The Development of Higher Order-Thinking Skills (HOTS) Instrument Assessment in Physics Study

K Arafah¹, B D Amin¹, S S Sari¹, and A Hakim¹

¹ Physics Department, Universitas Negeri Makassar, Jl. Daeng Tata Raya, Makassar, Indonesia

Abstract. This study aims to determine the characteristics of the test instruments in the mechanics of SMA to measure higher order thinking skills. This study also aims to see the validity, reliability, level of difficulty, and differentiation of items using the classical theory approach. This study adapted the development model from Mardapi. The items developed were tested on 114 students at one of High School in Makassar, South Sulawesi. The results show that the items that have been developed have content validity with a coefficient of 0.860. Of the 35 test items developed, there were 7 invalid items, with a test reliability of 0.593. The proportion of item difficulty level has the distribution of 0% hard items, 74.3% medium items, and 25.7% easy items. The distinguishing power of test items has shown good criteria.

1. Introduction

The most recent revision in the 2013 Curriculum that was put in effect was focused on improving two major parts of the curriculum, namely content standards and assessment standards. The content standard is designed so that students are able to think critically and analytically in accordance with international standards which is done by reducing irrelevant material and deepening and expanding relevant material for students, while the assessment standard is carried out by adapting the international standard assessment model gradually. Assessment of learning outcomes focuses more on Higher-Order Thinking Skills (HOTS).

Curriculum refinement to improve students’ high-order thinking skills is carried out based on the results of the International Program for International Student Assessment (PISA) study which shows data that achievement of reading literacy, mathematical literacy, and scientific literacy of participants in Indonesian education is very low [1–3]. This is the reason for the need for system changes in learning.

Higher order thinking skills can be trained in the learning process in the classroom. Learning that is carried out must provide space for students to find activity-based and meaningful knowledge concepts. Therefore, the 2013 K revision emphasizes having to integrate (Higher Order Thinking Skills / HOTS) in learning. This shows that learning must provide training not only for students' basic conceptual understanding of learning, but also for students' high-level abilities. Furthermore, to facilitate understanding in this article HOTS will write Higher Order Thinking Skills. But in fact, the application of HOTS learning is not easy for teachers to do. The teacher must really master the material and learning strategies and the teacher is also faced with challenges with the student environment. Learning will be meaningful if students are invited to think in higher order. The success of mastering a concept will be obtained when students are able to think at a high level, where students can not only remember and
understand a concept [4], but students can analyze and synthesize, evaluate, and create a concept well, the concept that has been understood can be embedded in memory students for a long time, so it is very important for students to have higher order thinking skills or HOTS.

The importance of assessment in the learning process is something that needs serious attention, considering that assessment is an integral part of the learning process [5–7]. Assessment is one of the activities carried out by teachers and students from a series of learning activities carried out. Teachers as learning managers are required to be able to compile and conduct assessments with correct procedures so that the learning objectives set are achieved [8,9].

The revised Bloom’s Taxonomy divides the thinking process into two, namely higher-order thinking skills or often referred to as Higher Order Thinking Skills (HOTS), and lower-order thinking skills, Lower Order Thinking Skills (LOTS). Low-order thinking skills involve the ability to remember (C1), understand (C2) and apply (C3) while in higher-order thinking skills involve analysis and synthesis (C4), evaluate (C5), and create or creativity (C6) [5,8,10–13].

Table 1. Cognitive Level and descriptors

| Cognitive Level | Description |
|-----------------|-------------|
| **Creating**    | Verbs: construct, design, create, develop, write, formulate. |
| **Evaluating**  | Make decisions |
| **Analyzing**   | Specifying aspects / elements. |
| **Applying**    | Use information on a different domain |
| **Understanding** | Explain the idea / concept. |
| **Knowing**     | Recalling. |

HOTS is very important to be applied and developed in learning. When learning activities focus on HOTS development targets, it greatly affects learning activities that are more effective, the intellectual abilities of teachers and students become more trained [14–16]. Evaluating HOTS requires questions that cannot be answered simply by students which will certainly have an effect on improving the quality of education.

The description above emphasizes the need for an assessment instrument to evaluate HOTS. In connection with the matters described above, this study is aimed at developing knowledge test items to measure higher order thinking skills.

2. Experimental

This type of research is Research and Development, which aims to develop test items for high school mechanics that can measure higher order thinking skills. The development of high school mechanics knowledge test items will follow the steps suggested by Mardapi [17]. The steps to be taken are: compiling test specifications, writing test items, reviewing test items, conducting test trials, analyzing test items, improving tests, assembling tests, and carrying out tests, and interpreting test results, as shown in the chart below.
The location of this research is in the Department of Physics, Faculty of Mathematics and Natural Sciences as a place for item development. The trial location for data collection was carried out at SMAN 8 Makassar. The characteristics of the test items in question are the distinguishing power and the level of difficulty. The level of problem difficulty is the opportunity to answer correctly a question at a certain ability level which is usually expressed in the form of an index. To calculate the difficulty level of the multiple choice form test is as follows [6,18]:

\[
TK = \frac{(WL + WH)}{(nL + nH)} \times 100\%
\]

Information: TK = Difficulty Level; WL = the number of students who answered incorrectly from the lower group; WH = the number of students who answered incorrectly from the top group; nL = number of lower groups; nH = number of upper groups.

The difficulty level criteria according to Arifin [6,11], are as follows.

| Percentage   | Criteria |
|--------------|----------|
| 0% – 27%     | Hard     |
| 28% – 72%    | Medium   |
| 73% – 100%   | Easy     |

The distinguishing power of a question is the ability of an item to distinguish between students who have good higher-order thinking skills and students who do not have good higher-order thinking skills. The calculation of the distinguishing power of multiple choice form questions is as follows:

\[
DP = \frac{(WL - WH)}{n}
\]

Information: DP = Discriminatory Power; WL = number of students who failed from the lower group; WH = number of students who failed from the top group; n = 27% x number of students.
The validity of the instrument content was obtained through expert judgment using the Gregory Internal Consistency Coefficient formula, as well as a qualitative analysis using panel techniques to obtain input / suggestions from experts regarding the preparation of instruments from the construction and language aspects.

Furthermore, the items were tested empirically to determine the validity of the criteria. In classical test theory, the scoring system that produces a score of 0 and 1 is known as a dichotomy score. In giving a score, the most important thing to pay attention to is the respondent's answer. For a scoring technique that follows no penalty, the correct answer is given a score of 1, while the wrong answer is given a score of 0. An example of this scoring technique is the multiple choice, true-false, and match-making test items. To calculate the validity of the items, the point biserial correlation formula is used [7] as follows.

$$ r_{pbi} = \frac{M_p - M_q}{s_t \sqrt{pq}} $$

Information: $r_{pbi} = \text{point biserial correlation coefficient}; M_p = \text{number of respondents who answered correctly}; M_q = \text{the number of respondents who answered incorrectly}; s_t = \text{standard deviation for all items}; p = \text{proportion of respondents who answered correctly}; q = \text{proportion of respondents who answered correctly}.$

This biserial point correlation coefficient is then compared with the criteria, which have been prepared in the table according to the level of reliability and the number of test subjects. The greater the number of test subjects involved, the smaller the value of the table correlation coefficient criteria. If the biserial point correlation coefficient obtained is greater than the criteria ($r_{table}$), then the items are said to be valid, and vice versa.

After the items are declared valid, then in stating the good and usable items is to determine the reliability status. The reliability coefficient can be calculated by removing all items that are declared invalid, then taking all items that are declared valid. Items scored using a dichotomy approach can be calculated using the Kuder-Richardson (KR) -20 formula. The KR-20 formula was first published in 1937, developed by Kuder and Richardson. The name of this formula comes from the fact that it is the twentieth formula discussed in Kuder and Richardson's seminal paper on reliability testing. The KR-20 is often called the Cronbach alpha special case, calculated for the dichotomy score. The KR-20 formula [7] can be presented as below.

$$ r_{ii} = \left( \frac{n}{n-1} \right) \left( \frac{s_t^2 - \sum qp}{s_t^2} \right) $$

Information: $r_{ii} = \text{instrument reliability coefficient}; n = \text{Number of items}; s_t^2 = \text{Total variance}; p = \text{proportion of subjects who answered correctly on item-i}; q = \text{the proportion of subjects who answered incorrectly on item-i (q = 1 - p)}.$

3. Result and Discussion
The knowledge test to measure higher order thinking skills in high school mechanics material developed refers to three indicators, namely analyzing, evaluating, and creating. Where the description for each indicator can be seen in Table 1. The distribution of test items for each indicator can then be seen in Table 3 below.
The results of the instrument trial data processing in determining the difficulty level of each item are shown in the following table.

**Table 4** Results of the Calculation of Problem Item Difficulty Levels

| Category     | Item Number | sum | Percentage (%) |
|--------------|-------------|-----|----------------|
| Easy         | -           | 0   |                |
| Medium       | 1, 3, 5, 6, 7, 9, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35 | 26 | 74.3%          |
| Hard         | 2, 4, 8, 10, 11, 12, 14, 24, 27 | 9  | 25.7%          |

From Table 4 it can be seen that of the 35 items there are 0 items in the easy category, 26 items in the medium category, and 9 items in the difficult category. A good question is one that has a difficulty level, neither easy nor difficult. So that the items that have a good difficulty level are 26 items. The percentage of the number of instrument items based on the level of difficulty is that there are 0% easy questions, 74.3% medium questions, and 25.7% difficult questions. The results of the calculation of grain differentiation can be seen in the table 5.

**Table 5. Categorization of Item Distinguishing Power**

| Category                          | Item Number | Sum | Percentage (%) |
|-----------------------------------|-------------|-----|----------------|
| very good question                | 1, 5, 10, 25, 29, 31 | 6   | 17.1           |
| good question, but may need to be fixed | 3, 7, 13, 16, 22, 30, 32, 33, 35 | 9   | 25.7           |
| enough question, needs to be fixed | 2, 6, 14, 15, 17, 20, 23, 24, 27, 28, 34 | 11  | 31.4           |
| bad questions, discarded or revised | 4, 8, 9, 11, 12, 18, 19, 21, 26 | 9   | 25.7           |

Based on Table 5, it can be seen that of the 35 items, there are 6 items that are classified as very good; 9 items are classified as good, but may need to be corrected; 11 items are in the sufficient category and need to be corrected, and 9 items are bad so they are discarded or revised. The percentage of the number of items based on the distinguishing power category was that 17.1% of the items were very good; 25.7% good questions; 31.4% enough questions; and 25.7% bad questions.

Based on the expert's assessment using a rating scale of 4, the inter-expert agreement model for content validity can be filled as follows [6,7].
The amount of the internal consistency coefficient by the expert is calculated using the following equation:

\[
\text{Content validity} = \frac{D}{(A + B + C + D)} = \frac{30}{35} = 0.86
\]

So based on Table 6, the content validity of the test instruments developed is in the very good category. In other words, the developed high school mechanics knowledge test items are very good at measuring higher order thinking skills.

Based on the calculation results, it can be seen that of the 35 test items, there are 28 items that fall within the valid criteria. Furthermore, in Table 4.8, it can be seen that for the indicators to analyze there are 8 valid items; evaluating indicators, there are 12 valid items; and the creating indicator there are 8 valid items.

| Table 6. Content Validity |
|---------------------------|
| Expert I                  |
| (1-2)                     |
| (3-4)                     |
| Pakar II                  |
| (1-2)                     |
| 2, 6, 24                  |
| (3-4)                     |
| 1, 3, 4, 5, 7, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35 |
| 8, 17                     |

The result of the calculation of the test item reliability coefficient for valid items is 0.593 and is in the sufficient category. The obtained reliability coefficient indicates that the high-school mechanical material knowledge test instrument to measure high-order thinking skills developed has been consistent and can be trusted to be used.

Based on the results of the student response questionnaire to the HOTS questions given at the trial stage, it was found that: (1) the questions given were relevant to everyday life; (2) interesting and fun questions; (3) students are interested and serious in solving all the questions given; (4) students feel they have to apply previous knowledge to solve the questions given; (5) students feel happy and challenged to do the questions given; (6) students want questions like this to be given in class.

Some considerations for obtaining content and construct validity include the level of relevance of basic competencies, mechanics materials that have been studied by students, HOTS indicators, and the equivalence of answer choices in item questions. These four main points are then taken into consideration by experts in validating the theoretical instrument.

Next will be explained the consideration of the general characteristics of measuring higher order thinking skills for each indicator. Assessing the quality of students' thinking in breaking information down into parts and reasoning with that information requires questions that require students to find or describe these parts and find out how they are related [10,11]. Analysis-level questions that have been
developed involve giving students mechanics material (or asking them to find mechanics material) to students, then presenting questions or presenting problems whose answers require differentiating or organizing pieces of information in a sensible way.

Assessing evaluations requires items that can assess how students assess material as well as methods for their intended purpose [9,13,16,19]. Students can assess material based on criteria. These criteria can be standards or criteria that students find themselves (in this case the element of creativity is also involved). The developed evaluation item items do not prioritize personal preferences, but rather reasoned evaluations that can be expressed as arguments or conclusions supported by evidence and logic. To assess how well students do the evaluation, items are made by providing material and asking students to rate their "value" against a goal.

Judging whether students can “create” in a taxonomic sense means assessing whether they can combine different things in new ways or rearrange existing things to make something new [10,11]. This is done by giving students problem items or problems to solve that include many solutions, planning procedures to achieve certain goals, or producing something new. These criteria has also been met by various HOTS question developments carried out by several researchers [5,20–24].

In general, the instrument for measuring high-order high-level thinking skills developed is empirically and theoretically valid and has high reliability. So that the results obtained from measurements using this instrument can be trusted.

4. Conclusion
The results of the development of a test instrument for knowledge of high school mechanics materials that have been developed in measuring higher order thinking skills consist of 28 valid and reliable items so that they are feasible to use. The test items developed have been able to measure higher order thinking skills and the results are reliable.

Acknowledgments
The author would like to say thank to the undergraduate students and lecturer team in our research classroom at one of High School in Makassar (SMAN 8 Makassar), who have given us permission and convenience during the research.

References
[1] OECD 2016 Result from PISA 2015 (Indonesia) (Paris: OECD publishing)
[2] Sani R A 2013 Pembelajaran Saintifik untuk Implementasi Kurikulum 2013 (Jakarta)
[3] Toharuddin U, Hendrawati S and Rustaman A 2011 Membangun Literasi Sains Peserta Didik (Humaniora)
[4] Arafah K, Rusyadi R, Arafah B, Arafah A and Arafah B 2020 The Effect of Guided Inquiry Model and Learning Motivation on the Understanding of Physics Concepts Talent Development and Excellence 12 4271–83
[5] Abosalem Y 2016 Assessment Techniques and Students’ Higher-Order Thinking Skills International Journal of Secondary Education 4 1
[6] Arifin Z 2009 Evaluasi pembelajaran: prinsip, teknik, prosedur : standar penilaian menurut BSNP, model evaluasi, instrumen evaluasi, penilaian berbasis kelas, penilaian portofolio, analisis kualitas tes refleksi pelaksanaan evaluasi (Remaja Rosdakarya)
[7] Anon Evaluasi proses dan penilaian hasil belajar | Perpustakaan Kementerian Pendidikan dan Kebudayaan
[8] Brookhart S M 2010 How to Assess Higher-order Thinking Skills in Your Classroom (ASCD)
[9] Trisnawaty W, Citrasukmawati A and Thohir M A 2017 Self Assessment For Student Performance Based on Higher Order Thinking Skills in Physics Learning Journal of Education and Learning (EduLearn) 11 446
[10] Krathwohl D R 2002 A Revision of Bloom’s Taxonomy: An Overview Theory Into Practice 41 212–8
[11] Bloom B S, Airasian P W, Cruikshank K A, Mayer R E, Pintrich P R, Raths J and Wittrock M C 2001 A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives (Longman)
[12] Ramos J L, Ramos S, Dolipas B and Villamor B 2019 Higher Order Thinking Skills and Academic Performance in Physics of College Students: A Regression Analysis 2013
[13] Schraw G and Robinson D H 2011 Assessment of Higher Order Thinking Skills (IAP)
[14] Supeno, Astutik S, Bektiarso S, Lesmono A D and Nuraini L 2019 What can students show about higher order thinking skills in physics learning? IOP Conf. Ser.: Earth Environ. Sci. 243 012127
[15] Richland L E and Simms N 2015 Analogy, higher order thinking, and education WIREs Cognitive Science 6 177–92
[16] Ramos J L S, Dolipas B B and Villamor B B 2013 Higher Order Thinking Skills and Academic Performance in Physics of College Students: A Regression Analysis 13
[17] Anon Teknik penyusunan Instrumen tes dan nontes / Djemari Mardapi, Prof. Ph.D. ; editor, Ari Setiawan , M.Pd. | OPAC Perpustakaan Nasional RI.
[18] Anon Dasar Pnulisn Karya Ilmiah (Ed.4) - Zaenal Arifin - Google Buku
[19] Trisnawaty W Analyze of Student’s Higher Order Thinking Skills to Solve Physics Problem on Hooke’s Law 6
[20] Kusuma M D, Rosidin U and Suyatna A The Development of Higher Order Thinking Skill (Hots) Instrument Assessment In Physics Study 7
[21] Merta Dhewa K, Rosidin U, Abdurrahman A and Suyatna A 2017 The Development of Higher Order Thinking Skill (Hots) Instrument Assessment In Physics Study IOSR Journal of Research & Method in Education (IOSR-JRME) 7 26–32
[22] Nirwana N, Rochman S and Zukmadini A Y 2019 An assessment of Higher Order Thinking Skills (HOTS) Based on Rasch Models of Student in Physics Learning International Conference on Educational Sciences and Teacher Profession (ICETeP 2018) (Atlantis Press) pp 162–5
[23] Serevina V, Sari Y P and Maynastiti D 2019 Developing high order thinking skills (HOTS) assessment instrument for fluid static at senior high school J. Phys.: Conf. Ser. 1185 012034
[24] Tanujaya B, Mumu J and Margono G 2017 The Relationship between Higher Order Thinking Skills and Academic Performance of Student in Mathematics Instruction IES 10 78