

Development of a Physics Diagnostic Assessment Framework based on Solo Taxonomy for Senior High School Student

Sabani¹, Irham Ramadhani², Wawan Bunawan³, Seila Agustina⁴, Elang Mulia Lesmana⁵, Dyah Inggit Murtinigrum⁶

^{1,2,3}Physics Education, Universitas Negeri Medan, Medan, Indonesia

⁴Bilingual Physics Education, Universitas Negeri Medan, Medan, Indonesia

^{5,6}Teacher Professional Education, Universitas Negeri Medan, Medan, Indonesia

¹sabani@unimed.ac.id, ²irhamramadhani@unimed.ac.id, ³wawanbunawan@unimed.ac.id,

⁴seilaagustina@mhs.unimed.ac.id, ⁵elang.mulia179@guru.sma.belajar.id, ⁶dyah.inggit14@guru.sma.belajar.id

ABSTRACT

The paradigm shift toward Deep Learning in Indonesia demands assessment tools that capture the quality of students' cognitive structures rather than mere procedural memorization. This study aims to develop and validate a physics diagnostic test framework based on the SOLO taxonomy to evaluate high school students' understanding of fluid concepts. Utilizing a Research and Development (R&D) design adapted from the Borg and Gall model, the study involved 80 eleventh-grade students in Medan. Data were analyzed using Classical Test Theory via SPSS and Microsoft Excel. Psychometric evaluations optimized the initial test instrument diagnostic based on SOLO Taxonomy into 12 valid items with validity coefficients ranging from 0.480 to 0.815. The final instrument demonstrated excellent internal consistency with a reliability coefficient of 0.887. Item discrimination was mostly very good (78.57%), with a balanced difficulty distribution (28.57% easy, 57.14% moderate, 14.28% difficult). The final framework successfully mapped cognitive structures across unistructural, multistructural, relational, and extended abstract levels. This valid and reliable diagnostic tool effectively measures students' structural understanding to support the deep learning paradigm.

Keywords: SOLO Taxonomy, Diagnostic Instrument, Fluid Concepts, Physics Assessment.



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Corresponding Author:

Irham Ramadhani,

Physics Education,

Universitas Negeri Medan,

Jl. William Iskandar Ps. V, Kenangan Baru, Kec. Percut Sei Tuan, Kabupaten Deli Serdang 20371, Indonesia.

irhamramadhani@unimed.ac.id

1. INTRODUCTION

Physics learning at the Senior High School (SHS) level aims to equip students with an understanding of scientific concepts related to natural phenomena. One of the essential topics is fluids, which include static and dynamic fluids and are closely related to everyday life, such as the working principles of pumps, buoyant force, blood pressure, and water flow. The topic of fluids is important because it provides a fundamental understanding of how pressure, buoyancy, and fluid motion govern various natural phenomena and technological applications. Understanding fluid concepts enables students to explain and analyze real-world phenomena, such as the operation of hydraulic systems, water distribution, aircraft lift, and the circulation of blood in the human body. Therefore, learning fluids is essential not only for developing students' conceptual understanding of physics but also for connecting scientific principles with practical applications in everyday life.

Ideally, physics learning should not only emphasize formulas and problem-solving procedures but also develop students' scientific thinking and deep conceptual understanding. However, various studies indicate that students' scientific literacy and conceptual understanding, particularly in fluid topics, remain low. data from the Programme for International Student Assessment (PISA) 2022 show that only about 34% of Indonesian students reached level 2 or higher in scientific literacy, far below the OECD average of 76%. This suggests that many students still struggle to interpret scientifically, reasoning.

These difficulties are also evident in students' understanding of static fluid concepts such as hydrostatic pressure, pascal's law, and archimedes' principle. Students are generally able to recall and apply formulas but struggle to integrate concepts into a coherent framework. From the perspective of the SOLO taxonomy (structure of the observed learning outcomes), students' understanding tends to remain at the unistructural and multistructural levels, where concepts are understood separately without

meaningful integration. previous studies have reported significant levels of misconceptions in fluid topics, such as hydrostatic pressure (16.19%), pascal's law (39.04%), and archimedes' principle (45.71%) (Ahmadi et al., 2024). Other studies also found that up to 88.5% of students experience misconceptions in static fluid materials (Atmaja & Samsudin, 2024; Hutasoit et al., 2025; Indrasari, 2025; Muki et al., 2025; Putri & Nasbey, 2023; Kurniawan, 2023; Koto & Ilhami, 2023; Fatmawati et al., 2025). This condition indicates that conventional assessments, which mainly focus on memorization and procedural skills, are insufficient to reveal the depth of students' conceptual understanding.

Initially observations through a survey revealed that most students actually demonstrate a tendency toward conceptual and analytical thinking. Approximately 76.7% of students reported that they understand physics concepts better through relationships between concepts rather than memorization, and 80% feel challenged by multi-step problems. Additionally, 76.6% stated that higher-order thinking activities improve their understanding. However, nearly half of the students believe that current assessments still emphasize memorization and do not fully reflect their thinking abilities. This indicates a mismatch between students' cognitive characteristics and the forms of assessment used in schools. one potential solution is the implementation of diagnostic tests based on the SOLO taxonomy. this taxonomy, developed by Biggs and Collis, classifies students' levels of understanding based on the complexity of their responses, ranging from prestructural, unistructural, and multistructural to relational and extended abstract levels (Biggs & Collis, 1982). Unlike bloom's taxonomy, which focuses on cognitive processes, the SOLO taxonomy emphasizes the quality and structure of understanding (Fernandez & Guzon, 2025). Therefore, SOLO-based diagnostic tests are more suitable for identifying students' conceptual development levels.

The urgency of developing assessment instruments capable of measuring depth of understanding is increasingly relevant in the context of national education transformation, particularly in the implementation of the Deep Learning paradigm, which is currently the policy direction for improving the quality of learning in Indonesia. Deep learning emphasizes a meaningful, conscious, and enjoyable learning process through the active involvement of students in acquiring knowledge, connecting various concepts, and transferring understanding to new situations (Sabani et al., 2026). In this context, learning success is no longer simply measured by the ability to recall information or solve routine problems, but rather by students' ability to demonstrate integrated conceptual understanding, conduct scientific reasoning, and apply concepts in various real-life contexts (Marton & Saljo, 1976; Biggs, 1987; Ministry of Education and Culture, 2025). The Deep Learning academic paper published by the Ministry of Primary and Secondary Education also emphasizes that this approach requires students to understand, apply, and reflect on knowledge meaningfully, rather than simply memorizing information.

This need is also reflected in the assessment system for the Teacher Professional Education Program (PPG), both the PPG for Specific Teachers and the PPG for Prospective Teachers. In recent years, evaluations of the implementation of the PPG (School Teacher Education Program) have shown that some assessment instruments used are still oriented toward measuring material mastery and procedural skills, while aspects such as depth of conceptual understanding, the ability to integrate various concepts, and the ability to reflect and transfer knowledge have not been optimally measured (Brookhrat, 2010; UNESCO, 2021). Prospective and professional teachers are required not only to master learning content but also to possess higher-order thinking skills that enable them to analyze student misconceptions, design meaningful learning, and develop adaptive learning strategies for various classroom contexts. The in-depth learning orientation emphasizes holistic understanding and the interconnectedness of concepts, further emphasizing the importance of assessments that capture the quality of this understanding (Biggs & Collis, 1982; Biggs & Tang, 2011).

In PPG assessments, participants are often challenged to demonstrate integrated pedagogical and professional skills. However, available instruments generally fail to map in detail the level of development of participants' understanding, from partial understanding to the ability to integrate various concepts and generate new generalizations. As a result, assessment results often only quantitatively describe the level of material mastery without providing diagnostic information about the cognitive structures underlying participants' responses. This situation presents a unique challenge because teachers, who will be tasked with developing critical thinking skills and in-depth understanding in students, must first demonstrate the characteristics of this understanding in their professional education process. In this perspective, the SOLO Taxonomy provides a highly relevant framework to support in-

depth learning-based assessment. The SOLO levels allow for the identification of the development of understanding quality ranging from pre-structural to extended abstract, thus providing a more accurate picture of the development of an individual's ability to connect, integrate, and generalize concepts (Biggs & Collis, 1982; Biggs & Tang, 2011). The alignment between learning principles and the characteristics of the SOLO Taxonomy makes this approach potentially a foundation for developing assessment instruments that are not only relevant for students in schools but also for PPG participants. Thus, the development of a SOLO Taxonomy-based diagnostic instrument for Fluids not only addresses the need for conceptual assessment at the high school level but also contributes to the development of assessment models that align with the demands of in-depth learning and the strengthening of professional teacher competencies in Indonesia.

Previous studies have shown that SOLO based instruments are effective in identifying misconceptions and students' thinking patterns in physics. for example, studies on fluid topics have demonstrated that SOLO based diagnostic tests can reveal misconceptions in hydrostatic pressure and bernoulli's principle (Ahmadi et al., 2024; Hartanti et al., 2021). however, most of these studies focus primarily on identifying misconceptions rather than developing and validating instruments that meet psychometric criteria such as validity, reliability, discrimination power, and difficulty level (Ole & Gallos, 2023; Sedhandoyo, 2024; Eryanto & Prasetyono, 2025; Guzon & Fernandez, 2025; Yuberti et al., 2021; Inggit et al., 2021). Furthermore, limited research integrates students' thinking characteristics into the design of SOLO based diagnostic instruments.

Based on these problems, this study aims to develop a diagnostic test instrument based on the SOLO taxonomy for fluid topics in senior high school and to evaluate its quality using classical test theory, including validity, reliability, discrimination power, and difficulty level. This study is expected to contribute to the development of assessment tools that can measure students' cognitive structures more comprehensively. The results of this study are expected to provide several benefits. For students, the instrument can enhance higher-order thinking skills and deepen conceptual understanding, for teachers, it can serve as an alternative assessment tool that provides more diagnostic and meaningful information about students' learning. For schools, it can support the improvement of physics learning quality. Additionally, this study can serve as a reference for future researchers in developing SOLO-based assessment frameworks in different contexts.

2. RESEARCH METHOD

This study employed a research and development (R&D) design aimed at developing and validating a diagnostic test instrument based on the SOLO taxonomy for fluid topics in senior high school physics. The development procedure followed the Borg and Gall model adapted by Sugiyono (2023), which consists of ten stages: (1) Identifying potential and problems, (2) Data collection, (3) Preliminary product development, (4) Design validation, (5) Design revision, (6) Limited trial, (7) Product revision, (8) Field testing, (9) Final product revision, and (10) Mass production. The research was conducted in April 2026 at a public high school in Medan, North Sumatra. This location was purposively selected because SOLO taxonomy-based assessments had not been previously implemented there, and the students demonstrated adequate academic readiness.

The subjects included 80 students selected using cluster sampling techniques, divided into three classes with first class 20 subjects, each 10 students as subject for first preliminary field testing and second preliminary field testing, second and three class 30 students for each class as wider field testing. The variables measured in this study included validity, reliability, item difficulty, and discrimination power. the developed instrument was a diagnostic test consisting of 17 multiple-choice items with various representations, including verbal, numerical, graphical, and visual forms, designed to assess students' conceptual understanding based on SOLO levels (S2–S5).

Data collection techniques included literature study and test administration. The instrument development process involved expert validation by two lecturers (assessment expert and physics expert), followed by a first preliminary field test with 10 students, second preliminary field test with 10 students, and wider field test with 60 students. the tools used in data processing included microsoft excel for discrimination power, and SPSS for statistical analysis of validity, reliability, and item difficulty.

Data analysis was conducted using both qualitative and quantitative approaches. Qualitative analysis involved expert judgment focusing on content, construction, and language validity using content validity ratio (CVR). The steps of the research procedure can be seen in Fig 1 below.

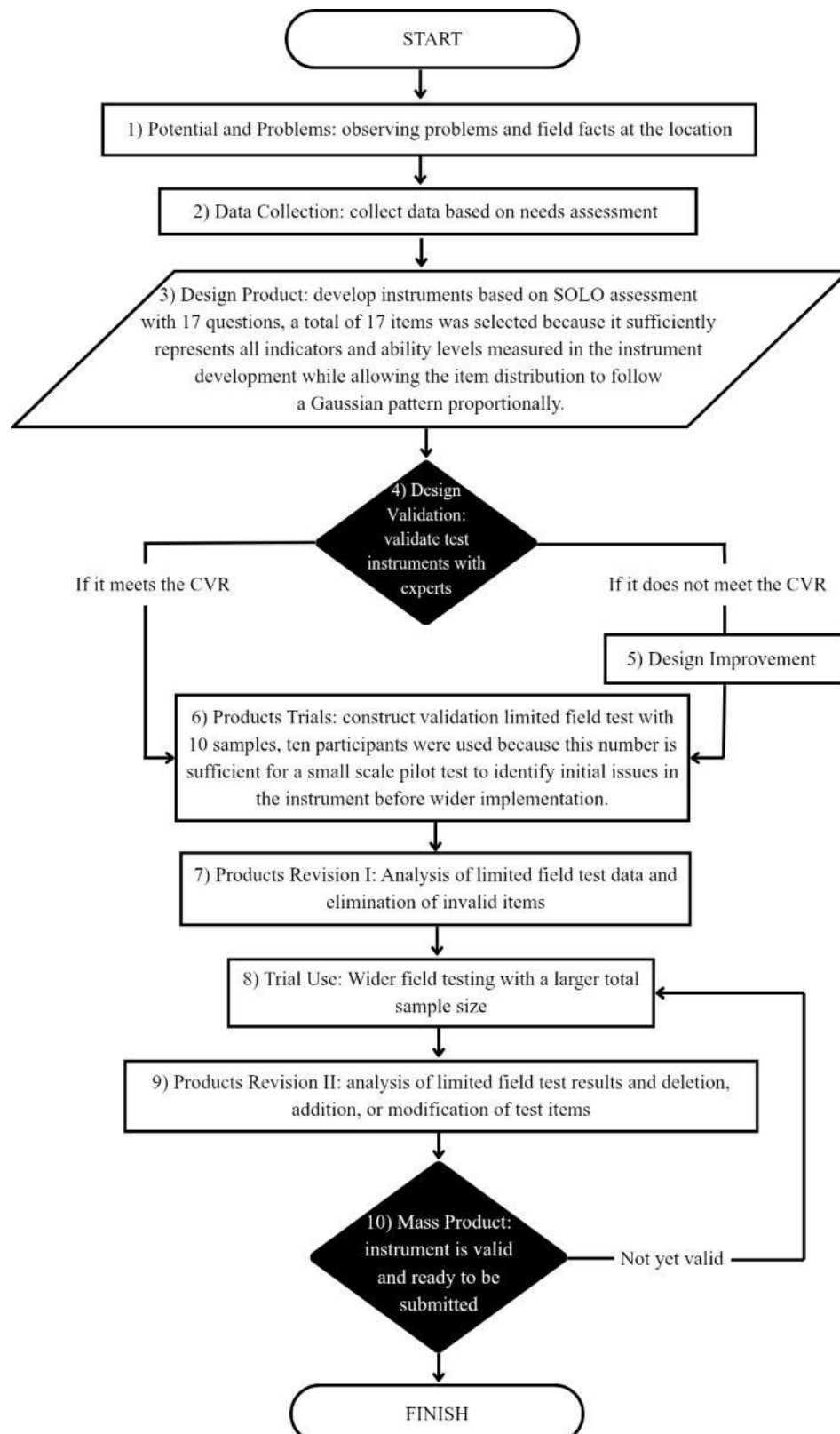


Fig 1. Research Procedure Steps

3. RESULTS AND DISCUSSION

In this instrument development study, the stages were carried out following the Borg and Gall procedure, starting from identifying potential and problems to producing a final product that was feasible for use. At the initial stage (potential and problem), it was found that the school already had assessment instruments to measure students' cognitive understanding; however, these instruments did not specifically classify the structure of understanding based on the SOLO Taxonomy levels. Therefore, an additional instrument was needed as a complement to provide a more structured representation of students' levels of understanding. At the data collection stage, literature studies and needs analysis were conducted to develop indicators and test blueprints.

A. Validation Results by Experts

At the product design stage, an initial instrument consisting of 17 items based on SOLO levels was developed. The instrument then underwent content validation by two experts, resulting in revisions that reduced the items to 14, all of which were deemed suitable for the trial. The product trial stage was conducted through the first preliminary field testing involving 10 students, and the results of the analysis showed that 8 items were invalid, namely items number 1, 2, 3, 5, 8, 9, 10, and 13, that have been studied by experts are processed using the CVR index with a scale of 1, 2, and 3, where if the conclusion of E/ Essential (can be used without revision) is on a scale of 3, ET/ Almost Essential (can be used with minor revisions) on a scale of 2, and T/ Not Essential (rejected) on a scale of 1.

If only 1 validator accepts:

$$CVR = \frac{n_e - \frac{N}{2}}{\frac{N}{2}} = \frac{1 - \frac{2}{2}}{\frac{2}{2}} = 0$$

If there are 2 validators who accept:

$$CVR = \frac{n_e - \frac{N}{2}}{\frac{N}{2}} = \frac{2 - \frac{2}{2}}{\frac{2}{2}} = 1$$

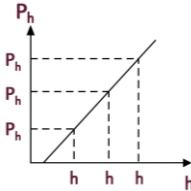
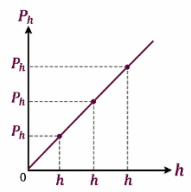
Because the calculation is valid based on the formula, there must be 2 validators who accept it so that one question item is declared valid. The categories of test item analysis results qualitatively presented in Table 1.

Table 1. Item Analysis Results Qualitatively

Category	Question Number	Questions Total
Accepted	1, 12, 14, 15, 16, 17.	6 quesrions
Revised	2, 3, 4, 5, 7, 8, 9, 10.	8 questions
Rejected	6, 11, 13.	3 questions

There were 3 items that did not meet the required criteria (rejected), namely item number 6 at SOLO level 5, as well as items number 11 and 13 at SOLO level 4. Thus, from the total items developed, 14 items remained and were considered suitable for use, with a distribution of 2 items at SOLO level 2, 5 items at SOLO level 3, 3 items at SOLO level 4, and 4 items at SOLO level 5. The rejection of these items was based on expert validation results. Item number 6 was rejected because the image provided was not consistent with the relevant theoretical concept, which could potentially lead to misconceptions. Meanwhile, items number 11 and 13 were rejected because they were considered illogical in terms of context and problem structure, making them unsuitable as instruments for measuring students' cognitive abilities. Therefore, these items were eliminated to ensure that the instrument met quality standards and could accurately assess students' understanding. After this stage, the researcher has 14 questions left that will be tested on students empirically with the order of the question numbers raised and rearranged after there are three questions that are deleted. Table 2 shows examples of question instruments before and after the revision.

Table 2. Revision of Test Instruments After Expert Validation

Initial Questions	Suggestions for Improvement	Questions After Revision
Grafik hubungan antara tekanan hidrostatik (P_h) dan kedalaman (h) disajikan sebagai berikut.  Berdasarkan grafik di atas, disimpulkan bahwa: 1. Antara tekanan hidrostatik (P_h) dan kedalaman (h) sebanding. 2. Tekanan hidrostatik besar (P_h) relatif terhadap lurus ke kedalaman (h). 3. Hubungan tekanan hidrostatik (P_h) dengan kedalaman (h) adalah linier. 4. Hubungan tekanan hidrostatik (P_h) dengan kedalaman (h) tidak linier. <i>Opsi:</i> A. Pernyataan 1, 2, dan 3 adalah benar. Pernyataan 4 salah. B. Pernyataan 1 dan 3 benar. Pernyataan 2 dan 4 adalah salah. C. Pernyataan 1 dan 3 adalah salah. Pernyataan 2 dan 4 adalah benar D. Pernyataan 4 benar. E. Tidak ada pernyataan yang benar.	Fix the graph so that the lines are linear from point 0, change statement 1 to make it more difficult.	Grafik hubungan antara tekanan hidrostatik (P_h) dan kedalaman (h) disajikan sebagai berikut.  Berdasarkan grafik di atas, disimpulkan bahwa: 1. Tekanan yang dialami suatu titik akan semakin besar jika titik tersebut berada pada kedalaman yang lebih tinggi dari permukaan fluida.. 2. Tekanan hidrostatik besar (P_h) berbanding lurus dengan kedalaman (h). 3. Hubungan tekanan hidrostatik (P_h) dengan kedalaman (h) adalah linier. 4. Hubungan tekanan hidrostatik (P_h) dengan kedalaman (h) tidak linier. Berdasarkan pernyataan di atas, manakah opsi yang benar di bawah ini? <i>Opsi:</i> A. Pernyataan 1, 2, dan 3 benar. Pernyataan 4 salah. B. Pernyataan 1 dan 3 benar. Pernyataan 2 dan 4 salah. C. Pernyataan 1 dan 3 salah. Pernyataan 2 dan 4 benar D. Pernyataan 4 benar. E. Tidak ada pernyataan yang benar.

B. Validity Pearson Correlation

In line with classical test theory, item validity was determined by comparing the calculated r -value (r -count) against the critical r -table value, alongside a significance threshold of $p < 0.05$. Following the first preliminary field testing, the analysis revealed that seven items had a very low correlation, two items had a low correlation, three items had a moderate correlation, and two items had a high correlation. Based on these results, the poorly performing items were removed, and a renumbering process was conducted to ensure that the deleted items did not reappear in subsequent analyses. The full validity interpretation for this first preliminary field testing is presented in Table 3.

Table 3. The interpretation validity for first preliminary field testing

Number of Questions	Pearson	Nilai Sig.	Summary	Interpretation
Q1	0	0	Not Valid	Very Low
Q2	-0.155	0.699	Not Valid	Very Low

Q3	0.203	0.573	Not Valid	Low
Q4	0.583	0.077	Valid	Enough
Q5	0	0	Not Valid	Very Low
Q6	0.62	0.056	Valid	High
Q7	0.785	0.007	Valid	High
Q8	-0.155	0.669	Not Valid	Very Low
Q9	0.155	0.669	Not Valid	Very Low
Q10	0	0	Not Valid	Very Low
Q11	0.807	0.005	Valid	Very Low
Q12	0.581	0.078	Valid	Enough
Q13	0.349	0.323	Not Valid	Low
Q14	0.426	0.219	Valid	Enough

The preliminary field testing that was carried out at first still had many invalid items, namely questions number 1, 2, 3, 5, 8, 9, 10, and 13. Therefore these questions are revised and then retested. However, because the limited test only used a small subject of 10, the average significant result could be above 0.05. The resulting correlation values are compared with the r table. The interpretation validity result for second preliminary field testing in Table 4.

Table 4. The interpretation validity for Second preliminary field testing

Number of Questions	Pearson	Nilai Sig.	Summary	Interpretation
Q1	0.583	0.077	Valid	Enough
Q2	0.653	0.041	Valid	High
Q3	0.515	0.128	Valid	Enough
Q4	0.730	0.016	Valid	High
Q5	0.515	0.128	Valid	Enough
Q6	0.730	0.016	Valid	High
Q7	0.620	0.056	Valid	High
Q8	0.408	0.242	Valid	Enough
Q9	0.467	0.173	Valid	Enough
Q10	0.454	0.188	Valid	Enough
Q11	0.515	0.128	Valid	Enough
Q12	0.678	0.031	Valid	High
Q13	0.515	0.128	Valid	Enough
Q14	0.675	0.032	Valid	High

The revision carried out was able to improve the quality of the instrument, so that the instrument was declared suitable for collecting field test research data more widely, the main field testing used 14 questions that had been revised based on the results of preliminary field testing. A main field testing was carried out in class XI with 60 subjects. The interpretation validity result for main field testing in table 5.

Table 5. The interpretation validity for Main Field Testing

Number of Questions	Pearson	Nilai Sig.	Summary	Interpretation
Q1	0.547	0.000	Valid	Enough
Q2	0.584	0.000	Valid	Enough
Q3	0.692	0.000	Valid	High
Q4	0.815	0.000	Valid	High
Q5	0.670	0.000	Valid	High
Q6	0.621	0.000	Valid	High
Q7	0.657	0.000	Valid	High
Q8	0.719	0.000	Valid	High

Q9	0.626	0.000	Valid	High
Q10	0.347	0.007	Not Valid	Low
Q11	0.480	0.000	Valid	Enough
Q12	0.795	0.000	Valid	High
Q13	0.179	0.172	Not Valid	Very Low
Q14	0.811	0.000	Valid	High

Based on the results of the field test analysis of 14 questions, validity test showed that 12 questions were declared valid with correlation coefficients in the category of sufficient to high (0.480–0.815), while two items, namely Q10 and Q13, were invalid because they had low to very low coefficients.

C. Reliability Cronbach's Alpha

According to Heale and Twycross (2015), an instrument is considered reliable if its Cronbach's alpha coefficient is ≥ 0.70 . Based on the first preliminary field testing, the SPSS analysis yielded an alpha value of 0.509, indicating that the initial items were unreliable. However, after revisions were made and the second preliminary field test was conducted, the Cronbach's alpha rose to 0.845, demonstrating acceptable reliability. This reliability was sustained into the main field testing, which produced a highly consistent alpha value of 0.876.

D. Distinguishing Power

The analysis of the distinguishing power of item question aims to distinguish between students who have high ability and students who have low ability, or to distinguish between the upper and lower groups (Çelen & Aybek, 2022; Rosyad, 2024). The results of the analysis of the distinguishing of test instruments in the first preliminary field testing based on the interpretation of the differential power index where 3 very good items (numbers 3, 7, and 11), 6 items are enough (numbers 4, 6, 8, 12, 13, and 14), 5 bad items (numbers 1, 2, 5, 9, and 10) as show in Table 6.

Table 6. Distinguishing Power for First Preliminary Field Testing

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
0	-0.2	0.4	0.2	0	0.2	0.4	0.2	-0.2	0	0.8	0.2	0.2	0.2

The results of the analysis of the distinguishing of test instruments in the second preliminary field testing where there are 8 very good questions (numbers 1, 2, 4, 6, 7, 8, 12, and 14), 3 enough questions (numbers 3, 5, and 9), and 3 bad questions (numbers 10, 11, and 13) as show in Table 7.

Table 7. Distinguishing Power for Second Preliminary Field Testing

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
0.4	0.8	0.2	0.8	0.2	0.8	0.4	0.6	0.2	0	0	0.6	0	0.8

The results for the analysis of the distinguishing test instruments after instrument revision in the main field testing based on the interpretation of the differential power index where there are 11 very good questions (numbers 1, 2, 3, 4, 5, 6, 7, 8, 9, 12, and 14), 1 question item was enough (number 10), and 2 items of bad questions (numbers 11 and 13) as show in Table 8.

Table 8. Distinguishing Power for Main Field Testing

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
0.81	0.75	0.81	0.93	0.75	0.68	0.81	0.87	0.87	0.25	0.18	0.62	0.18	0.62

E. Level of Difficulty

An ideal test item was one that possesses a moderate level of difficulty (Çelen & Aybek, 2022; Rosyad, 2024). In this study, the difficulty level (P) was classified as moderate if it fell within the range of $0.31 < P \leq 0.70$. Based on this interpretation, the first preliminary field testing revealed 4 difficult items (numbers 5, 9, 10, and 13), 2 moderate items (numbers 11 and 12), and 8 easy items (numbers 1, 2, 3, 4, 6, 7, 8, and 14) as show in Table 9.

Table 9. Level of Difficulty for First Preliminary Field Testing

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
1.000	.900	.800	.700	.000	.900	.800	.900	.100	.000	.600	.500	.100	.900

Following revisions, the second preliminary field testing showed a shift in the difficulty distribution, resulting in 1 difficult item (number 3), 10 moderate items (numbers 2, 4, 5, 6, 7, 8, 9, 10, 12, and 14), and 3 easy items (numbers 1, 11, and 13) as show in Table 10.

Table 10. Level of Difficulty for Second Preliminary Field Testing

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
.800	.400	.100	.600	.500	.600	.600	.700	.700	.600	.800	.500	.800	.600

Finally, the main field testing produced a distribution of 2 difficult items (numbers 10 and 13), 8 moderate items (numbers 1, 2, 3, 5, 6, 7, 8, and 9), and 4 easy items (numbers 4, 11, 12, and 14) as show in Table 11.

Table 11. Level of Difficulty for Main Field Testing

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
.566	.483	.583	.700	.666	.583	.583	.633	.350	.116	.950	.800	.083	.816

F. Data Analysis After Item Deleted

Based on the results of the field test analysis of 14 questions, the quality of the instrument in general was classified as good because it has met most of the criteria of validity, reliability, differentiation, and difficulty. The results of the validity test showed that 12 questions were declared valid with correlation coefficients in the category of sufficient to high (0.480–0.815), while two items, namely Q10 and Q13, were invalid because they had low to very low coefficients. Cronbach's Alpha reliability value of 0.887 indicates that the instrument has good internal consistency and suitable for use as a measuring tool. In terms of differentiation, most of the items are in the very good category so that they are able to distinguish between students with high and low ability, but Q10 has enough differentiation and Q11 and Q13 have low differentiation. Based on the level of difficulty, the majority of the questions are in the medium category which shows that the composition of the questions is quite proportional, although there are some easy questions (Q4, Q11, Q12, Q14) and difficult questions (Q10 and Q13) that need to be adjusted to make the distribution of difficulty levels more balanced. Overall, the instrument can be declared suitable for use for research data collection, but items Q10 and Q13 can be deleted because they do not meet the valid category so that the quality of the instrument becomes more optimal and are able to measure students' abilities more accurately, as well as low differentiation, this question are also classified as too difficult and almost impossible for students to answer, therefore the following are the results of validity, reliability, distinguishing power, and difficulty level for the fix question after the question item was deleted in table 12.

Table 12. Data Analysis After Item Deleted

No	Level SOLO	Validity	Reliability	Distinguishing Power	Difficulty Level
1	S4	0.547	0.887	0.75	0.5667
2	S5	0.584		0.75	0.4883
3	S4	0.692		0.875	0.5833
4	S3	0.815		0.875	0.7000
5	S4	0.670		0.75	0.6667
6	S2	0.621		0.75	0.5833
7	S5	0.657		0.75	0.5833
8	S3	0.719		0.875	0.6333
9	S5	0.626		0.875	0.3500
11	S5	0.480		0.1875	0.9500
12	S2	0.795		0.6875	0.8000

14	S3	0.811	0.6875	0.8167
----	----	-------	--------	--------

Therefore, it can be concluded that the total number of questions that are suitable for use is 12 items with a distribution of 2 questions at the SOLO 2 (unistructural) level, 3 questions at the SOLO 3 level (multistructural), 3 questions at the SOLO 4 (relational) level, and 4 questions at the SOLO 5 (extended abstract) level.

The results indicate that the SOLO Taxonomy-based assessment instrument developed in this study demonstrates satisfactory psychometric quality and can be used to assess students' conceptual understanding of fluid mechanics. The finalized instrument comprised 12 valid items, with validity coefficients ranging from 0.480 to 0.815 and a reliability coefficient of 0.887. These findings demonstrate that the instrument has adequate consistency in measuring the intended construct and can provide reliable information about students' levels of understanding. Biggs and Collis (1982) explained that assessment using the SOLO Taxonomy is intended to examine not only whether students arrive at correct answers but also the complexity and structure of the thinking underlying those responses. In this regard, the acceptable validity and reliability obtained in the present study indicate that the instrument was able to identify differences in students' cognitive development across the SOLO levels.

Further evidence of the instrument's quality is provided by the item discrimination results. The mean discrimination index of 0.734 demonstrates that the items were generally effective in distinguishing students with higher conceptual understanding from those with lower understanding. This result supports the findings of Mindayani et al. (2019), who stated that SOLO-based assessment should be able to differentiate students' cognitive performance because each SOLO category reflects a distinct level of conceptual development. The relatively high discrimination values obtained for most items indicate that the instrument was sufficiently sensitive to differences in students' mastery of fluid concepts, making it useful for both diagnostic and evaluative assessment.

The SOLO analysis revealed that the multistructural level (S3) was the most dominant, while the percentages at the relational (S4) and extended abstract (S5) levels were comparatively lower. This distribution provides an important description of how students understood fluid concepts. In general, students were able to recognize several relevant concepts, equations, and variables, but they were not always able to establish meaningful connections among them. Biggs and Collis (1982) described multistructural thinking as a stage in which learners can identify or use several relevant elements but have not yet integrated them into a coherent whole. Thus, students may demonstrate knowledge of different aspects of a topic while their overall conceptual structure remains fragmented.

The predominance of the multistructural level also indicates that students had developed considerable factual and procedural knowledge but had not consistently progressed toward relational reasoning. Kaharuddin and Hajeniati (2020) similarly found that students were often able to identify relevant information but experienced difficulties when they had to establish systematic relationships between concepts. Mulbar et al. (2017) likewise reported that learners may remain at intermediate SOLO levels when their learning is dominated by procedural approaches rather than conceptual integration. In the present study, students appeared to understand individual principles involving fluid pressure, buoyancy, and fluid equilibrium, but difficulties became more apparent when several principles had to be combined to explain a more complex physical situation.

The results at the relational and extended abstract levels nevertheless provide a positive indication. The difference between S4 (61%) and S5 (59%) was relatively small, suggesting that students who were able to establish relationships among concepts also showed potential to extend and generalize their understanding to new situations. Fernandez and Guzon (2025) emphasized that relational understanding can function as an important foundation for developing higher levels of abstraction because students who successfully integrate concepts are better positioned to transfer their knowledge across different contexts. Consistent with this perspective, Tian et al. (2024) found that progression from relational to extended abstract thinking is associated with the ability to construct generalized explanations rather than depending primarily on memorized procedures.

Despite the relatively favorable overall results, the decreasing achievement across the higher SOLO levels indicates that students continued to encounter difficulties with increasingly complex forms of thinking. Such a pattern is consistent with the hierarchical nature of the SOLO Taxonomy, in which each successive level requires greater cognitive complexity. Ahmad et al. (2024) explained that

movement between SOLO levels represents a progression from basic recognition and identification toward integrated and abstract reasoning. The findings of this study therefore suggest that students faced greater challenges when they were required to synthesize information, develop explanations, and apply their understanding beyond familiar situations.

These difficulties may also be related to the conceptual characteristics of fluid mechanics. Previous research has identified fluid concepts as a challenging area in physics learning. Hutasoit et al. (2025), for example, identified misconceptions involving hydrostatic pressure, Pascal's principle, buoyancy, and fluid equilibrium. Atmaja and Samsudin (2024) similarly reported that misconceptions may remain even after students have received formal instruction, potentially limiting their ability to establish meaningful relationships among concepts. As a result, students may be able to solve problems involving individual principles while struggling with tasks that require several principles to be considered simultaneously.

Similar conceptual difficulties have been reported in studies of both static and dynamic fluids. Kurniawan (2023) found that students could sometimes use appropriate formulas for static fluid problems despite having only a partial conceptual understanding of the underlying principles. Fatmawati et al. (2025) also reported a discrepancy between students' procedural performance and their deeper conceptual mastery. In a similar manner, Muki et al. (2025) observed that students' understanding of fluid concepts could remain superficial and fragmented, particularly when learning activities placed greater emphasis on mathematical procedures than on conceptual reasoning.

The challenges identified in this study are also consistent with findings from contexts outside Indonesia. Koto and Ilhami (2023) reported that secondary students continued to encounter conceptual difficulties in fluid dynamics following online learning during the pandemic. Putri et al. (2023) likewise identified substantial misconceptions in dynamic fluid concepts through the use of a four-tier diagnostic test. Taken together, these findings indicate that difficulties in understanding fluid concepts are not restricted to the students involved in the present study but represent a broader issue in physics education.

Differences between the upper and lower groups provide further evidence of variation in students' conceptual understanding. Students in the upper group demonstrated almost complete mastery of the concepts assessed, whereas the lower group obtained correct responses on fewer than one-quarter of the items. This considerable performance gap indicates that students did not develop conceptual understanding at the same level. Hasan (2017) stated that students functioning at higher SOLO levels tend to possess more interconnected cognitive structures, allowing them to interpret a problem from a more comprehensive perspective. Conversely, learners at lower SOLO levels tend to concentrate on individual pieces of information and have greater difficulty identifying relationships between concepts. The findings obtained in this study are consistent with this theoretical explanation.

The item discrimination results also contribute to the interpretation of students' conceptual profiles. Items with particularly high discrimination values were predominantly associated with questions that required conceptual reasoning and the integration of several ideas. This observation is consistent with Hartanti et al. (2021), who found that higher-order items based on the SOLO Taxonomy can more effectively differentiate students according to the complexity of their thinking. Claudia et al. (2020) similarly reported that tasks involving relational and extended abstract reasoning are more capable of revealing differences in conceptual understanding than items primarily requiring simple recall.

From an instructional perspective, the dominance of multistructural thinking indicates the need for greater attention to learning approaches that promote deeper conceptual understanding. Sabani et al. (2026) described deep learning as a process in which students actively establish relationships between concepts, critically evaluate information or evidence, and apply their knowledge flexibly in different contexts. The present findings suggest that many students had not yet fully achieved this form of learning because their knowledge remained relatively compartmentalized. Instruction should therefore move beyond the accumulation of factual information and provide greater opportunities for students to connect and integrate related concepts.

Inquiry-based learning is one approach that may support this development. Damopolii et al. (2020) found that inquiry-oriented learning can improve students' SOLO levels by encouraging them to investigate relationships between concepts and develop explanations independently. Such learning experiences can provide opportunities for students to engage in reasoning processes associated with the

relational and extended abstract levels. Accordingly, the incorporation of inquiry activities, problem-based tasks, conceptual discussions, and reflective questioning into physics instruction may help students develop more integrated and sophisticated understanding.

The findings are also relevant to the broader context of international science education. The OECD (2023) PISA report indicated that Indonesian students continue to face challenges in applying scientific knowledge to unfamiliar and complex situations. The dominance of multistructural responses in the present study reflects a related pattern, in which students demonstrate knowledge of individual concepts but experience greater difficulty when required to transfer and apply that knowledge in new contexts. These findings reinforce the importance of strengthening higher-order thinking through instructional practices that emphasize conceptual understanding and the application of knowledge.

Overall, the developed SOLO Taxonomy-based assessment instrument provides detailed information about students' conceptual understanding of fluid mechanics and the structure of their thinking. Rather than merely indicating whether students answered correctly, the instrument was able to reveal differences in the complexity of their understanding across SOLO levels. The predominance of the multistructural level indicates that students possessed a substantial foundation of conceptual knowledge but still required support to connect, integrate, and extend that knowledge. Future instructional efforts should therefore emphasize relational and extended abstract thinking through inquiry-based, conceptually oriented, and reflective learning activities that enable students to construct deeper and more transferable scientific understanding.

4. CONCLUSION

This development research successfully produced a SOLO Taxonomy-based diagnostic assessment framework on fluid mechanics for high school students using the revised Borg and Gall model. The development involved rigorous expert validation and field testing with 80 students. The results indicated that 14 items were evaluated, with 12 items (85.71%) declared valid and 2 items (14.29%) rejected based on validity, reliability, discrimination power, and difficulty level criteria. The final instrument demonstrated high reliability (0.887). Furthermore, most items exhibited 'very good' discrimination power (78.57%), complemented by a balanced difficulty distribution comprising easy (28.57%), moderate (57.14%), and difficult (14.29%) items.

REFERENCES

- Ahmad, F. A. K., Meshkat, M., & Taghipour, M. (2024). Using SOLO model for evaluation of students' physics learning. *Physics Journal Farhangian University*, 1(1), 55–65. <https://doi.org/10.48310/esip.2024.17397.1006>
- Atmaja, D. Y. S., & Samsudin, A. (2024). Are there misconceptions in my class? Misconception analysis with T3-SF (three tier-test on static fluid). *Tarbiyah: Jurnal Ilmiah Kependidikan*, 13(2), 155–170. <https://doi.org/10.18592/tarbiyah.v13i2.13518>
- Biggs, J. B & Collis, K. F. (1982). *Evaluating the Quality of Learning: The SOLO Taxonomy (Structure of The Observed Learning Outcome)*. Academic Press.
- Biggs, J. (1987). *Student Approaches to Learning and Studying*. Australian Council for Educational Research.
- Biggs, J., & Tang, C. (2011). *Teaching for Quality Learning at University (4th ed)*. McGraw-Hill Education.
- Brookhart, S. M. (2010). *How to Assess Higher-Order Thinking Skills in Your Classroom*. ASCD.
- Çelen, Ü., & Aybek, E. C. (2022). A novel approach for calculating the item discrimination for Likert type of scales. *International Journal of Assessment Tools in Education*, 9(3), 772–786. <https://doi.org/10.21449/ijate.1173356>
- Claudia, L. F., Kusmayadi, T. A., & Fitriana, L. (2020). The SOLO taxonomy: Classify students' responses in solving linear program problems. *Journal of Physics: Conference Series*, 1538(1), Article 012107. <https://doi.org/10.1088/1742-6596/1538/1/012107>
- Damopolii, I., Nunaki, J. H., Nusantara, E., & Kandowangko, N. Y. (2020). The effectiveness of inquiry-based learning to train students' thinking skill based on SOLO taxonomy. *Journal of Physics: Conference Series*, 1567(4), Article 042025. <https://doi.org/10.1088/1742-6596/1567/4/042025>

- Eryanto, D., & Prasetyono, H. (2025). Efektivitas pembelajaran mendalam dalam implementasi kurikulum merdeka di SD Negeri Cikampek Selatan I. *Jurnal Evaluasi Pendidikan*, 16(2). <https://doi.org/10.21009/jep.v16i2.60722>
- Fatmawati, W., Parno, & Diantoro, M. (2025). Identification of students' mastery level of physics concepts in static fluid material. *Jurnal Pendidikan Fisika dan Teknologi*, 11(1), 26–33. <https://doi.org/10.29303/jpft.v11i1.8360>
- Fernandez, P. J. M., & Guzon, A. F. H. (2025). A SOLO taxonomy-based rubric for assessing conceptual understanding in applied calculus. *Journal on Mathematics Education*, 16(2), 559–580. <https://doi.org/10.22342/jme.v16i2.pp559-580>
- Hartanti, D., Djudin, T., & Mursyid, S. (2021). Analisis tingkat berpikir dalam menyelesaikan soal gerak parabola menggunakan taksonomi Structure of the Observed Learning Outcomes (SOLO) siswa kelas X MIA SMAN 1 Sungai Raya. *Jurnal Inovasi Penelitian dan Pembelajaran Fisika*, 2(1), 1–9. <https://doi.org/10.26418/jippf.v2i1.44591>
- Hasan, B. (2017). Karakteristik respon peserta didik dalam menyelesaikan soal geometri berdasarkan taksonomi SOLO. *JINoP (Jurnal Inovasi Pembelajaran)*, 3(1), 449–458. <https://doi.org/10.22219/jinop.v3i1.4282>
- Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence-Based Nursing*, 18(3), 66–67. <https://doi.org/10.1136/eb-2015-102129>
- Hutasoit, Y., Nahor, D. B., Sianturi, E. S., Sati, L., & Simatupang, N. T. (2025). Studi literatur miskonsepsi siswa pada materi fluida dan tekanan zat cair. *Sindoro: Cendikia Pendidikan*, 18(1). <https://doi.org/10.9644/sindoro.v3i9.252>
- Indrasari, N. (2025). Analysis of senior high school students' science literacy in understanding static fluids. *Journal of Advances in Physics Education*, 1(1), 40–47. <https://doi.org/10.23960/jape.v1i1.376>
- Inggit, S. M., Liliawati, W., & Suryana, I. (2021). Identifikasi miskonsepsi dan penyebabnya menggunakan instrumen Five-Tier Fluid Static Test (5TFST) pada peserta didik kelas XI sekolah menengah atas. *Journal of Teaching and Learning Physics*, 6(1), 49–68. <https://doi.org/10.15575/jotalp.v6i1.11016>
- Kementerian Pendidikan Dasar dan Menengah. (2025). *Naskah akademik pembelajaran mendalam (deep learning) menuju pendidikan bermutu untuk semua*. Kemendikdasmen.
- Kaharuddin, A., & Hajeniati, N. (2020). An identification of students' responses based on SOLO taxonomy in mathematics learning toward learning activities and learning outcomes. *Al-Jabar: Jurnal Pendidikan Matematika*, 11(2), 191–200. <https://doi.org/10.24042/ajpm.v11i2.6270>
- Koto, I., & Ilhami, D. (2023). High-school students' conceptual understanding of fluid dynamics following online learning during the coronavirus pandemics. In *Proceedings Mathematics and Science Education International Seminar* (pp. 117–125). Universitas Negeri Semarang.
- Kurniawan, R. V. (2023). Analisis pemahaman konsep siswa pada topik fluida statis. *TEACHING: Jurnal Inovasi Keguruan dan Ilmu Pendidikan*, 3(1), 67–73. <https://doi.org/10.51878/teaching.v3i1.2165>
- Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I—Outcome and process. *British Journal of Educational Psychology*, 46(1), 4–11. <https://doi.org/10.1111/j.2044-8279.1976.tb02980.x>
- Mindayani, N., Hadi, W., & Ambarita, B. (2019). The development of SOLO taxonomy based assessment tool on text of observation reports of X grade students of senior high school 1 NA. IX-X. *Budapest International Research and Critics in Linguistics and Education (BirLE) Journal*, 2(4), 371–381. <https://doi.org/10.33258/birle.v2i4.527>
- Muki, B. G., Aviyanti, L., & Nawas, A. (2025). Investigation of students' conceptual understanding of fluids in Kupang City and Sabu Raijua District. *Momentum: Physics Education Journal*, 9(1), 144–162. <https://doi.org/10.21067/mpej.v9i1.11056>
- Mulbar, U., Rahman, A., & Ahmar, A. S. (2017). Analysis of the ability in mathematical problem-solving based on SOLO taxonomy and cognitive style. *World Transactions on Engineering and Technology Education*, 15(1), 68–73. <https://doi.org/10.26858/wtete.v15i1y2017p6873>
- OECD. (2023). *PISA 2022 Results Factsheets Indonesia*. OECD Publishing
- Ole, F. C. B., & Gallos, M. R. (2023). Impact of formative assessment based on feedback loop model on high school students' conceptual understanding and engagement with physics. *Journal of Turkish Science Education*, 20(2), 333–355. <https://doi.org/10.36681/tused.2023.019>

- Putri, R. A., Permana, H., & Nasbey, H. (2023). Identifikasi miskonsepsi materi fluida dinamis dengan menggunakan tes diagnostik four-tier untuk siswa SMA kelas XI. *Joint Prosiding IPS dan Seminar Nasional Fisika*, 11(1), PF181–PF186. <https://doi.org/10.21009/03.1102.pf25>
- Sabani., Panggabean, D. D., & Ramadhani, I. (2026). *Asesmen Pembelajaran Mendalam pada Bidang Studi Fisika*. PT MAFY Media Literasi Indonesia.
- Sedhandoyo, C. (2024). *Pembelajaran Poeaw Berbantuan Simulasi Phet untuk Meningkatkan Pemahaman Konsep Siswa SMA pada Materi Fluida Dinamis*. Universitas Pendidikan Indonesia.
- Sugiyono. (2023). *Metode Penelitian Kuantitatif Kualitatif*. Alfabeta.
- Tian, P., Fan, Y., Sun, D., & Li, Y. (2024). Evaluating students' computation skills in learning amount of substance based on SOLO taxonomy in secondary schools. *International Journal of Science Education*, 46(15), 1578–1600. <https://doi.org/10.1080/09500693.2023.2291691>
- UNESCO. (2021). *Reimagining Our Futures Together: A New Social Contract for Education*. UNESCO.
- Yuberti, Y., Suryani, Y., & Kurniawati, I. (2020). Four-tier diagnostic test with certainty of response index to identify misconception in physics. *Indonesian Journal of Science and Mathematics Education*, 3(2), 245–253. <https://doi.org/10.24042/ij sme.v3i2.6061>