Calibration for Instrument Argumentation Skills on the Subject of Fluid Statics Using Item Response Theory

Harry Affandy 1, Dewanta Arya Nugraha 1, Scundy Nourma Pratiwi 2 and C. Cari 3

1 Doctorate Program of Science Education, Universitas Sebelas Maret, INDONESIA
2 Physics Education, Master Program, Universitas Sebelas Maret, INDONESIA
3 Physics Department, Universitas Sebelas Maret, INDONESIA

E-mail: *cari@staff.uns.ac.id

Abstract. This study aims to calibrate the instrument of argumentation skills on the subject of Fluid Statics using item response theory (IRT). The blueprint instrument developed based on the aspects and sub-aspects of argumentation skills, totaling 20 items. The scoring uses the Partial Credit Model (PCM) principle in 4 categories. Subjects involved in the study were 206 undergraduate students at two different universities. The validation of the instrument using expert judgment techniques, which consisted of linguists, physicists, and educational measurement experts. Instrument calibration for argumentation skills is carried out by (1) analysis of item fit with PCM by testing goodness of fit items, (2) testing prerequisite analysis, (3) reliability, (4) difficulty level, and (5) Standard Error of Measurement (SEM). The test developed is valid with Aiken V index values in the range 0.93 to 1. The analysis of IRT assumptions uses two parameters, the adequacy test of the sample and the factor analysis. The Chi-Square value in the Bartlett test is 2202.67 with df 190, and a p-value value (Sig.) <0.05 indicates that the sample size used has met the unidimensional requirements. The instrument developed has met the criteria for scoring the Polytomous items, based on a) the MNSQ INFIT value is in the range of 0.70 to 1.30, and b) the difficulty index test is in the range -2.00 <b <2.00. The results of the reliability estimate on the test equipment obtained a reliability value of 0.86, which means that the reliability of the test is in the high category. The graph of the information function and SEM converge on the ability scales of -2.8 and +1.5, meaning that the skills test kit argues it is suitable for testees in the low to moderate ability range.

1. Introduction

Science and technology developed through education are long-term resources that are very useful for sustainable development [1]. Today the world is in an era of globalization, where competition is quite fierce, namely competition for the quality of human resources (HR). The level of education supports a good quality of human resources. The development of science and technology can encourage a more applicable and exciting learning process as an effort to improve the quality of education [2]–[4]. Efforts to improve the quality of knowledge can pursue by increasing the quality of the assessment [5].

Assessment is the process of determining the value of an object [6], [7]. In conducting the evaluation, there must be an object that is the target of the evaluation. In education, the item that uses as an assessment is student learning outcomes. To be able to assess student learning outcomes requires a measure or criterion; for example, it can be evaluated as good, moderate, it is not necessary to have
explicit provisions or standards of what is right, middle, and what is lacking. Assessment requires data of good quality so that it needs to be supported by a fair assessment process [5].

The assessment process can provide information on the effectiveness of a learning process so that it can be used as material to improve the learning process [8]. The results of the assessment can also be used as feedback of lecturers to enhance learning activities [9], [10]. Based on the results of a preliminary study by interviewing a necessary physics lecturer in the Physics Study Program of Sebelas Maret University, the assessment given to students is an assessment of learning outcomes. Assessment of learning outcomes carried out on the aspect of knowledge using an assessment instrument in the form of an essay/description. Assessment of learning outcomes carried out to determine and make decisions about student success in achieving predetermined competencies [3], namely about understanding student concepts [11].

Students experience difficulties in understanding concepts due to several factors, namely the lack of supporters in mastering concepts, the learning process, and environmental factors. Conceptual understanding is considered a complex phenomenon, namely factual, procedural, and conditional knowledge [12]. One of the physics courses that require a lot of experience of concepts and describes physics concepts in everyday life is static fluid material. Fluid Statics material is a subject related to everyday life. However, in reality, it is still difficult for students to understand. After all, there are misconceptions from the start. Students think that when the fish are in the cave, they get tremendous pressure because they are closed compared to fish outside the cave.

Physics is a basic science from a branch of science that base on experimental observation and quantitative measurement [13]. Experimental observation is an effort to understand the concepts in fundamental physics courses so that students can use thinking skills to be able to explain natural phenomena based on understanding these concepts [12]. Understanding the concept in question emphasizes the ability of students to apply scientific phenomena learned in everyday life in terms of obtaining information about a concept and the role of argumentation skills to strengthen information [14].

Argumentation skills are activities to facilitate understanding of cognitive activities in building scientific knowledge [14]. Good and complex student arguments will gain a good understanding of the concept. Students' scientific opinions can develop a better understanding of scientific concepts and processes so that students can understand scientific phenomena that exist in everyday life [15].

Osborne et al. [15], explain the difficulty of lecturers in assessing proper arguments. Science knowledge that defines in class is sometimes consider something solid, and a fact students cannot debate that. Toulmin's [16] argumentation scheme consists of 1) claim as the center of the argument; 2) data supporting claims; 3) warrant, the relationship between data and claim. 4) backing, quality, and type of reason; 5) qualifiers, statements assigning properties, limits, or explicit conditions that are part of the claim; 6) rebuttal, identifying exceptions to alternative claims or claims. The quality of the argumentation base on Toulmin's argumentation model. This is indicating a weakness in empowering students' argumentation skills in learning. The results of research conducted by Viyanti, et al. [14], which states that students' argumentation skills are still low in education because students trainers to argue scientifically [17].

Argumentation skills can provide trainers by providing questions about argumentation skills in the assessment process. The process of integrating argumentation skills in the assessment cannot separate from the relationship between the concept and teacher performance in the classroom. The relationship between epistemological aspects supports this. They are perspectives in appreciating criticism and argumentation as an attempt to build scientific knowledge [18]. The design of the instrument for assessing student's argumentative skills on static fluid material use to guide students in explaining and proving science. So that it can diagnose the extent to which students understand the concept of fluid statics material [17]. The results of research by Wahdan et al. [17], state that understanding of student concepts is said to be good if students can answer questions by explaining and proving science based on data. Questions or assessment techniques are not only used to evaluate the teaching and learning process, but the
Assessment is one of a series of activities to improve quality, performance, or productivity in implementing a program [5], [19].

Assessment in education uses two kinds of test theory to meet the criteria as a useful measuring tool, namely the classical test theory (Classical Test Theory) and modern measurement theory or item response theory (Item Response Theory: IRT) [20]. According to the classical test theory, the scoring of the test results base on the steps to answer an item correctly. The scoring model, which carries out per step by step, and the ability to be estimated by the raw score is not necessarily correct because the difficulty level of each step is not to take into account [21]. One of the scoring models that can use in the assessment is the polytomous scoring model.

The assumptions used in the polytomous scoring model developed to analyze the test items that require several steps of completion, known as the Partial Credit Model (PCM). The score on the PCM shows the number of steps to complete the item correctly. A higher category score indicates a greater ability than a lower category score. Thus PCM is suitable for use in learning outcome tests, including physics questions that require a problem identification stage to provide solutions or problem-solving. PCM is a development of the dichotomous Rasch model, which applies to the polytomous [8], [22].

Assessment in education as a scientific discipline is relatively new. In the scope of formal education starting in the 19th century, it practiced in schools [23]. Curriculum changes influence the current educational assessment paradigm. Therefore, lecturers are required to have adequate abilities, both conceptually and practically in the field of learning assessment to find out whether the students' understanding of concepts has a master or not.

2. Methods
The research subjects were second-semester students who were taking Physics courses. This research was conducted at two universities, as many as 206 undergraduate students from Universitas Sebelas Maret and Sunan Kalijaga State Islamic University. The preparation of the test blueprint becomes a guide for researchers in writing items so that they have a clear picture of the scope of the test material and the writing of the items will be well directed—the blueprint for the argumentation skills assessment instrument present in Table 1.

| Aspect of Argumentation Skills | Sub Aspects of Argumentation Skills | Hydrostatic Pressure | Archimedes Law | Pascal’s Law | Application in life | Number of items |
|-------------------------------|-------------------------------------|----------------------|----------------|--------------|---------------------|-----------------|
| Claim                         | Claims based on facts               | -                    | 1              | 10           | -                   | 4               |
|                               | Claims by definition                | -                    | -              | -            | -                   |                 |
|                               | Claims based on a causal relationship| -                    | -              | -            | 16                  |                 |
| Data                          | Data in the form of facts           | 2                    | -              | -            | -                   | 3               |
|                               | Data in the form of research results| -                    | 4              | -            | 9                   |                 |
| Warrant                       | General Warrant                     | 3                    | -              | -            | -                   | 4               |
|                               | Warrant as a guide                  | -                    | 7              | -            | -                   |                 |
|                               | Warrant as an analogy                | -                    | -              | 11           | 15                  |                 |
| Backing                       | Backing reinforces explanation      | 5                    | 20             | -            | -                   | 5               |
|                               | Backing strengthens the relationship of the statement | - | - | 17 | 12 |                 |
| Qualifier                     | Qualifier as attribute delimiter    | 8                    | -              | 18           | -                   | 4               |
|                               | Qualifier implies a statement specification | - | 14 | - | 13 |                 |
| Number of items               |                                     | 6                    | 5              | 4            | 5                   | 20              |
In a submerged object, the weight $\vec{w}$ is greater than the buoyancy force $\vec{F_B}$ and it can be written: the condition for the object to sink writes $\vec{w} > \vec{F_B}$. By considering that the conditions for a floating object are the same as for a floating object, that is, the weight of the object is equal to the buoyancy force ($\vec{w} = \vec{F_B}$) but has a difference in the volume of the object immersed in the fluid: a floating object is only partially immersed in the fluid so that $V_{bf} < V_b$, while the object is floating, all parts of the object are immersed in a fluid so that $V_{bf} < V_b$.

Eko assigns his students the task of investigating the blocks that will sink or float. The table of data on the results of student investigations is as follows:

|        | Mass (kg) | Volume (m$^3$) | Density (kg/m$^3$) |
|--------|-----------|----------------|---------------------|
| Water  | 0.97      | 1.03           | 1.0                 |
| Beam 1 | 0.53      | 0.13           | 0.07                |
| Beam 2 | 1.2       | 1.041          | 1.25                |
| Beam 3 | 0.98      | 0.87           | 0.85                |

The block that will sink is ……

a. Beam 1  

b. Beam 2  

c. Beam 3  

d. Beam 1 and 2 

e. Beam 2 and 3 

Reason: ………………………………………………………

Why is that? ……………………………

Information/evidence/data that convince you: ……….

Paint the direction of the forces acting on the object as it hovers, floats, and sinks: …………………

Figure 1. Example of Writing Test Items for Data Aspects

The instrument of argumentation skills assessment validates by seven experts consisting of linguists, material experts, educational evaluation experts, and practitioners. The validation by experts and practitioners is the validity of the content. The validity of the content of an instrument is related to a rational analysis of the domain to be measured to determine the representation of the instrument with the ability to be measured [24]. The content-validity coefficient, which is based on the results of the assessment of the expert forum, as many as seven people on items measuring constructs can be measured using the V statistic [25].

The criteria used to determine the content of valid items is to compare the value V count with the V table in the Right-tail Probabilities (p) table for selected values of the validity coefficient (V). Index V values range from 0 to 1, a high value indicating that an item has high content validity or that a set of items has high content validity in the judgment of an expert [26].

The test instrument that has been declared valid by the expert is then tested on empirically on students. Test items were analyzed using modern test theory or what is known as the Item Response Theory (IRT) Rasch model (1PL). The scoring of the test items uses the Partial Credit Model (PCM) technique which is a development of the 1-PL model and belongs to the Rasch model family.

IRT with PCM 1-PL which affects the performance of the testee, so in this model it can be said that all items in the instrument of arguing skills have the same difference in power. This is based on the opinion of Azwar [27], which states that because there is only one item parameter that affects the subject, in the IRT 1-PL model all items in the test have the same difference power. Other experts also stated that the assumptions used in PCM were that each item had the same distinctive power [24].

The selection of data analysis techniques using the IRT 1-PL model is based on the number of samples in the field trial of 100 students. This is supported by experts who state that the specific sample size for the IRT 1-PL model ranges from 30 to 300 with an INFIT t limit of -2 to +2 [28]. The results of
field trial data were analyzed using the Quest and Parscale programs to determine the characteristics of the argumentation skills test.

The feasibility analysis of the items with PCM 1-PL was carried out quantitatively by testing the goodness of fit items: If the INFIT MNSQ approaches 1.0 with a standard deviation approaching 0.0, it can be said that the overall test item is fit with PCM.

3. Results and Discussion

3.1. Content Validity

The results of the development of an argumentation skills assessment instrument before being tested must go through the item validation stage. This stage aims to improve the initial design of the assessment instrument. The assessment is carried out by experts and lecturers in basic physics subjects (practitioners). The criteria used to determine the content of valid items is to compare the value V count with V table, in the Right-tail Probabilities (p) Table for selected values of the validity coefficient (V) with a probability of 0.05 for the rating category 4 (four) and the number of rater (n) = 5 (five), the table shows the price of 0.87. If V count > V table (0.87), then the instrument is content valid. The results of Aiken's content validity analysis show that, overall, the items developed by the argumentation skills assessment instrument have met the content validity requirements.

3.2. Prerequisites Test Item Response Theory Analysis

The assumptions used in modern test theory, IRT are a test device capable of measuring one dimension [5]. The statistical techniques used to prove unidimensional use factor analysis [24], using the SPSS program. Factor analysis used the KMO test and the Bartlett test, which was preceded by testing the adequacy of the sample used in the study. The analysis showed that the Chi-Square value in the Bartlett test was 2202.671 with 190 degrees of freedom and a p-value (Sig.) <0.05. These results indicate that the sample size of 206 used in this study has met the unidimensional requirements. Based on the results of the factor analysis, it shows that the student response data to the instrument of argumentation skills assessment contains 5 eigenvalues which are more than 1, so it can be said that the developed diagnostic test contains 5 factors. The scree plot of the analysis results is presented in Figure 2.

![Scree Plot](image)

Figure 2. Scree plot of Exploratory Factor Analysis Results
Based on Figure 2, it appears that there is 1 dominant factor in the assessment instrument developed, and other factors contribute significantly to the variance.

3.3. Feasibility Analysis of Question Items with PCM 1-PL

The feasibility analysis of the items using the PCM 1-PL was carried out quantitatively using the QUEST program. The feasibility of the items is carried out in two stages of testing, namely: (a) testing the goodness of fit items: if the INFIT MNSQ approaches 1.0 with a standard deviation of close to 0.0, it can be said that the overall test item is fit with PCM; (b) testing the item difficulty index or difficulty index—analysis of the difficulty index using the QUEST program. The item is said to be good if the difficulty index is more than 2.00 or less than 2.00, which can be expressed by (-2.00 < b < 2.00). The instrument that was tested consisted of 20 items. The complete results of the instrument trial are presented in Figure 4.2. An item is said to be fit with the PCM 1-PL analysis model if the INFIT MNSQ value is in the range of 0.70 to 1.30.

![Figure 3. MNSQ INFIT Analysis Output Results](image)

Based on Figure 3 shows that the calibration results, as a whole the items have great support for the total score. This data illustrates that the overall items developed are accepted and compatible with PCM 1-PL. The analysis of the feasibility of the next item is the analysis of the difficulty level or difficulty index.

The item difficulty level provides information on the difference in the student's ability to answer the questions, items that are too easy or too difficult cannot provide sufficient information about the test so that the items need to be discarded or revised. Analysis of the level of difficulty using the QUEST program is based on the results of the difficulty estimation which is the average difficulty index at score 1, score 2, score 3, and score 4 on each item.

The delta-1 value shows the value required for students to move from category-1 (score 1) to category-2 (score 2), the delta-2 value shows the value required for students to move from category-2 (score 2) to category-3 (score 3), and a delta-3 value indicates the score required for students to move from category-3 (score 3) to category-4 (score 4). The results of the analysis of the difficulty level of the test items are presented in Table 2.
Table 2. Results of the Analysis of the Level of Difficulty for Each Question Item

| Item | Difficulty | $\delta_1$ | $\delta_2$ | $\delta_3$ | Item | Difficulty | $\delta_1$ | $\delta_2$ | $\delta_3$ |
|------|------------|------------|------------|------------|------|------------|------------|------------|------------|
| 1    | 0.11       | 0.27       | -0.91      | 0.64       | 11   | -0.89      | -1.67      | -0.24      | 1.91       |
| 2    | 0.50       | 0.30       | -0.86      | 0.56       | 12   | 1.03       | 0.06       | -2.12      | 2.06       |
| 3    | 0.62       | -0.27      | -0.66      | 0.93       | 13   | 1.24       | -0.84      | -1.49      | 2.33       |
| 4    | -0.84      | -2.02      | 0.64       | 1.38       | 14   | 0.49       | -1.52      | -0.14      | 1.66       |
| 5    | -0.04      | -1.02      | -0.05      | 1.06       | 15   | -0.87      | -1.48      | 0.31       | 1.17       |
| 6    | 0.05       | -1.06      | -0.13      | 1.19       | 16   | 0.11       | 0.15       | -0.85      | 0.70       |
| 7    | 0.05       | -1.26      | -0.10      | 1.37       | 17   | 0.40       | -0.70      | -0.48      | 1.18       |
| 8    | 0.62       | 0.10       | 0.22       | -0.21      | 18   | 0.24       | -0.87      | -0.16      | 1.03       |
| 9    | -0.93      | -1.85      | 1.05       | 1.8        | 19   | 0.14       | -1.31      | 0.14       | 1.17       |
| 10   | 0.91       | -2.26      | 0.35       | 1.91       | 20   | 0.12       | -1.05      | 0.01       | 1.03       |

Items 12 and 13 have the highest level of difficulty based on Table 2, with a value for each of the items 1.03 and 1.24. The overall difficulty value for each item is still in the range of values -2 < $\delta$ < +2. Based on this data, it can be said that the difficulty level of the developed questions has met the difficulty index value category, so it can be said that all the items are good. The results of item estimation on the argumentation skills test are presented in Table 3.

Table 3. Testee and Item Estimation Results

| Parameter                      | Estimates for Items | Estimates for Testee |
|--------------------------------|---------------------|----------------------|
| Average and standard deviation | 0.01±0.70           | 0.53±0.58            |
| Reliability                    | 0.73                | 0.86                 |
| The average value and standard deviation INFIT MNSQ | 1.00±0.19 | 1.02±0.43 |
| The average value and standard deviation OUTFIT MNSQ | 1.01±0.21 | 1.01±0.41 |
| The average value and standard deviation INFIT t | 0.04±1.25 | 0.08±1.54 |
| The average value and standard deviation OUTFIT t | 0.03±1.79 | 0.02±1.13 |

The argumentative skills test kit has a mean INFIT MNSQ value of 1.00 and a standard deviation (SD) of 0.19, meaning that the overall items developed are fit with the Rasch model. This is in accordance with the opinion of Aminah [20] which states that with a mean INFIT MNSQ of 1.01 and SD 0.09, it means that overall items are in accordance with the Rasch model. Another expert stated that, with a mean of INFIT MNSQ 1.00 and SD 0.6, it means that all items are in accordance with the Rasch model [29].

Reliability estimation using the QUEST program is calculated based on items called the item spacing index and based on testees is called person separation [20]. The higher the test reliability index, the more precisely all items analyzed based on the PCM 1-PL model. The results of the reliability estimation on the argumentation skills test tool obtained the person reliability (testee) value of 0.86, and the item reliability value of 0.73. It can be concluded that the consistency of the answers either from the testee or from the items in the argumentation skills test kit has high reliability.

Based on the results of testing the feasibility of the items with the PCM 1-PL with two steps, namely the INFIT MNSQ and the difficulty index, the items developed in the argumentation skills test kit were accepted as a whole 20 items. The results of calculations using QUEST and additional analysis using Excel obtained a display of the difficulty level of the questions based on the aspects and sub-aspects of the instrument. The value of the level of difficulty per aspect and sub-aspect of argumentation skills is presented in Table 4.
Based on Table 4, it is known that the level of difficulty of the hardest questions is the qualifier aspect with the sub-qualifier aspect implying a specification statement. The level of difficulty in these aspects and sub-aspects is 0.86 in the medium category. Then the backing aspect with the backing sub-aspect strengthens the relationship of the statement, has a difficulty level of 0.71 in the medium category. Overall, the aspect of argumentation skills is in the medium category, namely in the difficulty value range -1.00 < b < 1.00. The lowest problem difficulty level is in the data and warrant aspects, while the highest problem difficulty level is in the qualifier aspect with the qualifier sub-aspect implying a specification statement. This indicates that the highest student's argumentation skills in the qualifier aspect implies the statement specification and the lowest student's argumentation skills on the data and warrant aspects. This is based on the opinion of Mardapi [5], which states that if the testee's ability is low, the item difficulty level will be high.

3.4. Standard Error of Measurement (SEM)

The value of the item parameter index and the ability of the testee is the estimation results, so the truth is probability and cannot be separated from measurement errors. In item response theory, measurement error or Standard Error of Measurement (SEM) is closely related to the item information function. The relationship between the information function and SEM, as shown in Figure 4.
The information function and SEM have an inversely proportional relationship, which is based on the results of the research in Figure 4. Hambleton, Swaminatan, & Rogers [28], which states that the information function with SEM has an inversely proportional relationship, meaning that the greater the information function, then SEM is getting smaller, and vice versa.

The value of the information function of the ability \(-\sim\) increases to reach the maximum value, then decreases until \(+\sim\). Meanwhile, the measurement error is, on the contrary, decreases to reach the minimum value, then rises again. The two graphs converge on the -2.8 and +1.5 ability scales. Between these two abilities, items have a higher information function value than their measurement error. Conversely, when the ability scale is less than -2.8 and more than +1.5 items have a higher measurement error with the information it provides. Thus, the argued skill test kit is suitable for testees in that ability range, which is suitable for students with low to moderate abilities.

4. Conclusions
Argument skills assessment instruments include aspects of Claim, Data, Warrant, Backing, and Qualifier. The test scoring procedure used in this study uses the analysis of polytomous items using the Partial Credit Model (PCM), which takes into account several stages in the processing of questions by students. The PCM pattern is applied to multiple-choice questions because it is reasonable as an alternative model in an effective and fair physics learning assessment. Result expert judgment shows that the argumentative skills assessment instrument used has met the content validity. The form of reasoned multiple-choice questions developed on Static Fluid has met the scoring criteria for polytomous items, based on a) the MNSQ INFIT value is in the range 0.70 to 1.30, and b) testing the item difficulty index or difficulty index is in the range \(-2.00 < b < 2.00\). The reliability estimate on the test equipment obtained a reliability value of 0.86, which means that the reliability of the test is in the high category. The graph of the information function and SEM converge on the ability scales of -2.8 and +1.5, meaning that the skills test kit argues it is suitable for testees in the low to moderate ability range.

References
[1] M. N. Shahib, *Pembinaan Kreativitas Menuju Era Global*. Bandung: P.T. Alumni, 2003.
[2] H. Affandi, N. S. Aminah, and A. Supriyanto, “Analisis Keterampilan Berpikir Kritis Siswa Pada Materi Fluida Dinamis Di SMA Batik 2 Surakarta,” *J. Mater. dan Pembelajaran Fis.*, vol. 9, no. 1, pp. 25–33, 2019.
[3] Afandi. and Sajidan, *Stimulasi Keterampilan Berpikir Tingkat Tinggi*, 1st ed. Surakarta: UNS Press, 2017.
[4] S. R. N. A. S. Easterbrooks, “Applying Critical Thinking Skills to Character Education and Values Clarification with Students Who are Deaf of Hard of Hearing,” *Am. Ann. Deaf*, vol. 149, no. 3, 2004.
[5] D. Mardapi, *Pengukuran, Penilaian, dan Evaluasi Pendidikan*, Yogyakarta: Nuha Litera, 2012.
[6] N. Sudjana, *Penilaian Hasil Proses Belajar Mengajar*, 7th ed. Bandung: PT Remaja Rosdakarya, 2003.
[7] N. Sudjana, *Metode Statistika*, 6th ed. Bandung: PT. Tarsito, 2005.
[8] C. Cari, S. N. Pratiwi, H. Affandy, and D. A. Nugraha, “Investigation of undergraduate student concept understanding on Hydrostatic Pressure using two-tier test,” *J. Phys. Conf. Ser.*, vol. 1511, no. 1, 2020, doi: 10.1088/1742-6596/1511/1/012085.
[9] N. Sudjana, *Penilaian Hasil Proses Belajar Mengajar*, Bandung: Remaja Rosdakarya, 2017.
[10] N. Sudjana, *Penilaian Hasil Proses Belajar Mengajar*, Bandung: Remaja Rosdakarya, 2017.
[11] E. Lopez, J. Kim, K. Nandagopal, N. Cardin, R. J. Shavelson, and J. H. Penn, “Validating the use of concept-mapping as a diagnostic assessment tool in organic chemistry: Implications for teaching,” *Chem. Educ. Res. Pract.*, vol. 12, no. 2, pp. 133–141, 2011, doi: 10.1039/c1rp90018h.
[12] S. N. Pratiwi, C. Cari, N. S. Aminah, and H. Affandy, “Problem-Based Learning with Argumentation Skills to Improve Students’ Concept Understanding,” J. Phys. Conf. Ser., vol. 1155, no. 1, 2019, doi: 10.1088/1742-6596/1155/1/012065.

[13] R. A. Serway and J. W. Jewett, Physics 6th. 2004.

[14] Z. K. Viyanti., Cari., Sunarno W. Prasetyo, “The Development Rubric Skill Argued as Alternative Assessment Floating and Sinking Materials,” J. Res. Sci. Teach., vol. 41, no. 10, pp. 994–1020, 2004.

[15] S. Toulmin, The Uses of Argument. Cambridge: Cambridge University Press, 1958.

[16] E. S. Osborne J and S. Simon, “Enhancing the quality of argumentation in School science,” J. Res. Sci. Teach., vol. 41, no. 10, pp. 994–1020, 2004.

[17] W. Z. Wahdan, O. Sulistina, and D. Sukarianingsih, “Analisis Kemampuan Berargumentasi Ilmiah Materi Ikatan Kimia Peserta Didik SMA, MAN, dan Perguruan Tinggi Tingkat I,” J. Pembelajaran Kim., vol. 2, no. 2, 2017.

[18] R. D. P. N. J. Osborne, “Establishing The Norms of Scientific Argumentation in Classrooms,” Sci. Educ., vol. 84, no. 3, pp. 287–312, 2000.

[19] Sukardi, Evaluasi Pendidikan: Prinsip dan Operasionalnya. Jakarta: Bumi Aksara, 2015.

[20] N. S. Aminah, Assesmen Pembelajaran Fisika, 1st ed. Surakarta: UNS Press, 2017.

[21] E. Istiyono, D. Mardapi, and Suparno, “Pengembangan Tes Kemampuan Berpikir Tingkat Tinggi (PysTHOTS) Peserta Didik SMA,” J. Penelit. Dan Eval. Pendidik., vol. 18, no. 1, pp. 1–12, 2014.

[22] E. Istiyono, “The application of GPCM on MMC test as a fair alternative assessment model in physics learning,” Proceeding 3rd Int. Conf. Res. Implement. Educ. Math. Sci. (ICRIEMS), 16-17 May 2017, no. May, pp. 25–30, 2016, [Online]. Available: http://seminar.uny.ac.id/icriems/sites/seminar.uny.ac.id.icriems/files/prosiding/PE-04.pdf.

[23] B. Sumintono and W. Widhiarso, Aplikasi Pemodelan RASCH Pada Assessment Pendidikan, Revisi. Cimahi: Penerbit Trim Komunikata, 2015.

[24] H. Retnawati, Validitas Reliabilitas dan Karakteristik Butir: Panduan untuk Peneliti, Mahasiswa dan Psikometrian. Yogyakarta: Parama Publishing, 2016.

[25] L. R. Aiken, “Content validity and reliability of single items or questionnaires,” Educ. Psychol. Meas., vol. 40, no. 4, pp. 955–959, 1980, doi: 10.1177/00131644800400419.

[26] L. R. Aiken, “Three coefficients for analysing Reliability and Validity of rating.,” Educ. Psychol. Meas., vol. 45, pp. 131–142, 1985, doi: 10.1177/07399863870092005.

[27] S. Azwar, Tes Prestasi; Fungsi dan Pengembangan Pengukuran Prestasi Belajar, Edisi II. Yogyakarta: Pustaka Pelajar, 2014.

[28] H. Hambleton, R.K. Swaminathan and H. J. Rogers, Fundamental of Item Response Theory. Newbury Park: Sage Publication Inc., 1991.

[29] B. Subali and P. Suyata, Panduan Analisis Data Pengukuran Pendidikan Untuk Memperoleh Bukti Empirik Keshahihan Menggunakan Program QUEST. Yogyakarta: Lembaga Penelitian dan Pengabdian pada Masyarakat Universitas Negeri Yogyakarta, 2011.