Developing an Essay Test Instrument for Measuring Diagram Representation and the Capability of Argumentation on Newton’s Law

A N Syarif and H Kuswanto
Physics Education, Postgraduate Program, Universitas Negeri Yogyakarta
Colombo 1 Street, Karangmalang, Yogyakarta, Indonesia
andi.nurfitri2016@student.uny.ac.id, herukus61@uny.ac.id

Abstract. The objective of this research was to develop an essay test instrument for measuring diagram representation and capability of argumentation of senior high school students. This research was conducted through three stages: (1) planning; (2) instrument experiment; and (3) manufacture. The instrument that would be administered was an essay test that consisted of 14 items for measuring diagram representation and capability of argumentation on the Newton’s law. This study implemented qualitative and quantitative validation analyses that have been provided by eight experts (lecturers, teachers, and peer reviewer) and an empirical validation using the Partial Credit Model (PCM) Item Response Theory (IRT) instrument with QUEST and PARSCALE application. The results of the quantitative validation were attained from Aiken’s V Index and showed this index of six items on the measurement of diagram representation is around 0.94-1.00. On the other hand, the results of the quantitative validation through Aiken’s V Index also showed that index of eight items on the measurement the capability of argumentation had is around 0.88-1.00. Then, the empirical validity test involved 250 Grade XI students from SMAN 8 Bulukumba, South Celebes. Instrument analysis was conducted using PCM IRT on the basis of the QUEST application, and the results of this analysis showed that the INFIT Means-Square (INFIT MNSQ) was around 0.77 to 1.33 and the level of difficulty was around -0.824 to 0.797. The item reliability of each test item for the diagram representation and capability of argumentation was 0.79 and 0.85 respectively. In conclusion, the test instrument was in accordance with the students’ capability, ranging from -2.542 to 1.922, with a Standart Error Measurement SEM of 0.222, based on the PARSCALE outcome.

1. Introduction
Physics is a branch of science that exists in each aspect of life. The meaningfulness of physics in learning processes is a very important matter. Physics is not only a matter of mastering knowledge in the form of facts, concepts and principles, but also a matter of mastering inventions. The process of learning physics emphasises the provision of direct experience in order to develop physical competencies so that students may explore and understand the surrounding environment scientifically. Students are directed to perform investigations and action so that they can attain a better in-depth understanding of their surrounding environment. In daily life, people encounter problems and phenomena that might be explained and solved by implementing physics.

The objective of the learning process of physics is to generate competent problem-solvers [1]. This learning process is similar to the development of problem-solving capability. Success in solving problems is measured by the number of problems that students have solved appropriately [2]. There are several cognitive capabilities that play a significant role in solving physics problems: (1) the capability of identifying and interpreting physics concepts and principles appropriately and (2) the capability of
describing and organising physics knowledge effectively [3]. These capabilities might be identified through the use of a physics representation form. Such representation form might consist of a textual form, a mathematical form, a graphical form, verbal form and the like [4]. The success of the students in solving a problem relies on their capability in using the form.

Multi-representation, such as diagrams or graphs and equations, has advantages that can be used in learning science [5]. Diagram representation is a representation that presents a physical concept or a process in the form of a real picture that is similar to the original one [6]. [7] in their research pointed out that the representation of a diagram is very important in teaching Newton’s law to first-year students who have a clear impact on student’s knowledge. On the other hand, diagram representation can provide an overview of a physical concept to be more contextual. Therefore, diagram representation has become one of the students’ solutions to understand and solve the physics problems that they face.

For active, creative, and enjoyable learning, one of the conditions that students must be able to think correctly and communicate well. The capability of argumentation trains the students in using their thinking skills. [8] argued that the capability of argumentation plays an important role in developing critical thinking patterns and adding a deep understanding of an idea or idea. [9] suggests that to present arguments, a diagram is needed. Tests are one of the instruments that have been designed for the objective of measurement. At the same time, they are also a measurement tool that has been implemented in order to attain the students’ learning results [10]. A test is considered well qualified if it meets the requirements of validity and reliability [11]. Test validity refers to the compatibility between the test and the components, the result’s truthfulness and the interpretation of the test [12]. On the other hand, test reliability refers to the score stability and consistency.

In order to attain a highly qualified test instrument, a theoretical analysis in the form of an item review based on the aspects of content, construction and language should be conducted, in addition to an empirical analysis in the form of an item review. Empirical validity is attained from the Partial Credit Model (PCM) Item Response Theory (IRT) analysis. Based on these explanations, in this research, we aim to (1) generate an essay test instrument that can measure the students’ capability of diagram and argumentation representation and (2) identify the quality of the instrument that has been developed.

1.1 Diagram Representation

One of the important skills that students should master is the problem-solving skill. This skill refers to the efforts that are pursued through a cognitive manner in achieving objectives or in solving problems when there are no other alternatives available [13]. [14] stated that the objective of the learning process of physics is to generate competent problem-solvers. Therefore, through this learning process, it is expected that students will find solutions in relation to the matters that they did not understand through the activities that led to the formation of direct experience. By shaping the students’ overall attitude in solving problems and in understanding physics, it is expected that students will have a representation capability [15]. The appropriate use and selection of a representation format might improve the students’ performance in solving physics problems [16]. Representation capability entails not only a mathematical capability, but also a verbal, pictorial, and graphical.

Diagram representation refers to the presentation of physical concepts into actual pictures that look similar to daily life phenomena [17]. On the other hand, according to [13], diagram representation refers to the communication of information through the use of pictures. The process of learning physics contains numerous forms of diagrams: movement diagrams, free body diagrams, field line diagrams and wave surface diagrams, among others. Table 1 is about indicators of diagram representation.

| Table 1. Indicators of Diagram Representation. |
The 2nd Annual International Conference on Mathematics and Science Education
IOP Conf. Series: Journal of Physics: Conf. Series 1227 (2019) 012030
doi:10.1088/1742-6596/1227/1/012030

3

No Indicators of Diagram Representation
1 Describing the diagram in assisting problem-solving efforts
2 Understanding inter-diagram relationships
3 Clarifying problems/perceptions

1.2 Capability of the Argumentation
A good learning process is learning that’s able to develop the skills needed in the 21st century namely critical thinking, problem solving, collaboration and communication [18]. Generally, science learning in the classroom emphasizes practical work rather than involving students in the thinking process through a series of scientific topics such as discussion, argumentation and negotiation [19]. It is hoped that the ideal physics teacher has the capability in scientific argumentation and can deliver physics material in various forms of representation (multiple representation) to students. It was found that students had problems with connecting data to support arguments [20].

Argumentation is an important component within scientific literacy. With a good argumentative capability, students are at least able to master physical concepts. Nowadays, the argumentation of the learning process of physics in classrooms tends to be a one-way [21]. Table 2 is about indicators of capability argumentation

| No | Indicators of the Capability of Argumentation |
|----|---------------------------------------------|
| 1  | Identifying facts/statements                |
| 2  | Identifying evidence in the form of data, concepts and ideas for supporting theories/statements |
| 3  | Identifying assumptions that support theories |
| 4  | Establishing arguments and drawing conclusions |
| 5  | Observing and clarifying phenomena under observations |

1.3 Analysis of the Test Instrument
Analysis of the test instrument refers to the activities that have been performed in order to improve the quality of the test items that have been designed. The quality of the instrument might be viewed from the perspective of validity and reliability [22]. Validity reflects how far the accuracy or the scale of a test instrument is in performing its measurement function [23]. On the other hand, reliability reflects how far the measurements are trustworthy. As a result, a test instrument is considered valid and reliable when it is able to perform its measurement function.

The objective of instrument analysis is to test the quality of the test items that might provide information about the test item characteristics, namely, the strength or the weakness of the test item, based on a quantitative analysis. In accordance with the theory that has been developing up to the present time, the test item analysis in the present research is based on the Classical Test Theory (CTT) or IRT [24]. In numerous cases, IRT has many strengths in comparison to CTT. In IRT, there are supportive assumptions that might indirectly be measured and confirmed, and these assumptions are the (1) uni-dimensional trait, (2) local independence and (3) parameter invariance.

The uni-dimensional trait refers to the fact that each test item only measures one competence. Local independence refers to the absence of a correlation between the students’ responses from one to another in different test items. Parameter invariance refers to the assumption that the capability of an individual will not change only because of different tests and vice versa. Therefore, implementing the IRT in the present research was an appropriate decision. The IRT model consists of a one-parameter (1P) model, a two-parameter (2P) model and a three-parameter (3P) model [25]. The 1P model allows students to respond correctly to a test item, but on the basis that their responses are defined by the characteristics of the items from the index of difficulty level.
suggests that one of the ways to develop students’ capability is to familiarize students with the problems related to these abilities. Therefore, to improve their ability, students need to be trained and checked for understanding. One of them is using the right measuring tools to measure students abilities. This study aims to develop a feasible essay test instrument to measure the representation of diagrams and capability of argumentation. The instruments developed is in the form of a description problem which is expected to provide a significant description of the student’s ability regarding these two variables. The feasibility of the test instrument is an important part of this study, because it is the first step to determine whether the problem is made feasible or not suitable for use.

2. Method

2.1 Research Method

This research is the development research. The steps of the test instrument was adopted test instruments of Mardapi, and Oriondo & Antonio for stages of the development the test instrumen. The diagram of the models for each stage can be seen in in Figure 1. Table 3 is about blueprint instrument specifications.

| Learning Indicators                                                                 | Diagram Representation | Capability of Argumentation |
|-------------------------------------------------------------------------------------|------------------------|-----------------------------|
| Reviewing, discover and explain from Newton’s work principles.                      | D2 13                  | A1 9, 1                     |
| Insert a diagrams to solve the problem                                              | D1 8                   | A2 6, 2                     |
| Describe a diagrams (force, mass and motion) that works to object                  | D1 7                   | A3 3                        |
| D3 11, 10                                                                           |                        | A4 14                       |
| Investigate the relation between force and acceleration and the action-reaction force through simulation | D2 12                  | A4 4                        |

2.2 Data Analysis
The test instrument must have been valid and reliable criteria. This test instrument is essay to measure diagram representation and capability of argumentation on material of Newton’s Law. The validity of test instruments is obtained through content and empirical validity. Content validity is done through validation activities with expert consisting of lecturers, teachers and peers. The expert will be provided with validation assessment instruments to conduct quantitative and qualitative assessments for existing test instruments (essay) in terms of materials, constructs and languages. The quantitative assessment from the experts provide an assessment with the category score of 1 to 4 on each item, with reference to the indicators of achievement scores that have been compiled on the instrument validation assessment. The results of the quantitative assessment given will be analyzed using Aiken's V formula then can be obtained Aiken's V index:

\[ V = \frac{\sum s}{n(c-1)} \]  

Information:
- \( s = r - l \)  
- \( n = \) number of rater  
- \( l = \) the lowest score in the scoring category  
- \( c = \) number of category select  
- \( r = \) score given by rater

After the scores were attained from the experts, quantitative data were converted into qualitative data in order to identify the product quality. This conversion followed the guideline presented in Table 4.

| Table 4. Categories of Validity Quality. |  |
| --- | --- |
| Validity test results | Validity criteria |
| 0.8 < V ≤ 1 | Very good |
| 0.6 < V ≤ 0.8 | Good |
| 0.4 < V ≤ 0.6 | Moderate |
| 0.2 < V ≤ 0.4 | Poor |

When the test instrument is considered theoretically valid, it is placed into a limited experiment. This limited experiment served as an empirical validation and involved 250 students using the PCM [27]. This is politomus data because of the test instrument is essay with 4 categories (Rubric scoring instrument test: 0, 1, 2, 3). Students’ answers will be analyzed on an IRT through the QUEST and PARSACLE programs. In QUEST program will be obtained (1) Goodness of fit using PCM model: good criteria if INFIT MNSQ = 1 or on the range 0.77 – 1.30 its mean overall item and case fit with PCM model and valid; (2) Estimation of reliability: good criteria if value between 0.6 – 1.0; (3) Estimation of difficulty level: good criteria if item difficulty level was in the range -2 ≤ b ≤ +2. In the PARSACLE program is reviewed the graphical output of information functions that present the students’ ability and standard error measurement (SEM). Information gained about the students’ abilities from the output of PARSACLE program states that the test instrument is suitable for use with students on that range.
2.3 Research Subject
This research was conducted from February until March 2018. The subjects were 250 Grade XI students and two physics teachers from SMAN 8 Bulukumba, South Celebes. The Grade XI students took part in the theoretical validation, whereas the physics teachers took part in the empirical and reliability validation. In addition, the subjects also involved lecturers and peer reviewers as validators.

3. Result and Discussion
3.1 Content Validity
The quality of an instrument can be viewed from the perspective of validity and reliability. The test instrument that was developed consisted of 14 items. This instrument was intended to measure the students’ diagram representation and capability of argumentation and was designed on the basis of the 2013 curriculum. It consisted of six test items that measured diagram representation and eight test items that measured capability of argumentation. The following are the results of the theoretical validation for the test instrument from eight validators. The results are displayed in Table 5.

| Test item | Aiken’s V | Category   | Test item | Aiken’s V | Category   |
|-----------|-----------|------------|-----------|-----------|------------|
| 1         | 0.94      | Very good  | 1         | 0.94      | Very good  |
| 2         | 1.00      | Very good  | 2         | 0.94      | Very good  |
| 3         | 0.94      | Very good  | 3         | 0.88      | Very good  |
| 4         | 1.00      | Very good  | 4         | 0.94      | Very good  |
| 5         | 0.94      | Very good  | 5         | 1.00      | Very good  |
| 6         | 0.94      | Very good  | 6         | 0.88      | Very good  |
|           |           |            | 7         | 1.00      | Very good  |
|           |           |            | 8         | 1.00      | Very good  |

Based on the results displayed in Table 5 with regard to the theoretical validity and quantitative assessment by several experts (lecturers, physics teachers and peer reviewers), Aiken’s V Index was attained. The analysis was then performed on the basis of this index after being adjusted to the criteria of the index. Then, from the results of the instrument analysis, it was found that the score range of item validity is around 0.88 and 1.00 with a ‘very good’ category. In other words, the test instrument that was developed can be used by students. In addition to dealing with quantitative validity, the test instrument should also deal with qualitative validity in the form of comments and suggestions from the experts, because the results of the validity test that were analysed using Aiken’s V Index were corrected on the basis of the results of the qualitative validation. As a result, the test instrument can be tested further with empirical validation.
3.2 Goodness of fit

The empirical test involved 250 students that were selected randomly, and the results of this test were analysed using the PCM IRT. Based on the IRT analysis, as displayed in Fig.3, the output of the QUEST program for the goodness of fit showed that the goodness of fit (INFIT MNSQ) for the test items that measured the diagram representation and capability of argumentation is around 0.77 and 1.33. The test item for measuring the diagram representation Numbers 4 and 5 had a moderately good score range: 0.96 (the best score range for the goodness of fit was 1.00). On the other hand, the test item for measuring the capability of argumentation Number 9 had the best goodness of fit of all items: 0.97. Although there were several items that were closer to the ‘good’ category according to the INFIT MNSQ score, all test items for measuring the capacity of diagram and argumentation representation were around 0.77 and 1.33; in other words, all of these test items were valid.

![Figure 3. Goodness of Fit of Item](image)

3.3 Reliability

The instrument analysis with PCM IRT made use of the QUEST application so that the reliability score can be identified. The results of the instrument analysis are shown in Tables 6 and 7, and the output of the QUEST program resulted in two different kinds of reliability: item reliability and test reliability (Case estimates). From Table 6 (i.e. the results from the test items that measured the capacity of diagram representation), the item reliability score and the test reliability score were higher than 0.60, which implied that the instrument that was developed was very reliable. The item reliability score was 0.84, whereas the test reliability score was 0.79. On the other hand, from Table 7 (i.e. the results from the test items that measured the capability of argumentation), the item reliability score was 0.93, whereas the test reliability score was 0.85. As a result, it can be concluded that the test items for measuring the diagram representation and capability of argumentation fell into the ‘very reliable’ category.

| Table 6. Reliability Estimation for the Test of Diagram Representation |  |
| --- | --- |
| Reliability | Item estimates | Case estimates |
| 0.84 | 0.79 |

| Table 7. Reliability Estimation for the Test of Capability of Argumentation |  |
| --- | --- |
| Reliability | Item estimates | Case estimates |
| 0.93 | 0.85 |

3.4 The Difficulty Index

The instrument that was developed was a 1P model. Consequently, the results of the test instrument analysis were influenced by one parameter: the difficulty level. The difficulty level was analysed using the PARSSCALE application. The good criteria for item difficulty level were in the range $-2 \geq b \geq 2$. In Figure 4, the $x$-axis represents each item and the $y$-axis represents the item difficulty level. The results
that are attained in Figure 3 showed that 14 items that were analysed had a good difficulty level because the score of these 14 items ranged from −0.82 and 0.80 (there were no items smaller than −2 or greater than 2). It can be seen from Figure 4 that the easiest test item is test item number 1, whose score was −0.82, whereas the most difficult test item is test item number 13, whose score was 0.80. The easiest test item, item number 1, was part of the test instrument that measured the diagram representation, whereas the most difficult test item, item number 13, was part of the test instrument that measured the capability of argumentation.

3.5 Information Function and SEM

In Figure 5, the x-axis displays the score of the students’ capacity and the y-axis displays the SEM. It was found that the test instrument that was developed ranged between -2.54 and 1.92 with an SEM of 0.22. The results of the IRT analysis using the QUEST program and the PARSCALE application were the goodness of fit on the PCM (item validity), item and test reliability and also difficulty level and information function (SEM) that observed the development of instrument quality for measuring the diagram representation and capability of argumentation.

4. Conclusions

(1) The procedures of model development for the instrument follow the stages of instrument development and include test planning, construct of the test and test experiments.
(2) The test instrument is considered valid according to the qualitative and quantitative validity. Therefore, the test instrument can be put into experiment.
(3) The test instrument is already compatible with the PCM.
(4) The test instrument has good reliability because the coefficient of reliability ranges from 0.79 to 0.85.
(5) The item difficulty level within the instrument falls into the ‘good’ category because the level ranges between −2 and +2.

5. References
[1] Ding, L., Reay, N., Lee, A., & Bao, L. (2011). Exploring the Role of conceptual Scaffolding in Solving Synthesis Problems. *Physics Education Research*, 7 (2): 1–11.
[2] Tolga Gok. (2010). The General Assessment Of Problem Solving Processes And Metacognition In Physics Education. *Eurasian Journal Of Physics Adn Chemistry Education*, 2 (2): 110–122.
[3] Mundilarto. (2001). Evaluasi Terpadu dalam Pembelajaran Fisika. Makalah, Yogyakarta State University.
[4] De Cock, M. (2012). Representation Use and Strategy Choice in Physics Problem Solving. *Physical Review Special Topics - Physics Education Research*, 8, 020117.
[5] Cheng, M., & Gilbert, J. K. (2014). Towards a better utilization of diagrams in research into the use of representative levels in chemical education. *Springer Netherlands, in Multiple representations in chemical education* (pp. 55-73).
[6] Etkina, E., & David T. B. (2011). In search of alignment:Matching learning goals and class assessment. *AIP Conferences Proceedings*.
[7] Ayesh, Q. N., Ti, N., & Abdelfattah, F. (2010). The effect of Student Use of Free Body Diagram Representation on Their Performance. *Educational Research*, 1 (10): 505–511.
[8] Song, Y., & Deane, P. (2014). A case study in principled assessment design: designing assessment to measure and support the development of argumentative reading and writing skills. *Psicologia Educativa*.
[9] Harvey, M., Walkerden, G., Semple, A.L., McLachlan, K., Lloyd, K., & Baker, M. (2016). A Song And Dance: Being Inclusive and Creative In Practicing and Documenting for Learning. *Journal of University Teaching & Learning Practice*, 13(2), 3.
[10] Widoyoko, E. P. (2014). Evaluasi Program Pembelajaran. Yogyakarta: Pustaka Pelajar
[11] T. Hidayat, E. Susilaningshih, and K. Cepi. (2018). The Effectiveness of Enrichment’Test Instrumens Design to Measure Students’ Creative Thinking Skills and Problem-Solving. *Think. Ski. Creat.*
[12] Angell, C., Kind, P. M., Henriksen, E. K., & Ø. Guttersrud. (2008). An empirical-mathematical modelling approach to upper secondary physics, *Physics Education*, 43 (3): 256–264.
[13] Liliarti, N., & Kuswanto, H. (2017). Pengembangan Perangkat Pembelajaran Fisika Berbasis Kearifan Lokal (Kapal Othok-othok) Berbantuan Android untuk Meningkatkan Kemampuan Representasi Diagram dan Argumentasi. *Tesis*. Universitas Negeri Yogyakarta.
[14] Nieminen, P., Savinainen, A., & Viiri, J. (2012). Relations Between Representational Consistency , Conceptual Understanding Of The Force Concept , And Scientific Reasoning, *Physical Review Special Topics-Physics Education Research*, 10123: 1–10.
[15] Docktor, Jennifer L & Mestre, Jose P. (2014). Synthesis of Discipline-based Education Research in Physics. *Physics Education Research*, 10.
[16] Kohl, P. B. (2007). “Towards an Understanding of How Students Use Representations in Physics Problem Solving” pp. 1–355.
[17] Widianingtyias, L. & Bakri, F. (2015). “Pengaruh Pendekatan Multi Representasi dalam Pembelajaran Fisika Terhadap Kemampuan Kognitif Siswa SMA” 1: 31–38.
[18] Kulsum, N. (2014). Penerapan Model Pembelajaran Cooperative Problem Solving untuk Meningkatkan Kemampuan Pemahaman Konsep dan Komunikasi Ilmiah Siswa pada Mata Pelajaran Fisika. *Unnes Physics Education Journal*. 
[19] Roshayanti, F. 2012. Pengembangan Model Asesmen Argumentatif Untuk Mengukur Keterampilan Argumentasi Mahasiswa pada Konsep Fisiologi Manusia. Disertasi. Universitas Pendidikan Indonesia, Bandung.

[20] Acar, O., L. Turkmen., & A. Roychoudhury. (2010). Student Difficulties in Socio-Scientific Argumentation & Decision-Making Research Findings: Crossing the Borders of two Research Lines. *International of Science Education*. 32 (9):1191-1206.

[21] Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing The Quality of Argumentation In School Science. *Journal Of Research In Science Teaching*, 41 (10): 994–1020.

[22] Anthony, J. N. (1996). “Educational Assessment of Students” (Merril an Imprint of Prentice Hall Englewood Cliffs, Ohio).

[23] Azwar, S. (2012). *Reliabilitas dan Validitas*. Pustaka Pelajar:Yogyakarta

[24] Mardapi, D. (2008). *Teknik Penyusunan Instrumen Tes dan Nontes*. Mitra Cendekia Press:Yogyakarta, p.5. (4).

[25] Retnawati, H. (2014). *Teori Respons Butir dan Penerapannya*. Yogyakarta: Parama Publishing.

[26] Tandilling, Edy. (2012). Pengembangan Instrumen untuk Mengukur Kemampuan Komunikasi Matematik, Pemahaman Matematik, dan Self-Regulated Learning Siswa dalam Pembelajaran Matematika di Sekolah Menengah Atas. *Journal Penelitian Pendidikan*. Vol. 13 No.1, p.24-31

[27] Adams, R. J. & Khoo, S. T. (1996). Quest: The interactive test analysis system version 2.1. Victoria: The Australian Council for Educational Research.