceivable that the capacity, or lack thereof, to effectively participate in the task of linking and transposing information while working with various representations might be influenced by cognitive structures that encompass intuitive elements.

Various approaches have been employed to investigate learners and their engagement with science. For instance, research has involved documenting, categorizing, and observing the science and mathematics-related endeavors of children aged 3 to 5 years (van Schijdel, van van Es, Franse, van Bers, & Raijmakers, 2018). Their findings indicated that these activities often encompass a sequence of science-based tasks primarily aimed at fostering investigative and observational skills while promoting critical thinking. In another study (Minogue & Borland, 2016), it was revealed that children are provided with opportunities to learn about subjects like water and air. Interestingly, children tend to share a common conceptual grasp of weight, size, and material in relation to buoyancy. Specifically, their understanding often aligns with the idea that buoyancy is linked to conditions of lightness, smaller dimensions, and the specific materials involved, such as wood.

When delving into the realm of floating and sinking, the enhancement of conceptual comprehension can be achieved through various instructional methods, each rooted in distinct approaches. An intervention study conducted by Smith, Maclin, Grosslight, & Davis (1997) encompassing pretest and posttest evaluations of student comprehension, discovered that students broadened their grasp of floating and sinking concepts in a more comprehensive manner when exposed to the "density approach." This approach, characterized by its focus on density-related explanations for floating and sinking, facilitated the eliminating of the 'weight-based model' previously held by students. Consequently, students began to view buoyancy as a phenomenon influenced by a multitude of physical properties, as elucidated by Chien et al. (2009).

In order to effectively transform concepts into scientific mental frameworks, learners must actively participate in a

process known as knowledge integration (Shen et al., 2017). This process entails linking newly developed scientific concepts with their existing understandings, refining or discarding the latter as necessary. A pivotal prerequisite for achieving conceptual change is learners' dissatisfaction with their current conceptions, prompting a desire for new explanatory frameworks. Central to this notion is the introduction of cognitive conflict, which acts as the catalyst for initiating conceptual change by challenging the initial notions a student might hold. For instance, to counter the misconception that all buoyant objects inherently contain air, an instructor might present students with objects that possess hollow spaces (and therefore contain air), yet sink, thereby sparking cognitive conflict and prompting the need for a revised understanding.

The aspects explained above require instrument support in the form of test for their exploration process. The criteria for the test are not only about the coverage of the context but also related to practicality of use and ease of analyzing output data. For this purpose, adaptation was made to a previous offline-based test (Kaharu & Mansyur, 2021) to an online format. The external representation patterns and mental models (Mansyur et al., 2022a; 2022b) found through the use of the test became the basis for developing a test in the form of two-tier test.

Two-tier test is a multiple-choice instrument consisting of items with two sub-items (tier/level). The first tier contains content questions and the second tier asks respondents to choose the best option that represents their thoughts/reasoning on the first tier. The advantage of O2Tt is that it is easy to administer like other multiple-choice instruments. Another advantage is that the instrument facilitates researchers/educators in understanding the respondent's reasoning (Ivanjek, Morris, Schubatzky, Hopf, Burde, Haagen-schützenhöfer, et al., 2021). Although there are criticisms of the two-tier test, the use of reasoning questions when answering multiple-choice instruments can be a sensitive and effective way to assess meaningful learning, while also overcoming the weaknesses of traditional multiple-choice instruments. There are criticisms related to the second-tier questions that are considered biased because the answer options have been predetermined. However, the two-tier test has been used in various fields such as biology education, chemistry education, and physics education.

As previously explained, an instrument has been developed to access mental models and representation patterns (Kaharu & Mansyur, 2021). However, the test has a weakness because it is difficult to administer. Thus, this study aims to cover this deficiency and develop O2Tt that can be widely used. The use of online platforms also supports the expansion of the use and administration of the instrument as respondents' answers are sent in real-time to the researcher's email/account.

# Method

## Development Process

This research focuses on the development of a test in the form of O2Tt by utilizing the findings and recommendations of previous studies. The goal is to obtain an online test on the mental model and external representation pattern of object context in static fluids. The development of O2Tt is based on the previous essay test (Kaharu & Mansyur, 2021) that was used offline and is considered as Version-0. Version-0 was then modified into an online Version-1 in the form of multiple-choice questions with options taken from the findings of Mansyur et al. (2022a; 2022b) and other previous studies. The development of Version-0 and Version-1 involved fifth and sixth-grade elementary school, junior high school, high school students, as well as first to fourth-year physics education students. In Version-1, respondents were asked to provide reasons for their choices. The respondent's choices in the Tryout-1 session become Tier-1, and the combination of the reasons for choosing options in Tryout-1 and the findings of previous studies (Mansyur et al., 2022a; 2022b) and other previous studies become Tier-2 in the online two-tier multiple-choice (Version-2). Tryout-1 involved 97 pre-service physics teachers. The respondents' reasons for choosing options in Tryout-1 were selected to enrich the Tier-2 options. Version-2 instrument went through an expert validation process.

Version-2 also went through a refinement process, which was a readability test (Tryout-2) with revisions based on feedback from tryout respondents. The readability test involved 12 students. This process resulted in Version-3 instrument.

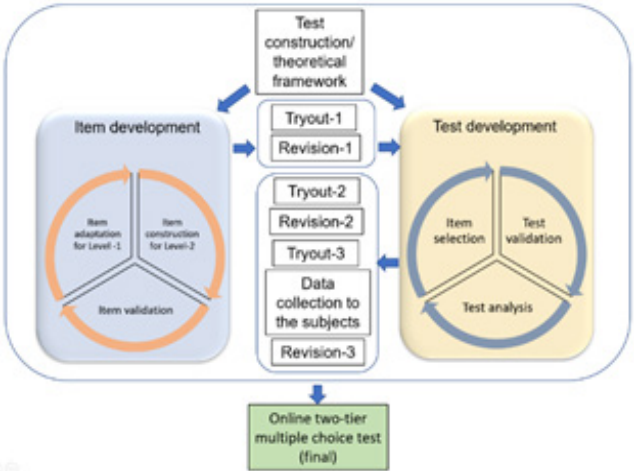


**Fig. 1:** Diagram of development process

**Table 1:** Four groups of the items

| *Aspect* | *Description of Item* | *Item* |
|---|---|---|
| Object position in liquid | Student distinguishes the position of floating, sinking and suspending objects. | 1, 2, 3 |
| | Student identifies the category of 3 floating objects based on their position in the liquid. | 4 |
| | Student identifies the category of 3 objects that are located between the bottom of the container and the surface of the liquid, related to density. | 5 |
| Object condition in liquid due to certain treatment | Student confirms the position of the object and the statement of the object's condition in the liquid. | 13 |
| | Student determines the object's condition in the liquid when a hole, hollow, hollow filled with air or cavity filled with water is made on the floating, sinking or suspending object | 6, 7, 8, 9, 14, 15, 16, 17, 18 |
| | Student determines the effect of adding water to the container on the properties and position of the object. | 23 |
| Object representation | Student identifies the category of object representation related to properties based on density. | 10, 11, 12, 13 |
| Effect of object volume | Student determines the properties and representation of the object pieces related to mass, volume, and/or density. | 19, 20, 21, 22 |

To obtain a description of the instrument characteristics based on item response theory (IRT), data collection was carried out (Tryout-3) involving 318 pre-service elementary school teachers. The results of the instrument characteristics data analysis were then used to revise some items, especially replacing options that did not work well. The results of this revision provide the Final Version instrument (O2Tt). In general, the test development process is presented in Figure 1. O2Tt consists of 23 two-tier item-items. These items are grouped into four that represent the four aspects of the basic concept of static fluids, particularly the concepts of floating, buoyancy, and sinking. The grouping of items is presented in Table 1.

## Data Collection

Data collection included some steps of the test development, namely: Tryout-1, Tryout-2, Validation, and Tryout-3 as explained in the development process.

## Content and Construct Validity

The Version-2 instrument obtained through several subsequent stages was further validated. The validity of the test was determined through expert validation involving four physics education lecturers. Two evaluation experts validated the construct (face validity) and content (content validity) aspects for both levels based on the principles of multiple-choice item construction. Two experts in media/computer-based systems/websites validated the online test aspects, including design and application, web devices, semantic web, respondent accessibility, online platform flexibility, ease of data extraction from the system output, and considerations for data analysis ease. All validators provided comments and suggestions regarding aspects that did not meet the validity criteria. The comments and suggestions from the validators served as the basis for revisions.

## Data Analysis

Prior to conducting the data analysis, a scoring procedure was undertaken, involving both separate and paired scoring. In the case of separate scoring, each tier was considered as an individual item and assigned a score of 1 for correct responses and 0 for incorrect ones (Xiao, Han, Koenig, Xiong, & Bao, 2018). Under paired scoring, respondents received a score of 1 only when both tiers were answered correctly, and a score of 0 if one or both tiers were answered incorrectly. Both of these scoring methods were subjected to analysis using the licensed WINSTEPS program to evaluate their compatibility with the Rasch model.

The test was designed to accommodate as many student ideas as possible, including both correct and incorrect answers, based on previous research findings and idea screening through pilot testing. Although uncommon, some items provided more than one correct answer, and respondents were allowed to select multiple answers they deemed correct for all items. With this pattern, in the initial stage, data processing was conducted in Excel to process raw data of respondent choices, such as A, B, C, D, E, AB, BDC, etc. The correct answers for these items were not just A, B, etc., but could also be combinations of options, such as BCE. The scoring was conducted according to the following criteria: the answer was considered correct with a score of 1 if the student's choice was a complete combination, i.e., BCE. If the student's answer was B, C, E, or a combination of BC, BE, or CE, a score of 0 was given.

This multiple-choice format is actually a modification of item format with options that are combinations of statements in the stem, for example: *1. Statement W, 2. Statement X, 3. Statement Y, 4. Statement Z. Options: Choose A if 1, 2, 3 are correct; choose B if 1, 3 are correct, choose C if 2, 4 are correct, choose D if 4 is correct, and choose E if all are correct.*

The test development data was analysed using WINSTEP 5.4.3.0 Version (licenced) for Rasch analysis. The Rasch model is a probabilistic model that describes what happens when a respondent interacts with test items and converts raw score calculations into a common scale for measuring the respondent's ability (Linacre, 1994). The model assumes that all items only investigate the measured variable (in this case, about mental models and patterns of representing objects in static fluids). Analysis of the O2Tt includes item reliability, person reliability, item distribution in the Wright map, and data fit with the model.

## Findings and Discussion

### Result from Test Development

As previously explained, the test development in this study was based on previous research (Kaharu & Mansyur, 2021). For example, the transformation process of one item from the essay version (offline) to the two-tier multiple-choice version (online) is presented. Figure 2 shows a sample of the essay test item from Version-0 (translated from Indonesian language).

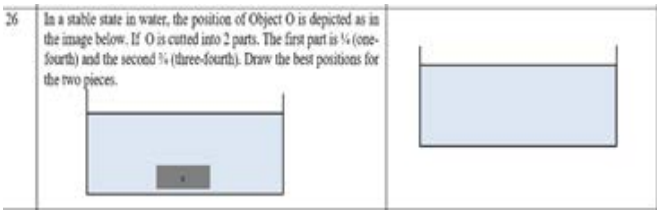


**Fig. 2:** Sample item Versi-0 (essay, offline)

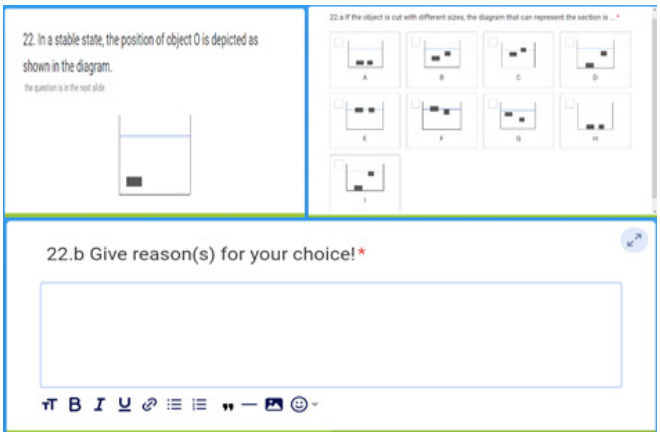


**Fig. 3:** Sample item of Version-1 (Tier-1 and space for developing Tier-2)

The item was then transformed into a multiple-choice format and presented and administered online. In this case, the item from Version-0 became Tier-1 in Version-1. The online Version-1 test (Figure 3) was then piloted with pre-service physics teachers.

To construct Tier-2 of the item, respondents were asked to provide the reasons for choosing the option in Tier-1. These reasons were then combined with the findings from previous research, particularly from Mansyur et al. (2022a) and Mansyur et al. (2022b). This process resulted in the development of Version-2 test (multiple-choice, two-tier, online) as shown in Figure 4.

Subsequently, Tryout-2 was conducted on Version-2 of the test, focusing on the readability of statements in Tier-1 and Tier-2. Tryout participants were asked to provide feedback on the clarity of the statements in each item. Their opinions were written on the provided paper sheets. Based on the tryout, some revisions were made. This process resulted Version-3 of the test.

### Results of Tes Analysis

The test analysis consists of qualitative-descriptive analysis and statistical analysis. The qualitative-descriptive analysis is related to the construct validity data and utilizes the data obtained from expert validation. The statistical analysis includes person reliability, item reliability, item separation index, and other aspects derived from the WINSTEPS output for Rasch model analysis.

### Content and Face Validity

Content and face validity is determined by referring to the guidelines for developing multiple-choice tests. Result of the validation is depicted in Table 2.

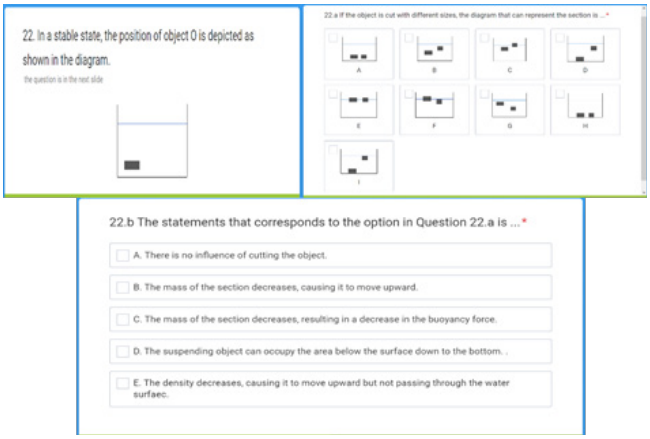


**Figure 4.** An example item of Version-2 (online, Tier-1 dan Tier-2). The first question measures student understanding and the second question asks the student to indicate the reason for choosing the answer in the first question

**Table 2**: Validation result (content and face validity) by two experts for multiple-choice test items.

| | *Average Score*** | |
|---|---|---|
| *Aspect* | *Expert-1* | *Expert-2* |
| Language use | 5,0 | 4,2 |
| Conceptual accuracy | 5,0 | 4,2 |
| Clarity of stem | 4,9 | 5,0 |
| Clarity of stem image* | 4,9 | 4,6 |
| Clarity of option image* | 5,0 | 4,6 |
| Option homogeneity | 4,5 | 4,3 |
| Ordering of options | 4,8 | 4,1 |
| Functionality of options | 4,6 | 4,0 |
| Relevance of Tier-1 and Tier-2 options | 4,8 | 4,8 |

*Specifically for items that include images

**From overall items, maximum score: 5 (excellent)

Overall, the scores by the two validators indicated that the items generally exhibit good language use, conceptual accuracy, clarity of stem and option images, with some space for improvement in terms of option homogeneity, ordering, and functionality. The options in both tiers are considered highly relevant to the test construct.

**Table 3**: Results of online system validation by two experts

| No. | Aspect | Score* Expert-1 | Score* Expert -1 |
|---|---|---|---|
| *a. Design and application* | | | |
| 1 | Layout (criteria: *Informative banner, interactive on each page, proportional, concise.*) | 5 | 4 |
| 2 | Content size (criteria: Aesthetic, symmetrical, balanced, consistent). | 5 | 5 |
| 3 | Color usage (criteria: Consistent, harmonious, vibrant, suitable background color) | 5 | 5 |
| 4 | Font type and size (criteria: Readable, proportional, appropriate heading and body, familiar) | 4 | 5 |
| *b. Web device* | | | |
| 5 | Navigation (criteria: Consistent, easy to understand, simple, efficient) | 5 | 5 |
| 6 | Page loading speed (criteria: Fast before cookie is stored, fast after cookie is stored, all content (images, text, and banners) loads simultaneously, using appropriate JavaScript in the menu) | 5 | 5 |
| 7 | Stability (criteria: Can be accessed multiple times, no server errors, all scripts function properly, content displays orderly) | 4 | 5 |
| 8 | Device compatibility (criteria: Laptop, personal computer, tablet computer, mobile phone) | 5 | 5 |
| 9 | Image (criteria: Readability with medium resolution, fast accessibility, proportional) | 5 | 4 |
| *c. Semantic web* | | | |
| 10 | Content (criteria: Easy to understand; proportional between text, images, and titles; uses language according to Indonesian language guidelines; neat structure and taxonomy) | 4 | 4 |
| 11 | Language use (criteria: Easy to understand, clear, communicative, effective) | 5 | 4 |
| 12 | Readability of text (criteria: Easy to understand, effective, does not cause multiple interpretations, readable) | 5 | 5 |

**Maximum score: 5 (excellent)*

The validators also provided notes and suggestions for test improvement. Examples of the suggestions given are: improvement in the instructions and inclusion of item numbers on each slide.

## Validators' suggestion:

a. The condition that is applied to answer the next question after answering an on-screen question can be bypassed by selecting pagination navigation below. If it is allowed, there is no need for any restrictions, or if leaving the answer blank is not permitted, then the pagination option below can be removed.

b. If possible, the questions can be numbered per section. Before item 1.a, there might be a statement representing the main topic of items 1.a-1.b and so on. Question content that includes statements on the following page should also be numbered, for example, the statement before item 5a can be numbered as 5.

## Rasch Analysis

The Rasch model is a probabilistic framework that elucidates the interaction between individuals (test takers or survey respondents) and the items within those tests or surveys. It is determined by two key parameters: item difficulty and person ability. Rasch measurement mirrors the principles of physical measurements, creating and employing linear measurements for both individuals and items that remain consistent regardless of the specific attributes of the sample or the test items. This is done along a single-dimensional construct (Planinic, Boone, Susac, & Ivanjek, 2019).

To evaluate the performance of the O2Tt test instrument, an analysis of the summary statistics, item fit statistics, and Wright map was conducted for both scoring models.

Initially, the item fit statistics were assessed for all 23 items across both scoring frameworks. In both scoring models, the collective infit values for MNSQ fell within the range of 0.5 to 1.5. However, two specific items (21 and 22) exhibited MNSQ outfit values surpassing 1.5. Despite their elevated outfit values, these two items were retained after undergoing revisions. This decision was grounded in the likelihood that the increased outfit values were attributed to issues with the combination of correct answers. Consequently, all items were retained for continued analysis.

Figures 5(a) and 5(b) illustrate bubble plots displaying the infit MNSQ of 23 items for both scoring models.

These plots visualize the test item fit by plotting the item difficulty level on the vertical axis and the item infit MNSQ on the horizontal axis (Ivanjek et al., 2021). Each circle represents an item, and its size corresponds to the standard error of calibration. The horizontal distance between the circle and the expected fit MNSQ of 1 indicates how well the item fits the model. Ideally, items should be near the center of the plot along the infit value. The vertical axis shows the item difficulty level in logits, which allows for easier comparison of difficulty levels between items (Planinic et al., 2019). Overall, this plot can be a useful tool for evaluating the performance of test items and identifying areas that need improvement in the test development process.

The outcomes of WINSTEPS analysis for person reliability, item reliability, and the separation index are summarized in Table 4. The evaluation encompassed both item and person reliability. For separate scoring model, a notably high item reliability of 0.94 was observed, signifying the capacity to effectively differentiate item parameters. The assessment of person reliability yielded a value of 0.56 for separate scoring and 0.00 for paired scoring, with corresponding person separation indices of 0.00 and 1.14, respectively. While these values are deemed satisfactory as per established standards (Boone, Satver, & Yale, 2014), it's worth noting that there are inherent limitations in obtaining dependable person measurements.

For each context, where each tier is considered as a separate item, the person separation index obtained is 1.14 (low category). Through the strata separation, H=4x1.14/3=1.52, which is rounded to 2. This means that the test can only divide respondents into two groups, namely the high and low ability groups.

The item separation index is 4.12 (very good category). Using the same method, H=5, which means that the test with tiers considered as separate items can be divided into five groups based on the difficulty levels of the items, namely very difficult, difficult, moderate, easy, and very easy.

The person reliability is 0.56 (low category), indicating that the consistency of respondents' answers is weak. However, from the aspect of item quality, the item reliability is 0.94 (very good), suggesting that the items are highly reliable.

Regarding the Infit MNSQ and Outfit MNSQ scores in the Table 4 pertaining to the average person, the recorded values are 1.00 and 1.03, respectively. The optimal value is 1.00, and the closer the values approach this benchmark, the

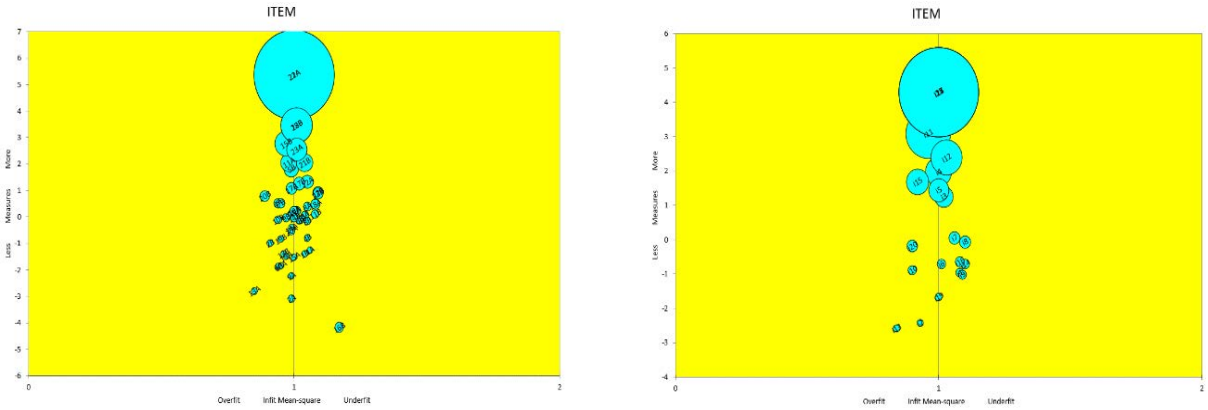


**Fig.5:** (a) Bubble chart for MNSQ Infit values for separate scores, (b) Bubble chart for MNSQ Infit values for paired scores.

more favorable it is. As for Infit ZSTD and Outfit ZSTD, the average person values are 0.0 and 0.1, respectively. Values that gravitate towards 0.0 are deemed more desirable. Based on these findings, it can be deduced that, on the whole, the fit of individual persons is considered acceptable.

Within the same table, the mean Infit MNSQ and Outfit MNSQ for the items stand at 1.00 and 1.03, correspondingly. The optimal target for these measurements is 1.00, and the greater proximity to this figure, the more favorable the fit of the items. Similarly, the average Infit ZSTD and Outfit ZSTD values for the item segment are -0.1 and 0.0, sequentially. Values that draw closer to 0.0 are indicative of a more suitable fit for the items. Based on this dataset, it can be inferred that the overall alignment of the items is considered acceptable.

In the context of paired scores (in Table 5), the person separation index has a value of 0.00. This indicates that the two-tier pairs cannot differentiate respondents' abilities. In other words, in this testing scenario, no significant ability differences can be distinguished among the respondents. Meanwhile, for the item separation index, a value of 2.22 is obtained. To calculate the number of groups that can be formed based on difficulty, the formula $H = 4x2.22/3 = 2.96$ (almost 3) is used. This means that in paired scores, the test can be divided into three groups based on the difficulty levels of the items, namely difficult, moderate, and easy groups.

When using paired scores, the value of person reliability is 0.00, indicating very weak consistency in respondents' answers. This indicates that the measurement of respondents' abilities is not consistently reliable. However, in terms of item reliability, a value of 0.83 is obtained. This indicates that the quality of the tier-pairings is good. It means that the difficulty levels of the items in this test can be consistently measured and effectively differentiated.

Overall, despite the very weak consistency in respondents' answers in paired scores, the quality of tier-pairings remains good. In this context, the test can still provide relevant information about the difficulty levels of the items, although consistency in measuring respondents' abilities needs improvement.

Infit MNSQ and Outfit MNSQ are metrics used in the analysis of item fit in psychometric testing. These variables provide information about how well an item in the person table fits the measurement model used. Infit MNSQ and Outfit MNSQ are used to measure the level of item fit with the desired measurement scale. In this case, the average value of Infit MNSQ is 1.01, and the Outfit MNSQ is 0.96. Ideally, the MNSQ values should be close to 1.00, indicating better item fit as they approach 1.00. In this case, the slightly lower average values of Infit MNSQ and Outfit MNSQ, which are close to 1.00, indicate that overall, the item fit is considered acceptable.

For Infit ZSTD and Outfit ZSTD, the average values in Table 5 are 0.1 and 0.2, respectively. Values close to 0.0 are considered better because they indicate a better fit of the items with the measurement model. In this case, although the average values of Infit ZSTD and Outfit ZSTD are slightly higher than 0.0, these values are still relatively low and indicate that the item fit is considered acceptable. Based on this data, it can be generally stated that the item fit is acceptable.

Despite some slight discrepancies, both Infit MNSQ and Outfit MNSQ, as well as Infit ZSTD and Outfit ZSTD, have average values that are relatively close to the ideal values. This suggests that overall, the items in the person table are consistent with the measurement model used.

Based on the data from Table 5, it can be concluded that overall, the item fit is considered acceptable. The average value of Infit MNSQ is 1.00, and the Outfit MNSQ is 1.08. Ideally, the MNSQ values should be close to 1.00, indicating good item fit with the model. In this case, the average value of Infit MNSQ being close to 1.00 indicates that most items in the item table have a reasonably good level of fit with the desired measurement scale. Although there are some slight deviations from the ideal value, these differences are still acceptable.

Furthermore, for Infit ZSTD and Outfit ZSTD, the average values in the person table are 0.0 and -0.1, respectively. Values close to 0.0 are considered better as they indicate a better fit of the items with the model. In this case, the average value of Infit ZSTD is close to 0.0, while the average value of Outfit ZSTD is slightly lower than 0.0. Although there are some deviations from the ideal value of 0.0, these values still

**Table 4:** Statistic summary by WINSTEPS for separate scoring

Output to E:\PRIVACY\PENELITIAN 2016-2025\HIBAH FAKULTAS 2023\ZOU703WS.TXT
WINSTEP-318 ORANG-AKUN PREMIUM.xlsx

| PERSON | 318 INPUT | | 318 MEASURED | | INFIT | | OUTFIT | |
|---|---|---|---|---|---|---|---|---|
| | TOTAL | COUNT | MEASURE | REALSE | IMNSQ | ZSTD | OMNSQ | ZSTD |
| MEAN | 10.1 | 46.0 | -1.82 | .46 | 1.00 | .0 | 1.03 | .1 |
| P.SD | 3.6 | .0 | .71 | .08 | .22 | .9 | .76 | .8 |
| REAL RMSE | .47 TRUE SD | | .54 | SEPARATION | 1.14 | PERSON RELIABILITY | | .56 |
| ITEM | 46 INPUT | | 46 MEASURED | | INFIT | | OUTFIT | |
| | TOTAL | COUNT | MEASURE | REALSE | IMNSQ | ZSTD | OMNSQ | ZSTD |
| MEAN | 69.8 | 318.0 | .23 | .29 | 1.00 | -.1 | 1.03 | .0 |
| P.SD | 68.5 | .0 | 1.94 | .36 | .06 | .8 | .25 | 1.2 |
| REAL RMSE | .46 TRUE SD | | 1.88 | SEPARATION | 4.12 | ITEM | RELIABILITY | .94 |

**Table 5.** Statistic summary by Winstep for paired scoring

Output to E:\PRIVACY\PENELITIAN 2016-2025\HIBAH FAKULTAS 2023\HASIL WINSTEP\ZOU
WINSTEP-318 ORANG-AKUN PREMIUM-Skor berpasangan.xlsx

| PERSON | 318 INPUT | | 318 MEASURED | | INFIT | | OUTFIT | |
|---|---|---|---|---|---|---|---|---|
| | TOTAL | COUNT | MEASURE | REALSE | IMNSQ | ZSTD | OMNSQ | ZSTD |
| MEAN | 1.8 | 23.0 | -3.32 | 1.13 | 1.01 | .1 | .96 | .2 |
| P.SD | 1.7 | .0 | 1.22 | .46 | .25 | .6 | 1.15 | .7 |
| REAL RMSE | 1.22 TRUE SD | | .00 | SEPARATION | .00 | PERSON RELIABILITY | | .00 |
| ITEM | 23 INPUT | | 23 MEASURED | | INFIT | | OUTFIT | |
| | TOTAL | COUNT | MEASURE | REALSE | IMNSQ | ZSTD | OMNSQ | ZSTD |
| MEAN | 25.3 | 318.0 | .93 | .66 | 1.00 | .0 | 1.08 | -.1 |
| P.SD | 30.9 | .0 | 2.26 | .65 | .08 | 1.0 | .72 | 1.4 |
| REAL RMSE | .93 TRUE SD | | 2.06 | SEPARATION | 2.22 | ITEM | RELIABILITY | .83 |

indicate that overall, the item fit is considered acceptable.

Outfit statistics tend to exhibit heightened sensitivity to outliers, whereas infit statistics are more responsive to respondents' reactions to items with difficulties that closely align with their abilities. Alternatively, outfit and infit can be evaluated based on ZSTD values, where ZSTD represents a standardized Z score of the residual. Conventionally, it's acknowledged that items displaying infit and outfit MNSQ values ranging from 0.7 to 1.3, as well as infit and outfit ZSTD values between -2 and 2, demonstrate a favorable fit with the model. Furthermore, items with infit and outfit MNSQ values spanning from 0.5 to 1.5 can be considered valuable for measurement purposes (Ivanjek et al., 2021; Boone et al., 2014). The evaluation of item functioning can be executed using fit statistics and point-measure correlations, which illuminate the degree to which a specific item contributes to the overarching person or item measure.

Figure 6 shows the Wright map, which illustrates the difficulty level of items (vertical axis) against Infit MNSQ item. In this Wright map, respondents with high abilities and items with higher difficulty levels are located at the top of the map. Conversely, respondents with low abilities and easier items are positioned closer to the bottom of the map. This indicates the relationship between the difficulty level of items and the abilities of respondents in that category.

Furthermore, in Figure 6 (left part), it can be observed that the mean (M) of respondent scores is below the mean (M) of item scores. This indicates that, overall, the test items (separate scores model) in that category are difficult for the respondents taking the test. The mean of respondents being lower than the mean of items suggests that most respondents face difficulties in answering the items in that test.

In the context of the paired scores model, Figure 6 (right part) shows that the items are extremely difficult for the respondents. This can be seen from the position of the item mean, which is significantly higher than the persons mean, indicating a high level of difficulty for respondents in answering those items. The figure provides a visual representation of the relationship between the difficulty level of items and the abilities of respondents in that category. The analysis of the Wright map can offer valuable insights into understanding response patterns and item fit in psychometric testing.

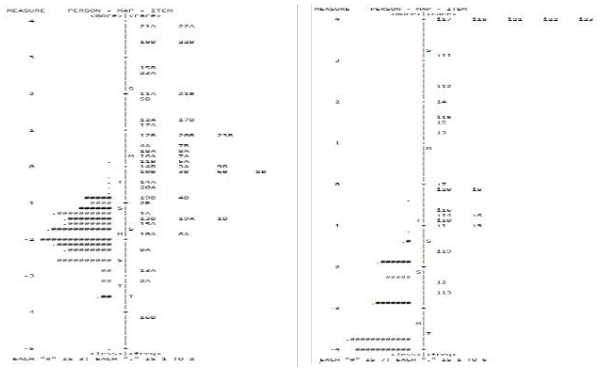

**Fig. 6:** Display of the Wright map for separate scores (left) and paired scores (right).

Figure 6 is a Wright map that presents item difficulty (vertical axis) against Infit MNSQ item. Respondents with high abilities and items with higher difficulty levels are positioned at the top of the map. Conversely, respondents with low abilities and easier items are closer to the bottom of the map. Figure 3 (left) also indicates that the mean (M) of respondent scores is below the mean (M) of item scores, suggesting that, overall, the test items (separate scores) are difficult for the respondents taking the test. For the paired scores model, the figure shows that the items are very difficult for the respondents.

Another piece of information from Rasch modeling is examining the quality of item fit to the model, abbreviated as item fit. Item fit assesses whether the items function as expected in the measurement process. According to Boone et al. (2014), criteria used to evaluate the level of item fit are outfit mean-square, outfit z-standard, and point measure correlation. If an item does not meet these criteria, it can be considered as not fitting well, indicating that the item may need to be revised or replaced. The criteria are:

a. 0,5 < MNSQ < 1,5
b. -2,0 < ZSTD < +2,0
c. 0,4 < Pt Measure Corr. < 0,85

**Table 6:** Two items do not meet the requirements for separate scoring

| INFIT | | OUTFIT | | PTMEASUR-AL | | EXACT MATCH | | |
|---|---|---|---|---|---|---|---|---|
| MNSQ | ZSTD | MNSQ | ZSTD | CORR. | EXP. | OBS% | EXP% | ITEM |
| MAXIMUM MEASURE | | | | .00 | .00 | 100.0 | 100.0 | 21A |
| MAXIMUM MEASURE | | | | .00 | .00 | 100.0 | 100.0 | 22A |
| 1.01 | .24 | .98 | .21 | .03 | .05 | 99.4 | 99.4 | 18B |
| 1.01 | .24 | .98 | .21 | .03 | .05 | 99.4 | 99.4 | 22B |
| .97 | .09 | .40 | -1.18 | .19 | .07 | 98.7 | 98.7 | 15B |
| 1.01 | .16 | 1.41 | .87 | .03 | .07 | 98.4 | 98.4 | 23A |

**Table 7**. Five items do not meet the requirements for paired scoring

| ENTRY NUMBER | TOTAL SCORE | TOTAL COUNT | JMLE MEASURE | MODEL S.E. | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD | PTMEASUR-AL CORR. | PTMEASUR-AL EXP. | EXACT MATCH OBS% | EXACT MATCH EXP% | ITEM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 17 | 0 | 318 | 4.30 | 1.83 | MAXIMUM MEASURE | | | | .00 | .00 | 100.0 | 100.0 | i17 |
| 18 | 0 | 318 | 4.30 | 1.83 | MAXIMUM MEASURE | | | | .00 | .00 | 100.0 | 100.0 | i18 |
| 21 | 0 | 318 | 4.30 | 1.83 | MAXIMUM MEASURE | | | | .00 | .00 | 100.0 | 100.0 | i21 |
| 22 | 0 | 318 | 4.30 | 1.83 | MAXIMUM MEASURE | | | | .00 | .00 | 100.0 | 100.0 | i22 |
| 23 | 0 | 318 | 4.30 | 1.83 | MAXIMUM MEASURE | | | | .00 | .00 | 100.0 | 100.0 | i23 |
| 11 | 1 | 318 | 3.09 | 1.01 | .96 | .29 | .14 | -1.50 | .13 | .06 | 99.6 | 99.6 | i11 |
| 12 | 2 | 318 | 2.39 | .71 | 1.03 | .28 | 3.69 | 2.31 | -.03 | .09 | 99.2 | 99.2 | i12 |
| 4 | 3 | 318 | 1.97 | .59 | 1.00 | .19 | .68 | -.18 | .12 | .10 | 98.8 | 98.8 | i4 |
| 15 | 4 | 318 | 1.68 | .51 | .92 | .00 | .36 | -1.03 | .20 | .12 | 98.3 | 98.3 | i15 |

If an item does not meet all three of these criteria, then the item is categorized as not good and needs to be revised or replaced. Based on these criteria, items 21A and 22A (Table 6) do not meet the requirements. However, instead of discarding or replacing these items, they will be revised by reducing the number of response options. Both items initially had 9 response options, and the revision will reduce them to 5 options.

Based on the criteria, the items that do not meet the requirements are item numbers 17, 18, 21, 22, and 23 (Table 7). However, instead of discarding or replacing these items, they have been revised. Items 21 and 22 have been revised by reducing the number of response options provided. Items 17 and 18 have been revised based on the homogeneity aspect of the options in Tier-2. Item 23 has been revised based on the homogeneity aspect of the options in both tiers.

The data analysis showed that both in separate and paired scores, the test in the category was very difficult for the group of test subject. The number and combination of correct answers reduced the chances for individuals to answer correctly. This is a rational argument because if there are five options and only one correct answer, the probability of each option being chosen is 20%. If the number of options is more than five, it will further reduce the probability of each option being chosen. This probability decreases even more if the correct answer is a combination of two or more options. When there are multiple correct answers or a combination of options that form the correct answer, the probability of selecting all the correct options decreases even further. This is because the individual needs to correctly identify and choose all the correct options from the available choices, which becomes more challenging as the number of options increases (Xiao et al., 2018). The rationale behind this decrease in probability is based on the principles of chance and decision-making. With more options to choose from, the chances of randomly selecting the correct answer(s) decrease. Additionally, the cognitive demands of identifying and combining multiple correct options add another layer of complexity to the decision-making process (Gierl, Bulut, Guo, & Zhang, 2017).

Based on these considerations, revisions were made to the item structure, especially regarding the number of options and the number of correct options in each item. Revisions were also made regarding the homogeneity of options to ensure that all options within an item are within the same context or concept domain. Through this entire process, a test was obtained that follows the general rules of multiple-choice test construction, including having the same number of options in all items, containing only one correct answer, and having homogeneous options within each item in terms of explored context and concept. The final version of the test is presented in the Supplemental Material, which can be accessed by permission through a link.

## Conclusion

In this study, we have developed an online diagnostic test on the basic concepts of static fluids that is limited to the properties of object in liquids. This test can be used for middle school students up to college level with some notion. The uniqueness of this test instrument lies in the fact that it is a two-tier test, with items that can accommodate the ideas of learners in the concept domain. Rasch analysis of the two-tier instrument showed that, overall, the test functioned well, but there were some items that did not function properly. The dysfunctionality of these items is suspected to be related to the variation in the number of options in both tiers and the presence of multiple correct answers. The instrument is very difficult for the sample of pre-service elementary school teachers. The mean distribution of student abilities is far from the logit score for separate scoring and even further below the mean difficulty of the items, indicating that the test is too difficult for the sample. The presence of highly difficult items, particularly in the domain of suspending objects, results in very low person reliability, especially when the paired scoring model is used. However, we have decided to include these more difficult items in the test with the consideration that the development of the test idea started from findings in that particular concept domain and representative items of the domain are needed.

## Suggestion

The process of developing the Tier-2 and readability of the test involved pre-service physics teachers, and the data for Rasch analysis included pre-service elementary school teachers. The research results demonstrate that despite the test having high item reliability, it is considered very difficult for pre-service elementary school teachers. The characteristics of the test can be further examined by involving physics students or pre-service physics teachers.

## Limitation

The scope of fundamental concepts of static fluids is quite extensive, however, the developed test is limited to concepts of floating, suspending, sinking, density, and factors related to the behavior of objects within a liquid fluid. Students' mental model, representation pattern or conception are not included in this article. They will be published in the next publication.

## Acknowledgements

This study was funded by DIPA 2023 of Tadulako University. We are grateful to students who were involved as participants, as well as and those that contributed to this study.

## References

Boone, W.J., Satver, J.R., and Yale, M.S. (2014). *Rasch Analysis in Human Sciences*. Dordrecht: Springer.

Canlas, I. P. (2019). Using visual representations in identifying students' preconceptions in friction. *Research in Science and Technological Education*, *39*(2), 1–29. DOI: 10.1080/02635143.2019.1660630

Castillo, R. D., Waltzer, T. & Kloos, H. (2017). Hands-on experience can lead to systematic mistakes: A study on adults' understanding of sinking objects, *Cognitive Research: Principles and Implications, 2*(1). DOI: 10.1186/s41235-017-0061-8

Chen, C., Bao, L., Fritchman, J. C., & Ma, H. (2021). Causal reasoning in understanding Newton's third law, *Physical review physics education research 17*, 010128

Chien, S., Hsiung, C., & Chen, S. (2009). The development of young children's science- related concept regarding floating and sinking, *Asia-Pacific Journal of Research in Early Childhood Education, 3*(2), 73–88. http://www.pecerajournal.com/detail/21791.

Djudin, T. (2021). Promoting students' conceptual change by integrating the 3-2-1 reading technique with refutation text in the physics learning of buoyancy. *Journal of Turkish Science Education, 18*(2), 290-303. DOI: 10.36681/tused.2021.66.

Gette, C. R., Kryjevskaia, M., Stetzer. M. R., & Heron, P. R. L. (2018). Probing student reasoning approaches through the lens of dual-process theories: A case study in buoyancy, *Physical Review Physics Education Research, 14*(1), 010113. DOI: 10.1103/PhysRevPhysEducRes. 14.010113.

Gierl, M.J., Bulut, O., Guo, Q., and Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research, 87*(6), 1082–1116 DOI: 10.3102/0034654317726529.

Havu-Nuutinen, S (2005). Examining young children's conceptual change processing floating and sinking from a social constructivist perspective, *International Journal of Science Education, 27*(3), 259–279, 2005. https://doi.org/10.1080/0950069042000243736

Hsin, C.T., & Wu, H.K. (2011). Using Scaffolding Strategies to Promote Young Children's Scientific Understandings of Floating and Sinking, *Journal of Science Education and Technology, 20*(5), 656–666. https://doi.org/10.1007/s10956-011-9310-7

Hooda, M. & Devi, A. (2018). Development of Reasoning Ability and Barriers among Students, *International Journal of Research in Engineering, IT and Social Sciences, 8*(11), 129-131.

Ivanjek, L., Morris, L., Schubatzky, T., Hopf, M., Burde, J., Haagen-schützenhöfer, C., Dopatka, L., Spatz, V., & Wilhelm, T. (2021). Development of a two-tier instrument on simple electric circuits, *Physical Review Physics Education Research,17*(2),20123. https://doi.org/10.1103/PhysRevPhysEducRes.17.020123

Kafiyani, F., Samsudin, A., & Saepuzaman, D. (2019). Development of four-tier diagnostic test (FTDT) to identify student's mental models on static fluid, *Journal of Physics: Conference Series, 1280*(5). DOI: 10.1088/1742-6596/1280/5/052030.

Kaharu, S. N. & Mansyur, J. (2021). The development of a test to explore the students' mental models and external representation patterns of hanging objects, *Pegem Journal of Education and Instruction, 11*(4), 110-125. DOI: 10.47750/pegegog.18.4.011

Karmiloff-Smith, A. (1990). Constraints on representational change: Evidence from children's drawing, *Cognition, 34*(1990), 57-83.

Linacre, J. M. 1994. *A User's Guide to WINSTEPS*, https://www.WINSTEPS.com/manuals.html.

Liu, Z., Pan, S., Zhang, X., & Bao, L. (2022). Assessment of knowledge integration in student learning of simple electric circuits, *Physical Review Physics Education Research 18, 020102.*

Mansyur, J., Werdhiana, I. K., Darsikin, D., Kaharu, S. N., Tadeko, N. (2022a). Students' mental models about the suspending objects in static fluid, *Journal of Turkish Science Education,19*(1), 253-283.

Mansyur, J., Werdhiana, I. K., Darsikin, D., Kaharu, S. N., Tadeko, N. (2022b). Students' external representation patterns of the suspending objects in static fluid, *European Journal of Educat ional Research, 11*(2), 805-820. DOI: 10.12973/eu-jer.11.2.805.

Minogue, J., Borland, D., Russo, M., Chen, S. T., & Grady, R. (2015). Investigating the influence of haptic technology on upper elementary students' reasoning about sinking & floating, *2015 Annual International Conference*, Chicago.

Minogue, J. & Borland, D. (2016). Investigating students' ideas about buoyancy and the influence of haptic feedback, *Journal of Science Education and Technology, 25*,187–202. https://doi.org/10.1007/s10956-015-9585-1.

Planinic,M., Boone, W.J., Susac, A and Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters, *Physical Review Physics Education Research, 15*, 020111.

Shen, J., Liu, O. L., & Chang, H. Y. (2017). Assessing students' deep conceptual understanding in physical sciences: an example on sinking and floating, *International Journal of Science and Mathematics Education, 15*(1), 57–70. https://doi.org/10.1007/s10763-015-9680-z.

Smith, C., Maclin, D., Grosslight, L., & Davis, H. (1997). Teaching for understanding: A study of students' pre-instruction theories of matter and a comparison of the effectiveness of two approaches to teaching about matter and density, *Cognition and Instruction,15*(3), 317–393. https://doi.org/10.1207/s1532690xci1503_2

Tao, Y., Oliver, M. C., & Venville, G. J. (2011). Chinese and Australian year 3 children's conceptual understanding of science: a multiple comparative case study, *International Journal of Science Education, 34*(6), 879–901. DOI:10.1080/09500693.2011.57867.

Teo, T. W., Yan, Y. K., & Ong, W. L. M. (2017). An investigation of Singapore pre-school children's emerging concepts of floating and sinking, *Pedagogies*. DOI: 10.1080/1554480X.2017.1374186.

van Schijndel, T. J. P., van Es, S. E., Franse, R. K., van Bers, B. M. C. W., & Raijmakers, M. E. J. (2018). Children's mental models of prenatal development. *Frontiers in Psychology, 9*(1835), 1–13. DOI: 10.3389/fpsyg.2018.01835.

Viyanti, V., Cari, C., Sunarno, W., & Prastyo, Z. K. (2017). The development rubrics skill argued as alternative assessment floating and sinking materials, *Journal of Physics: Conference Series, 909*(1), 012057. DOI: 10.1088/1742-6596/909/1/012057

Vosniadou, S. & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology, 24*(4), 535–585. DOI: 10.1016/0010-0285(92)90018-W.

Xiao, Y., Han, J., Koenig, K., Xiong, J., and Bao, L. Multilevel Rasch modeling of two-tier multiple choice test: A case study using Lawson's classroom test of scientific reasoning, *Physical Review Physics Education Research, 14*, 020104. DOI: 10.1103/PhysRevPhysEducRes.14.020104.

Yin, Y., Tomita, M. and Shavelson, R. (2008). Diagnosing and dealing with student misconceptions: floating and sinking, *Science Scope, 31*(8), 34–39. w