Phys. Comm. 5 (2) 2021 : 33 - 43

Physics Communication

http://journal.unnes.ac.id/nju/index.php/pc

Development of a Testlet Model Physics Learning Assessment to Know Students' Cognitive Profiles on Static Fluids

Rachma Afifah, Putut Marwoto, Ellianawati

Postgraduate Universitas Negeri Semarang, Semarang, Indonesia

Article Info

Article history:
Submitted 31 May 2021
Revised 15 August 2021
Accepted 18 August 2021

Keywords:
assessment, testlet, cognitive profile, concept understanding.

Abstract

The ability of students to achieve learning objectives can be measured by conducting an assessment. There are various forms of assessment, and the assessment in the form of a testlet is believed to have advantages in identifying the students' conceptual understanding achievements. This study aims to develop a testlet model physics instrument that is valid, reliable, and feasible to use. The development procedure used is 4D which consists of 4 stages, namely: Define, Design, Develop, and Disseminate. However, this research is reduced to 3D only limited to the Develop stage. The research was conducted in a high school located in the city of Semarang. The small-scale trial sample involved 20 students who were selected using a random sampling technique. The results showed that the content validity analysis by the expert obtained an average result of 0.9 in the valid category. The results of the small-scale product test analysis showed that 92% of the items were declared valid and 8% of the items were declared invalid. The question is declared to have very high reliability with a value of 0.974. Analysis of the level of difficulty obtained criteria for 8% difficult questions and 92% moderate questions. The discriminating power of the questions shows the percentage of 2% of the questions being rejected, 2% of the questions being corrected, 2% of the questions being accepted well with improvements, and 94% of the questions being well-accepted.

✉Correspondence:
Postgraduate Universitas Negeri Semarang
Jalan Kelud Utara III No.37, Kota Semarang, Indonesia 50237
e-mail: rachmaafifah@students.unnes.ac.id

p-ISSN 2528-5971
e-ISSN 2528-598X

INTRODUCTION

Education in the 21st century must face the challenges of an increasingly advanced era and industrial needs with high abstraction tasks so that understanding and problem-solving abilities are needed in a person. The ability of a person can be known by conducting an assessment according to the learning objectives to be achieved. Assessment or assessment has a function in the form of (1) knowing the extent to which students have succeeded in following the lessons given by the teacher and (2) helping teachers to find out the difficulties and successes of students in mastering the material (Arikunto, 2013). Programmed assessment in learning is an assessment approach that is focused on measuring competence to assess students thoroughly and meaningfully (Schuwirth & Vleuten, 2019).

According to Lee et al. (2019), assessment in learning places students at the center of learning and is considered a powerful alternative assessment approach that can maximize student learning. The assessment also has considerable potential in improving the quality of learning (Wiliam, 2011). According to Care et al. (2019), the diversity of evaluation tools is not only seen from the dimensions of authenticity and assessment design but how teachers can design learning that can facilitate the development of 21st-century skills.

An instrument is a tool that can be used to measure the level of attainment of student competence (Trianto, 2013), the process of collecting data on the description of student learning development (Nurhadi, 2009), and to investigate the understanding of knowledge concepts and make connections between knowledge concepts (Asmalia et al., 2015). Mardapi (2017) explains that there are nine steps in developing and compiling test instruments. The nine steps include: (1) compiling test specifications; (2) writing tests; (3) reviewing the test; (4) conducting test trials; (5) analyzing the test items; (6) improving the test; (7) assemble test; (8) carry out the test; and (9) interpreting the test. The development of test instrument forms is increasingly diverse and is adapted to the needs and developments of the times.

Testlet is a test that was developed and designed to represent essay questions because in a set of questions related to one another (Slepkov & Shiell, 2014) and each question represents the steps in solving the problem (Shiell & Slepkov, 2015) by sharing a common stimulus. such as graphics or text (Frey et al., 2016). According to Mindyarto et al. (2017), the testlet variant has pedagogical features to stimulate qualitative analysis in solving physics problems.

Instruments in the form of objective tests have been developed. Testlet is a form of objective test that was developed. The testlet is a series of question items that have conceptual attachments and are hierarchical with a tiered multiple-choice model (Kusumaningrum et al., 2015). The research that has been done by Wahyuni et al. (2015) revealed that the testlet model test instrument developed had good content validity, high reliability, and could be used to determine the abilities and learning difficulties experienced by students. According to Mindyarto et al. (2017), the testlet variant has pedagogical features to stimulate qualitative analysis in solving physics problems. This means the testlet test has the opportunity to be used to evaluate students' thinking processes in solving physics problems.

The understanding of students' concepts of static fluid material is also still low. Subtopics with low conceptual understanding are hydrostatic pressure, Pascal's law, and Archimedes force (Adisna et al., 2019; Putri et al., 2017; Yadaeni et al., 2016). Students assume that hydrostatic pressure is influenced by the shape of the vessel (Putri et al., 2017), volume (Yadaeni et al., 2016), and cross-sectional area (Adisna et al., 2019). Efforts made to reduce misconceptions include using interactive learning media when delivering static fluid material (Zukhruf et al., 2016), the Predict Discuss Explain Observe Discuss Explain (PDEODE) learning model assisted by PhET animation to remediate misconceptions in static fluid material (Anggraeni, 2018), and Treffinger's learning model with a STEM (Science, Technology, Engineering, and Mathematics) approach to remediate the static fluid misconception (Zaqiyatunnisak, 2019).

The ability to understand students' concepts is very important to support life in the future, so an in-depth study is needed to determine the cognitive profile of students' conceptual understanding with a test instrument with a STEM-based testlet model.

METHODS

This research was carried out at SMA N 11 Semarang. The small-scale trial sample involved 20 students who were selected using a random sampling technique. The design of this research is Research and Development which refers to the research design (Sugiyono, 2017). According to Thiagarajan (1974), Define, Design, Develop, and Disseminate. However, this research is reduced to 3D only limited to the Develop stage.

The Define stage includes literature review, pre-research, and needs analysis. The literature review stage was carried out by reviewing some of the literature on test instruments using the testlet method and several studies on cognitive profiles. The Design phase begins with determining the construct, determining the objectives, determining the format of the questions, and determining the assessment guidelines. The Develop stage is carried out after the question assessment instrument is completed and then tested by experts and then revised according to the suggestions for improvement given by the experts. The revised instrument was then piloted to students on a small scale. Small-scale trials were conducted to analyze the validity of the items, reliability, level of difficulty, and discriminatory power. The results of small-scale trials that are used to measure the feasibility of the developed instrument, if improvements are needed, must be carried out so that the instrument is as expected. The small-scale trial stage was also given a questionnaire to students to find out whether the instrument was by the objectives.

The instruments used were in the form of a testlet and questionnaire. Questionnaire data were obtained by validating 5 expert validators consisting of 3 lecturers and 2 teachers. The results of the validation and analysis with V Aiken are used to improve the developed product. Calculation of expert validity test using Aiken's V index formula according to (Aiken, 1985).

$$V = \frac{\Sigma s}{n(c-1)} = \frac{r - l_o}{n(c-1)} \qquad (1)$$

where

V : consistency of content validity

s : the score given by the rater minus the lowest score

r : the score given by the rater

l_o : lowest validity assessment score

c : highest validity assessment score

n : number of raters

RESULTS AND DISCUSSION

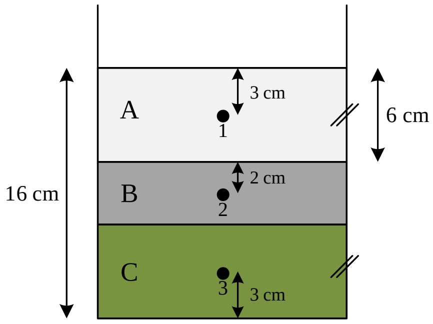
The products produced in the research include testlet model assessment instruments on static fluid material consisting of: 1) question grids, 2) question texts, and 3) answer keys and discussion. The product design of the assessment instrument developed was in the form of a multiple-choice testlet consisting of 15 main questions consisting of 50 questions. The product developed has gone through an assessment of core competencies and basic physics competencies in static fluid materials. The basic competencies are then broken down into 15-question indicators. An example of writing multiple-choice testlet questions can be seen in Table 1.

Table 1. Example of Writing Multiple Choice Questions for the Testlet Model

Test *Testlet*

Main Problems (Problems 2.1 – 2.5 regarding Hydrostatic Pressure)

A vessel contains 3 types of liquids with different densities $\rho_A = 0{,}8\ \text{g/cm}^3$, $\rho_B = 1\ \text{g/cm}^3$, $g = 9{,}8\ \text{m/s}^2$ in the vessel there are 3 points of different depths.



Testlet Integration 2

2.1 The distance of point 3 from the surface of the liquid is ….

a. 3 cm

b. 7 cm

c. 8 cm

d. 10 cm

e. 13 cm

2.2 The pressure that occurs at point 2 is ….

a. 6,272 Pa

b. 6,664 Pa

c. 627,2 Pa

d. 666,4 Pa

e. 784 Pa

2.3 The difference in pressure $P_2 – P_1$ is ….

a. 392 Pa

b. 431,2 Pa

c. 490 Pa

d. 548,8 Pa

e. 784 Pa

2.4 If the pressure $P_3 – P_2$ = 4194,4 Pa, then ρ_C is ….

a. 13,6 g/cm^3

b. 800 kg/m^3

c. 1250 kg/m^3

d. 1360 kg/m^3

e. 13600 gr/cm^3

2.5 The pressure that occurs at point 3 is ….

a. 1019,2 Pa

b. 1274 Pa

c. 1732,6 Pa

d. 3998,4 Pa

e. 4860,8 Pa

Content Validity

Expert validation was carried out by five validators consisting of 3 lecturers and 2 teachers with quantitative and qualitative data. Quantitative data is obtained by means of the validator filling out the validation sheet. Qualitative data were obtained from suggestions or comments given by 5 validators in order to improve the product before conducting a small-scale test. The suggestions from the validator in detail are presented in Table 2.

Table 2. Suggestions for Testing the Validity of Testlet Instruments

Validator	Suggestion
Validator 1	The problem is good, it just needs improvement in the illustration image
Validator 2	1. Writing the number of questions and multiple choices to tidy up 2. The image of mercury is clarified (figure 1) 3. Images are clarified with clearer shading/colors (figures 2 and 3) 4. The beam image is clarified/shaded (figure 4) 5. Figure 5 is clarified to distinguish water from blocks, the FA image in question number 5 does not come out above the water and the gravity vector w should not disappear after immersing in water 6. The image of the water is clarified (figures 6, 7, 8, 9, 10, 13) 7. The image of the iron ball (picture 11) is clarified/don't leave it blank
Validator 3	1. Complete Bloom's taxonomy level of each question. 2. Avoid word problems because they can have a double meaning.
Validator 4	The questions are good enough, it just needs improvement in the illustrations provided.
Validator 5	1. Some of the question indicators do not include the competencies that must be achieved by students. 2. Some questions are not appropriate in the placement of cognitive profile indicators for understanding concepts.

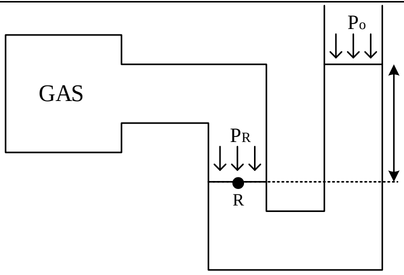
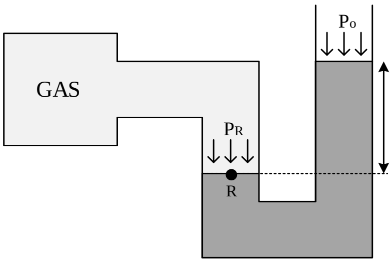
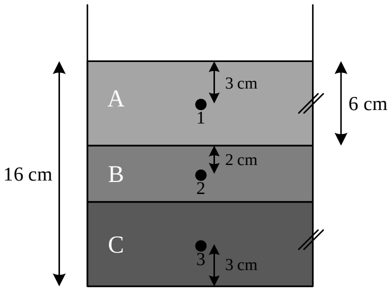
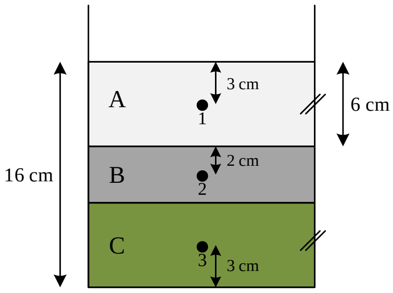
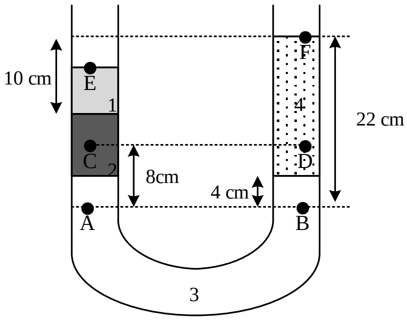
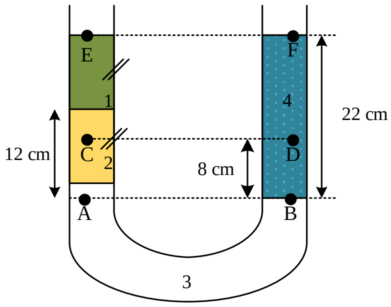
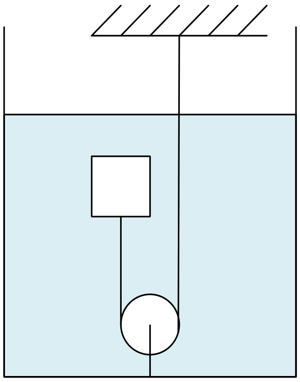
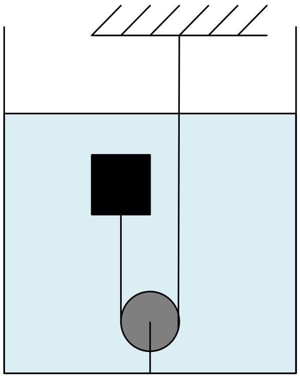
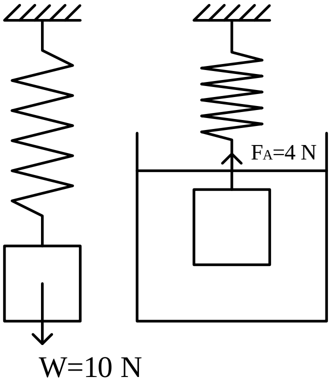
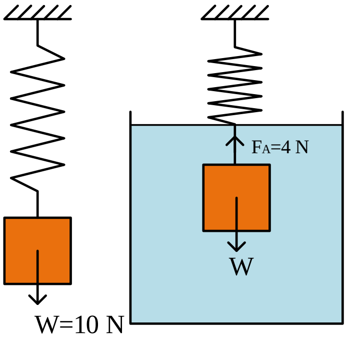
Quantitative data on the validation of the testlet model of the physics learning assessment instrument that was analyzed was obtained from the results of the instrument assessment by experts in the field of physics. The score obtained will be analyzed using V Aiken validity analysis. The results of the analysis can be said to be valid if it meets the V Aiken coefficient limit. The V Aiken limit requirement for 4 categories and 5 raters is 0.87 with a probability of 0.021. The results of the analysis of the average expert validation obtained 0.9, then it can be declared valid so that it is feasible to use. This is also in line with Retnawati (2016) statement, Aiken's V score whose value is more than 0.8 is said to be very valid. Aiken's V mean scores are presented in Table 3.

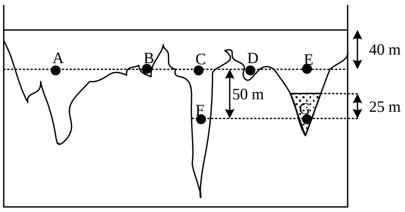
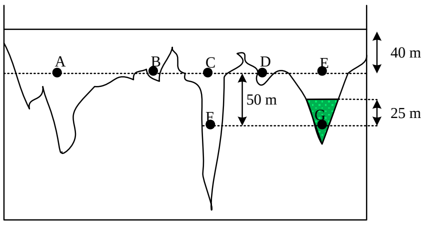
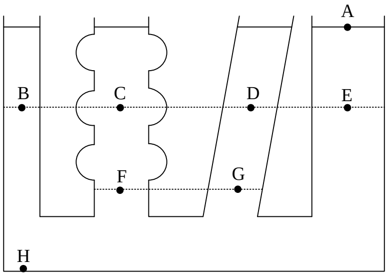
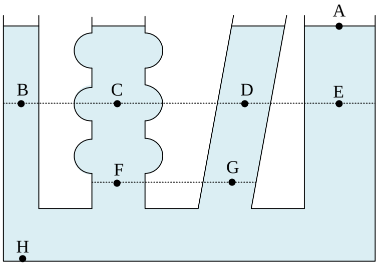
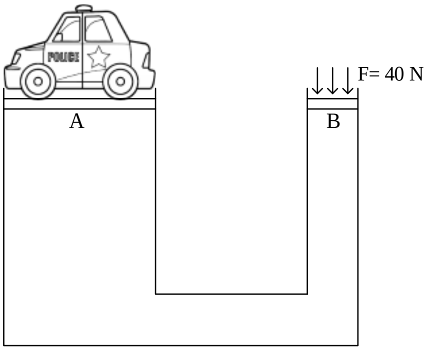
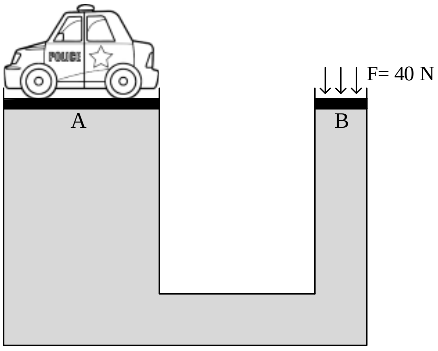
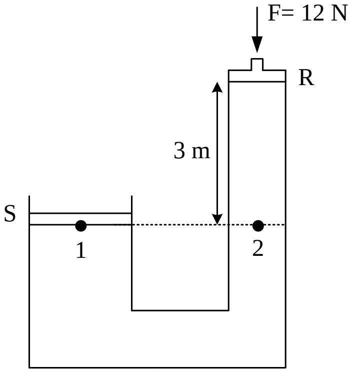
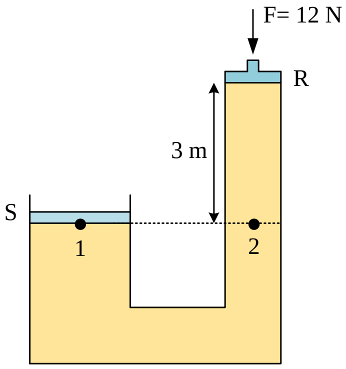
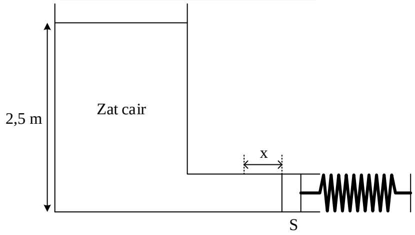
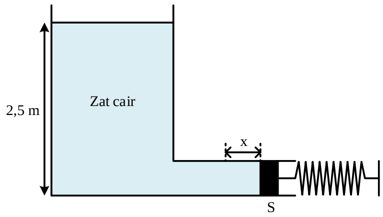
Table 3. Aiken's V score for expert validation of the assessment instrument

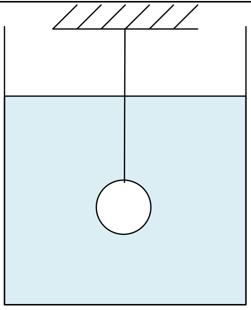
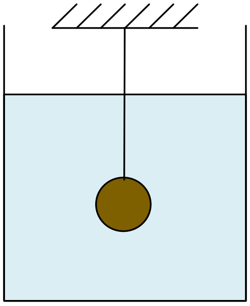
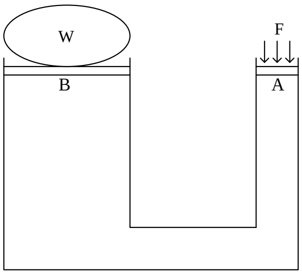
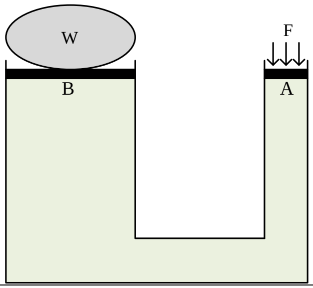
	Content Aspect	Construct Aspect	Language Aspect	Average
V Score	0.87	0.92	0.89	0.90

Suggestions from validators were considered and implemented to improve the STEM-based testlet instrument. Some questions that need to be revised and illustrated images are also corrected according to the suggestions given by the validator. The changes to several items that were revised are presented in Table 4.

Table 4. Revision Results in the Testlet Instrument

No	Before Revision	After Revision
1		
2	 A 1 3 cm 6 cm 16 cm B 2 2 cm C 3 3 cm	 A 1 3 cm 6 cm 16 cm B 2 2 cm C 3 3 cm
3	 10 cm E 1 C 2 8cm 4 cm A B 3 F 4 D 22 cm	 E 1 12 cm C 2 8 cm A B 3 F 4 D 22 cm
4		
5		

No	Before Revision	After Revision
6		
7		
8		
9		
10		

No	Before Revision	After Revision
11		
13		

Test Results

A small-scale test of the testlet model assessment instrument was conducted to obtain the data used to analyze the validity, reliability, level of difficulty, discriminating power of the items, and cognitive profile analysis of concept understanding. A large-scale test was carried out to obtain data that was used to analyze the reliability and cognitive profile analysis of concept understanding. The data from the analysis of the validity, reliability, level of difficulty, and discriminating power of the items that were are presented in Table 5.

Table 6. The results of the analysis of validity, reliability, level of difficulty, and discriminating power of items

Item Validity		
Category	Total	Percentage (%)
Valid	46	92
Invalid	4	8
Reliability		
	Small Scale Test	Large Scale Test
r_{count}	0.974	0.939
Level of Difficulty		
Category	Total	Percentage (%)
Easy	0	0
Medium	46	92
Difficult	4	8
Discriminating Power		
Category	Total	Percentage (%)
Not used	1	2
Repaired	1	2
Accepted but repaired	1	2
Well accepted	47	94

The results of the analysis of the validity of the items were influenced by several factors including items that were not understood by the respondents and many students answered correctly. For example, in question number 3.1, only 8 students answered correctly.

The results of the reliability analysis on the small-scale test and large-scale test obtained a reliability value of more than 0.70. Instruments that have a reliability value of more than 0.70 are declared reliable for use (Kurniawan & Lestari, 2019; Setiawan & Faoziyah, 2020; Shobrina et al., 2020). The results of the reliability analysis of the small-scale test data are greater than the large-scale test, because in the small-scale test the number of questions is 50 questions, while in the large-scale test the number of questions is 44 questions. A test with a large number of questions is certainly more reliable than a test that only consists of a few items (Arifin, 2014; Retnawati, 2016) because many samples are being measured and the proportion of correct answers is increasing (Arifin, 2014). The level of difficulty with difficult and easy categories for a group will result in low reliability (Arifin, 2014)

The results of the analysis of the level of difficulty obtained the percentage of questions in the difficult category of 8%, medium 92%, and easy 0% where the proportion of the level of difficulty of the questions was not balanced. The proportion of the difficulty level of a good question to obtain maximum learning achievement (Arifin, 2014) is that the proportion of difficulty level must be spread in a balanced way, such as 25% difficult questions, 50% moderate questions, and 25% easy questions.

Cognitive Profile of Concept Understanding

The results of the cognitive profile analysis of concept understanding can be seen in Table 7.

Table 7. Results of Percentage Analysis of Cognitive Profile Indicators

Indicators	Small Scale Test (%)	Large Scale Test (%)
Interpreting	44	37
Classifying	48	42
Inferring	43	37
Comparing	51	37
Explaining	48	37

The results of the analysis of the percentage of cognitive profile indicators on the small-scale test all indicators fall into the fairly high category, while on the large-scale test the indicators classify in the fairly high category and the other 4 categories in the low category. Based on the results of the analysis of conceptual understanding from the large-scale and small-scale test data, it can be concluded that there is a decrease in the percentage of each indicator of the cognitive profile of concept understanding. The decrease in the percentage of each indicator is because some questions were tested on the small-scale test that was not tested on the large-scale test. The decrease in percentage also occurred due to several factors including the number of respondents in the large-scale test more than in the small-scale test so that the abilities of the respondents vary, the psychological condition of the respondent also affects the results of the test, and the environmental conditions when the test takes place.

CONCLUSION

The results of the content validation of the five expert validators found that the items developed had an average V Aiken of 0.9 which were included in the very valid category. The comments and suggestions obtained include illustrations of the images in the questions that need to be clarified. The results of the analysis of the validity of the items in the small-scale product test showed that 46 items were declared valid and 4 items were declared invalid. The question is declared to have very high reliability with a value of 0.974. Analysis of the level of difficulty obtained criteria for difficult questions 4 and medium questions 46. The discriminating power of questions showed the percentage of 1 question being rejected, 1 question being corrected, 1 question being accepted well with improvement, and 47 questions being accepted well.

REFERENCES

Adisna, Q. D. P. P., Wahyuni, A., & Suyudi, A. (2019). Analisis Pemahaman Konsep Fisika Siswa pada Pokok Bahasan Fluida statis. *Jurnal Ilmu Fisika Dan Pembelajarannya*, *3*(2), 68–75.

Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, *45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Anggraeni, Y. M. (2018). *Remediasi Miskonsepsi dengan Model Pembelajaran Predict-Discuss-Explain-Observe-Discuss-Explain (PDEODE) Berbantuan PhET Simulation pada Materi Fluida*.

Arifin, Z. (2014). *Evaluasi Pembelajaran*. Remaja Rosdakarya.

Arikunto, S. (2013). *Dasar-Dasar Evaluasi Pendidikan Edisi 2*. Bumi Aksara.

Asmalia, I., Fadiawati, N., & Kadaritna, N. (2015). Pengembangan Instrumen Asesmen Berbasis Keterampilan Proses Sains pada Materi Stoikiometri. *Jurnal Pendidikan Dan Pembelajaran Kimia*, *4*(1), 299–311.

Care, E., Kim, H., Vista, A., & Anderson, K. (2019). Education System Alignment for 21st Century Skills: Focus on Assessment. In *Center for Universal Education at the Brookings Institution.* (Issue January). https://www.brookings.edu/wp-content/uploads/2018/11/Education-system-alignment-for-21st-century-skills-012819.pdf

Frey, A., Seitz, N. N., & Brandt, S. (2016). Testlet-Based Multidimensional Adaptive Testing. *Frontiers in Psychology*, *7*(NOV), 1–14. https://doi.org/10.3389/fpsyg.2016.01758

Kurniawan, R. Y., & Lestari, D. (2019). The Development Assessment Instruments of Higher Order Thinking Skills on Economic Subject. *Dinamika Pendidikan*, *14*(1), 102–115. https://doi.org/10.15294/dp.v14i1.19226

Kusumaningrum, L., Yamtinah, S., & Saputro, A. N. C. (2015). Pengembangan Instrumen Tes Diagnostik Kesulitan Belajar Kimia SMA Kelas XI Semester I Menggunakan Model Teslet. *Jurnal Pendidikan Kimia*, *4*(4), 36–45.

Lee, I., Mak, P., & Yuan, R. E. (2019). Assessment as Learning in Primary Writing Classrooms: An Exploratory Study. *Studies in Educational Evaluation*, *62*, 72–81. https://doi.org/10.1016/j.stueduc.2019.04.012

Mardapi, D. (2017). *Pengukuran, Penilaian, dan Evaluasi Pendidikan*. Parama Publishing.

Mindyarto, B. N., Mardapi, D., & Bastari. (2017). Development of a Testlet Generator in Re-Engineering the Indonesian Physics National-Exams. *AIP Conference Proceedings*, *1868*(August 2017), 1–9. https://doi.org/10.1063/1.4995178

Nurhadi. (2009). *Pendekatan dalam Penilaian*. Pustaka Sinar Harapan.

Putri, U. D., Parno, & Supriana, E. (2017). Identifikasi Pemahaman Konsep Siswa SMA pada Materi Fluida Statis. *Prosiding Seminar Pendidikan IPA Pascasarjana UM*, *2*, 316–324. https://doi.org/10.21067/mpej.v1i1.2216

Retnawati, H. (2016). *Analisis Kuantitatif Instrumen Penelitian*. Parama Publishing.

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2019). How 'Testing' Has Become 'Programmatic Assessment for Learning.' *Health Professions Education*, *5*(3), 177–184. https://doi.org/10.1016/j.hpe.2018.06.005

Setiawan, D., & Faoziyah, N. (2020). Development of a Five-Tier Diagnostic Test to Reveal the Student Concept in Fluids. *Physics Communication*, *4*(1), 6–13.

Shiell, R. C., & Slepkov, A. D. (2015). Integrated Testlets: A New Form of Expert-Student Collaborative Testing. *Collected Essays on Learning and Teaching*, *VIII*, 201–210.

Shobrina, N. Q., Sakti, I., & Purwanto, A. (2020). Pengembangan Desain Bahan Ajar Fisika Berbasis E-Modul Pada Materi Momentum. *Jurnal Kumparan Fisika*, *3*(1), 33–40. https://doi.org/10.33369/jkf.3.1.33-40

Slepkov, A. D., & Shiell, R. C. (2014). Comparison of Integrated Testlet and Constructed-Response Question Formats. *Physical Review Special Topics - Physics Education Research*, *10*(2), 1–15. https://doi.org/10.1103/PhysRevSTPER.10.020120

Sugiyono. (2017). *Metode Penelitian Pendidikan Kuantitatif, Kualitatif, dan R&D*. Alfabeta.

Thiagarajan, S., & Etc. (1974). *Instructional Development for Training Teachers of Exceptional Children: A Sourcebook*. National Center for Improvement of Educational. https://doi.org/10.1016/0022-4405(76)90066-2

Trianto. (2013). *Model Pembelajaran Terpadu*. Bumi Aksara.

Wahyuni, I., Yamtinah, S., & Utami, B. (2015). Pengembangan Instrumen Pendeteksi Kesulitan Belajar Kimia Kelas X Menggunakan Model Testlet. *Jurnal Pendidikan Kimia*, *4*(4), 222–231.

Wiliam, D. (2011). What is Assessment for Learning? *Studies in Educational Evaluation*, *37*(1), 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001

Yadaeni, A., Kusairi, S., & Parno. (2016). Studi Kesulitan Siswa dalam Menguasai Konsep Fluida Statis. *Prosiding Semnas Pendidikan IPA Pascasarjana UM*, 59–65.

Zaqiyatunnisak. (2019). Remediasi Miskonsepsi Melalui Model Treffinger dengan Pendekatan STEM (Science, Technology, Engineering, and Mathematics) pada Materi Fisika SMA. In *Universitas Islam Negeri Raden Intan Lampung*.

Zukhruf, K. D., Khaldun, I., & Ilyas, S. (2016). Remediasi Miskonsepsi Dengan Menggunakan Media Pembelajaran Interaktif Pada Materi Fluida Statis. *Jurnal Pendidikan Sains Indonesia*, *04*(02), 56–68.