Constructing Reasoning Multiple Choice Test to Measure Bloomian Higher Order Thinking Skills in Physics of XI Grade Students

Muthmainnah* and Edi Istiyono
Yogyakarta State University, Indonesia

*Corresponding Author: muthmainnah.2016@student.uny.ac.id

Abstract. A commonly used assessment to measure students' abilities is to provide a simple multiple-choice test or a brief entry. It is not able to accurately measure students' abilities. It is necessary to develop a test which capable to measure students’ ability accurately, especially their high-level thinking (HOTS) in physics learning. The purpose of this study is to develop a test to measure Bloomian HOTS in Physics. The participant of this research were XIth grade students purposively selected based on the recent UN score Bloomian HOTS includes the ability to analyse, evaluate and create in subject material of equilibrium of rigid body, elasticity and Hooke’s law, static fluid, dynamic fluid, and temperature and heat. The test developed in the form of reasoning multiple choice (RMC) items, which requires students to not only think about the answers but also the reasons for the answers. The development model used is adapted from the Wilson and Oriondo-Antonio’s model, which consists of three stages: 1) designing the test, 2) validating the test, and 3) assembling the test. Four sets of tests have been assembled, each consisting of 40 items with 8 anchor items. Validation was divided into two, by expert and trial test. The result of expert’s judgement obtained value of Aiken V was 0.87 to 0.93. The result of trial test was analysed using Item Response Theory (IRT). The results of this study indicated that the RMC test that has been developed valid and reliable to be used to measure Bloomian HOTS in Physics of XI grade students.

Keywords: Assessment; Bloomian HOTS; RMC; IRT; Physics test.

1. Introduction
Physics is a unique discipline that is difficult to teach and even harder to understand [1] so that it can be considered as HOT, or Higher Order Thinking [2]. Therefore, an educator is also required to have a Higher Order Thinking Skills (HOTS) so that it can carry out learning physics [3] [4] [5] and the assessment of student’s right skills and appropriate [6]. Further, the assessment is one of the important processes in the learning process [7] [8]. Students are expected to have the ability to think critically, [9], especially HOTS Bloomian covering analyse, evaluate, and create [2] [10] [11] [12] because of Bloom's Taxonomy is widely applied in learning process in Indonesia.

Multiple choice test is the test most commonly used [13], because these tests can be used to test the mastery of materials with a broad scope [9]. In addition, the scoring of multiple choice tests easier. However, multiple choice tests in general are sometimes less able to measure students HOTS [14]. Therefore, it is necessary to develop a test that requires students to reasoning [15]. The multiple-
choice tests reasoned (Reasoning Multiple Choice or RMC) is capable of demanding students to think about the reasons underlying their choice in answering test.

The research purpose is to develop a test to measure Bloomian HOTS in Physics of XI grade students, which includes the ability to analyse, evaluate and create in subject material of equilibrium of rigid body, elasticity and Hooke’s law, static fluid, dynamic fluid, and temperature and heat.

The rest of this paper is organized as follow: Section 2 describes the proposed research method. Section 3 presents the obtained results and following by discussion. Finally Section 4 concludes this work.

2. Research Method
This section describes the proposed research method.

2.1. Development Model
This research is a development research. The development model used is adapted from the Wilson and Oriondo-Antonio’s model [16], which consists of three stages: 1) designing the test, 2) validating the test, and 3) assembling the test. The product of this research is four set of RMC test in Physics.

The first stage of development was designing the test. In this stage, the items test was constructed. The items were constructed based on the contents of Physics syllabus 2013 [17] in Indonesia and integrated with Bloomian HOTS aspects and sub aspects [18]. Bloomian HOTS aspects are analysing, evaluate and create. Sub aspects of analyse were distinguish, sort and articulate. Sub aspect of evaluate were check and criticize. And sub aspects of create were formulate, plan and produce.

The second stage of development was validating the test. Validating proses was divided into expert judgement and trial test. The experts will have judge the content and construct of the preliminary test. After passed the experts judgement, the test will have been tested to

2.2. Participants
The participate of this research were 298 of XIth grade students purposively selected from three high schools in Yogyakarta, Indonesia. The three high schools were selected based on the result of the recent national examination score. The time were during the second semester of academic year 2017-2018.

2.3. Data Analyze Technique
Data analyze were used to obtained the item’s characteristic and the test’s reliability. Data obtained from expert were analyzed using Aiken V method [19], and trial test result were analyzed using Quest and Parscale program.

3. Result and Discussion
The items test was constructed based on the contents of Physics syllabus 2013 and been integrated with Bloomian HOTS aspects and sub aspects. The origin Bloom HOTS in Bloom taxonomy were analysis, synthesis and evaluation. Then those aspects were revised by Anderson and Krathwol [10], and became analyze, evaluate and create as Bloom HOTS or in this research been called Bloomian HOTS. There were 136 items were constructed covering five materials in Physics of XI grade second semester. The test items were divided based on aspects and sub aspects of Bloomian HOTS. The number items of each aspects and sub aspects are shown in Table 1.

| Aspects | Sub Aspects | Number Items |
|---------|-------------|--------------|
| Analyze | Distinguish | 19           |
|         | Sort        | 17           |
|         | Articulate  | 16           |
The items form was modified multiple choice. An item contained the main question, five choices of answer, and five choices of reasoning answer. An example of an item test is shown in the Appendix. The student had to choose one of five choices of answer and one of five choices of reasoning answer for each item. Therefore, the student chose two right answers for each item. The scoring method of answer were based on the answer and reasoning which had been chosen by students. An item will be scored 4 if the answer and reasoning answer were correct. An item will be scored 3 if the answer was incorrect but reasoning answer was correct. An item will be scored 2 if the answer was correct but reasoning answer was incorrect. And last, an item will be scored 1 if the answer and reasoning answer were incorrect.

The preliminary items have been validated by five experts. The experts validated the suitability of all 136 items to be used as a measure instrument in Physics, especially to measure Bloomian HOTS in Physics. The validity aspect included content, construct, language and appearance of the test items. The V Aiken score obtained from expert judgments were 0.87 to 0.93. This indicated that all of 136 items of Bloomian HOTS test were good.

The items of test were divided into four sets to be tested in trial test. Each sets contained forty items with eight anchor items. The four sets have been tested in three schools. The result has been analyzed to obtained item characteristic.

The analysis of statistical fit is a check on the internal validity. The internal validity of test was assessed in terms of the statistical fit of each item to the model. The form of the item in the integrated assessment is RMC test items with five possible choices of answers and five possible choice of reasons. The data was in the polytomus, which was appropriate to be analyzed using the Partial Credit Model 1 Parameter Logistic (PCM 1-PL) in IRT. Therefore, the model in this research referred to the PCM 1-PL. The item fit describes whether the item is functioning normally in measuring or not. If the item is outliers or misfits, it indicates that there is a misconception on the students to the item. According to Korashy [20], if the fit statistic of an item is acceptable, then the item can be said as a valid item. The criteria used for the suitability of the outliers was if the value of infit mean square (MNSQ) is 0.50 to 1.00 and standard deviation around 0.00. In this research, the data analyzed by using Quest and Parscale program. Based on result of data analyze, the value of infit MNSQ was 0.99 ±0.09 and standard deviation was 0.00 ± 0.48. The result shown the value of goodness of fit of all items were in interval 0.83 to 1.27. It can be said that all items were fit to the models. Distribution of test items for items 1 to 23 according to Quest's output is more clearly shown in Figure 1.
Figure 1. Distribution of first 23 items of 136 items according to Quest

There were no items were discarded on this analysis. It indicated that each item has been fit with the model of PCM 1-PL. Therefore, all items were feasible to be used. In other words, all items can be said to be the final product which has been proved valid empirically.

Further analysis was to determine the difficulty level of items. The items difficulty level was conducted to establish an additional measure of the characteristic items. The difficulty level was computed to get an idea of the proportion of the test takers correctly responding to the item. The interpretation of level for difficulty according to Adedoyin and Mokobi [21] with addition. The items are categorized very easy if the value of b (measure) is less than -1. Items are categorized easy if the value of b (measure) at interval $-1 < b < -0.5$. Items are categorized medium if the value of b (measure) at intervals of $-0.5 < b < 0.5$. Items are categorized difficult if b at interval $0.5 < b < 1$. Items are categorized very difficult if b is more than 1. The difficulty level of the 136 items at interval $-1.29$ to $1.14$. There were ten items with highest difficulty level as shown in Table 2.

Table 2. Ten items with highest difficulty level

| No | Item Number | $b$   |
|----|-------------|-------|
| 1  | 129         | 1.14  |
| 2  | 77          | 1.13  |
| 3  | 100         | 1.10  |
| 4  | 122         | 1.07  |
| 5  | 5           | 1.01  |
| 6  | 45          | 0.81  |
| 7  | 118         | 0.81  |
| 8  | 70          | 0.76  |
| 9  | 27          | 0.74  |
| 10 | 29          | 0.74  |

Based on the data in the Table 2, it can be concluded that item 129 has the highest level of difficulty, which is equal to 1.14. The item with the level of difficulty was an item of create aspects and sub aspects of planning. The item was item number 1 in set D. And ten items with lowest difficulty level as shown in Table 3.
Table 3. Ten items with lowest difficulty level

| No | Item Number | b   |
|----|-------------|-----|
| 1  | 23          | -1.29 |
| 2  | 103         | -1.17 |
| 3  | 24          | -1.00 |
| 4  | 116         | -1.00 |
| 5  | 87          | -0.97 |
| 6  | 55          | -0.93 |
| 7  | 12          | -0.82 |
| 8  | 25          | -0.81 |
| 9  | 26          | -0.72 |
| 10 | 47          | -0.67 |

Based on the data in the Table 3, it can be concluded that item 23 has the lowest difficulty level, which is equal to -1.29. The item with the lowest level of difficulty was an item of analyze aspects and sub aspects of distinguish. This item was item number 18 in set A.

All items’ difficulty level was distributed normally. Based on difficulty categories, the number item with the very easy category was six items with the percentage of 4%. The easy category was fourteen items with the percentage of 10%. The medium category was ninety-six items with the percentage of 71%. The difficult category was thirteen items with the percentage of 10%. And last, the very difficult category was seven items with the percentage of 5%. Overall, the difficulty level of the item was divided in five categories as presented in Figure 2.

![Figure 2. Percentage of difficulty categories of Bloomian HOTS test items](image)

All items were developed based on aspects and sub aspects of Bloomian HOTS. Therefore, the difficulty level for each aspect and sub aspects of Bloomian HOTS were count. The result shown in table 4.
Table 4. Items Difficulty Level Mean based on Bloomian HOTS Aspects and Sub Aspects

| Aspects | Sub Aspects | Difficulty Level (b) | Category |
|---------|-------------|----------------------|----------|
|         |             |                      | 1 (%)    | 2 (%)    | 3 (%)    | 4 (%)    |
| Analyze | Distinguish | -0.02                | 41.8     | 18.5     | 9.3      | 30.3     |
|         | Sort        | -0.35                | 24.1     | 13.1     | 23.6     | 39.2     |
|         | Articulate  | -0.08                | 34.1     | 17.4     | 21.0     | 27.6     |
| Evaluate| Check       | 0.09                 | 37.7     | 21.7     | 18.4     | 22.1     |
|         | Criticize   | -0.03                | 32.6     | 16.1     | 24.2     | 27.1     |
| Create  | Formulate   | 0.25                 | 44.0     | 11.2     | 25.9     | 18.9     |
|         | Plan        | 0.06                 | 34.4     | 10.5     | 34.5     | 20.5     |
|         | Produce     | 0.10                 | 43.2     | 14.5     | 22.2     | 20.1     |

Based on Table 4, it can be seen that the difficulty level data is more detailed for each aspect and sub aspects. The table also shown the percentage of students' answers in each scoring category for each aspects and sub aspects. If the percentage of students answered in category 1 at most, then the item was difficult. Conversely, if the percentage of students answers in category 4 at most, then the item was easy. The items that are categorized as the most difficult are in the create aspect and sub aspects of formulate with difficulty level of 0.26 and category 1 of 44.0 %. The items that are categorized as the easiest are easily found in the analyze aspect and sub aspect of sort with difficulty level of -0.35 and category 4 of 39.2%.

Based on analyzing data using Quest obtained item estimate score was 0.87 and case estimate score was 0.86. Both were above 0.80 and accepted. It can be concluded that all items were reliable. Parscale output shown the relationship between information functions and SEM as shown in Figure 3.

![Figure 3](image.png)

Figure 3. The relationship between information functions and SEM

A good relationship between information functions and SEM is inversely proportional, where based on the results of research in figure 3. This is in accordance with the theory of Hambleton and Swaminathan [22] which states that if the SEM relationship and the function of true information are inversely proportional. The higher the SEM, the lower the information function will be obtained, and vice versa. From this relationship it will be known that the test can be used to measure the ability of
students in the range of -2.0 to 2.0 (-2.0 ≤ θ ≤ 2.0). Therefore, it can be said that the test can be done by students with moderate to high abilities and categorized into tests that require Higher Order Thinking Skills. This developed Bloomian HOTS test have been proven to be able to measure students’ Bloomian HOTS in Physics. By using this kind of test, the students were not only think about a certain answer to solve the question but it is also needed think about the rational reasons (reasoning) for the answer they have choose. This test was better than a common multiple choice test. However, this test still has its deficiency. The test can only provide a closed-answer form as rational reasoning choices. This deficiency cannot cover the whole mindset of the student when they answering questions.

4. Conclusion
This paper has presented constructing reasoning multiple choice test to measure Bloomian higher order thinking skills in physics of XI grade students. The result of this research ascertained that the developed Bloomian HOTS test which contained 136 items has relatively high validity and reliability. The value of V Aiken is at interval 0.83 to 0.97. The test is suitable to measure the students’ higher order thinking skill. There are three aspects and eight sub aspects of Bloomian HOTS in Physics that can be measure by using this test. The valid test can be applied in assessment proses to measure Bloomian HOTS in Physics of XIth grade students.

Acknowledgment
Acknowledgments are conveyed to the student’s senior high schools in region Yogyakarta, Indonesia, who have helped the trial test of Bloomian HOTS test in Physics and to five experts who helped researchers to developed a valid test.

References
[1] Mustansyir R and Munir M 2015 Filsafat Ilmu (Yogyakarta: Pustaka Pelajar)
[2] Ramos J L S, Dolipas B B and Villamor B B 2013. Higher Order Thinking and Academic Performance in Physics of College Students: A Regression Analysis. International Journal of Innovative Interdisciplinary Research 4: 48-60
[3] Adawati 2014 Exploring Teachers’ Experiences on Integration of Higher Order Thinking Skills (HOTS) in Teaching of Science Journal of Southeast Asian Education 2(1): 1-21
[4] Yen T S and Halili S H 2015 Effective Teaching of Higher-Order Thinking (HOT) in Education. The Online of Distance Education and e-Learning 3(2): 41-47
[5] Saputra H 2016 Pengembangan Mutu Pendidikan menuju Era Global: Penguatan Mutu Pembelajaran dengan Penerapan HOTS (Higher-Order Thinking Skills) (Bandung: Smart Learning Solution)
[6] Niepel C, Mustafic M, Greiff S, and Roberts R D 2015 The dark side of creativity revisited: Is students’ creativity associated with subsequent decreases in their ethical decision making? Thinking Skills and Creativity :1-10.
[7] Atkin J M and Coffey J E 2008 Everyday Assessment in the Science Classroom (Virginia: NSTA Press)
[8] Haviland D 2013 Building s Culture of Faculty-Owned Assessment College Science Teachers Guide to Assessment (4):7-14
[9] Evans J and Stanovich K E 2013 Dual process theories of higher cognition: Advancing the debate Perspectives on Psychological Science 8: 223-241
[10] Anderson L W and Krathwohl D R 2010 Kerangka Landasan untuk Pembelajaran, Pengajaran, dan Asesmen: Revisi Uksonomi Pendidikan Bloom (Yogyakarta: Pustaka Pelajar)
[11] Shaheen R 2010 Creativity and education Creative Education 1(3):166-169
[12] Assaly I R and Smadi O M 2015 Using Bloom’s Taxonomy to Evaluate the Cognitive Level of Master Class Textbook’s Questions English Language Teaching 8(5):100-110
[13] Sani R A 2016 Penilaian Autentik (Jakarta: Bumi Aksara)
[14] Piaw C Y 2010 Building a Test to Assess Creative and Critical Thinking Simultaneously. Procedia Social and Behavioral Sciences 2: 551-559
[15] Zohar A 2013 Scaling Up Higher Order Thinking in Science Classrooms: The Challenge of Bridging The Gap between Theory, Policy and Practice Thinking Skills and Creativity 10: 168–172.
[16] Oriondo L L and Antonio E M D 1984 *Evaluating Educational Outcomes (Test, Measurement, And Evaluation)* (Manila: Rex Book Store)

[17] Ministry of Education and Culture Indonesia 2013 *Regulation of Ministry of Education and Culture Number 66 in 2013 regarding the Standard for Education Assessment* (Jakarta: Ministry of Education and Culture Indonesia)

[18] Miri B, David B and Uri Z 2007 Purposely Teaching for the Promotion of Higher-Order Thinking Skills: A Case of Critical Thinking *Journal Res Sci Educ* 37: 353-369.

[19] Aiken L R 1985 Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement* (45), 131-142.

[20] Korashy A F 1995 Applying the Rash model to the Selection of items for mental ability test. *Educational and Psychological Measurement*, 55(5), 753-763.

[21] Adedoyin O O and Mokobi T 2013 Using IRT Psychometric Analysis in Examining the Quality of Junior Certificate Mathematics Multiple Choice Examination Test Items *International Journal of Asian Social Science* 3(4), 992-1011.

[22] Hambleton R K and Swaminathan H 1985 *Item Response Theory Principles and Applications* (Boston: Kluwer Nijhoff Publising)
Appendix

Item number 15 set A

15. Rina diminta untuk melakukan percobaan mengenai hukum Hooke. Alat dan bahan yang
disediakan adalah statif, pegas, beban dan alat ukur. Rencana langsung kerja yang
benar yang akan diikuti Rina adalah ... .
A. Menggunakan statif pada pegas – Menggunakan belon pada statif – Menarik beban –
   Memukul pertumbuhan panjang pegas
B. Menggunakan belon pada pegas – Melihat pertumbuhan panjang pegas – Mengukur
   panjang statif – Mengukur pertumbuhan panjang pegas
C. Mengukur panjang pegas – Mengukur panjang statif – Menggunakan belon pada pegas –
   Memukul pertumbuhan panjang pegas dan statif
D. Memasang pegas pada statif – Menggunakan belon pada pegas –
   Memukul pertumbuhan panjang pegas
E. Mengukur panjang statif – Menggunakan pegas pada statif – Menggunakan belon pada
   pegas – Mengukur pertumbuhan panjang pegas

Alasan
A. Panjang statif mempengaruhi panjang pegas
B. Pegas tidak memiliki panjang mulia-mula
C. Belon mempengaruhi panjang pegas
D. Belon mempengaruhi panjang statif
E. Belon tidak dipengaruhi gaya gravitasi bumi

Item number 17 set B

17. Dua benda ujung berbeda massa (A dan B) dipanaskan selama 23 menit seperti pada
   gambar. Kemudian kedua ujung dimasukkan ke dua wadah yang berbeda yang beris
   dengan volume dan suhu yang sama. Setelah beberapa lama, sudu air pada kedua wadah
diukur, dan dinyatakan bahwa suhu kedua suhu sama. Jika dikaitkan dengan konsep
   perpindahan kalor, maka penjelasan tersebut ...

A. Salah. Benda A suhunya lebih rendah karena lebih sedikit menyerap kalor saat
   dipanaskan
B. Salah. Benda A suhunya lebih tinggi karena lebih sedikit menyerap kalor saat
   dipanaskan
C. Salah. Benda B suhunya lebih rendah karena menyerap banyak kalor saat dipanaskan
D. Benar. Suhu pada wadah benda A dan B sama, karena kalor yang diserap benda seta
   sama
E. Benar. Suhu pada wadah benda A dan B sama, karena suhu yang dipindahkan benda
   sama

Alasan
A. Massa benda tidak mempengaruhi besarnya kalor yang diserap benda
B. Benda bernama lebih kecil lebih banyak melekas kalor ke lingkungan disekitarnya
daripada benda bernama besar
C. Massa benda mempengaruhi besarnya kalor yang diserap benda, semakin besar massa
   benda, semakin banyak kalor yang diserap
D. Benda bernama kecil lebih cepat menyerap kalor daripada benda bernama besar ketika
   dipanaskan
E. Benda bernama lebih besar lebih cepat melepas kalor daripada benda bernama kecil
   ketika dipanaskan