The Developing and Calibration of PhysEDiTHOTS Based on IRT and IQF for Students’ HOTS Diagnostic

Edi Istiyono* and Suyoso
Yogyakarta State University, Indonesia

*E-mail: edi_istiyono@uny.ac.id

Abstract. This research aims at developing a diagnostic test based on Item Response Theory (IRT) to measure the weakness of physics students’ ability in higher order thinking skills (PhysEDiTHOTS) and to find out the characteristics of PhysEDiTHOTS. The instrument is essay test developed from the sub-aspects of higher order thinking skills and sub-materials of fundamental physics based on Indonesian Qualification Framework (IQF). The instrument comprises 24 valid items according to physics experts, physics education experts, and the experts of the measurement of physics education, which is analyzed using Aiken’s V (0.92 to 1.00). The instrument was tested on 271 students of Physics Education Department, The Faculty of Mathematics and Natural Sciences, Yogyakarta State University. The students were chosen using proportional random sampling techniques considering students level of study, i.e. first, third and fifth semester. The response is polytomous data in four categories that were analyzed according to Partial Credit Model (PCM). The research result shows that the 24 items of PhysEDiTHOTS fit to PCM with a reliability coefficient of 0.77, a difficulty index ranging between -0.97 and 1.87, as well as based on the information function and SEM, convey that PhysEDiTHOTS can be appropriately used to measure weakness in higher order thinking skills among students whose range of ability from -1.6 to 3.0. Therefore, PhysEDiTHOTS is an instrument that is good and appropriate to measure the weakness of physics students’ higher order thinking skills.

Keywords: Physics diagnostic test development; HOTS; IRT; IQF; PCM.

1. Introduction
Learning quality contributes to the quality of education. Therefore, the improvement of education quality can be done to improve learning quality. The efforts to improve learning quality can be initiated by setting up learning objectives that fit the graduate competence standards. The competence standards for high school physics are (1) living with positive attitudes with an ability to think critically, innovatively, and collaboratively, provided with honesty and openness, based on process potentials and physics products, as well as (2) recognizing and appreciating the roles of Physics in solving problems faced by mankind [1]. Therefore, the students are expected to be able to develop their thinking skills. The students are expected not only to develop their lower order thinking skills (LOTS), but also to develop their higher order thinking skills (HOTS). This is due to the significant relation between HOTS and physics students’ performance [2]. Thus, HOTS is required in the success the students to become perfect and successful humanbeings.

Concerning the higher order thinking skills, the facts show that the Indonesian physics students’ achievement is at the rank of 40 out of 42 countries [3]. A similar result is stated by Ridwan Efendi [4]
based on TIMSS. He said that (1) the average achievement of Indonesian physics students within the cognitive domain (knowledge, application, logics) is still low, and the Indonesian students’ ability is in the average category (knowledge and logic); (2) there is a tendency that Indonesian physics students’ achievement is degrading in every cognitive aspect. The knowledge of Physics among Indonesian students must be improved in all aspects, especially in the aspect of logic, through the training process of higher order critical thinking skills. This is in accordance with a statement saying that the higher order thinking skills among high school physics student in Yogyakarta is categorized as medium and low, and the distribution is very low (1.91%), low (16.03%), medium (61.11%), high (20.75%), and very high (0.19%) [5]. Therefore, the high school physics students learning result concerning higher order thinking skills can be considered low at the national as well as international level. This result is due to an appropriate learning process or inappropriate evaluation model.

The higher order thinking skills among the students taking physics subject in high school is related to the applied teaching model. Madhuri, Kantamreddi & Goteti [6] state the dominant teaching model is determined by the teacher. If the teacher does not apply the model that does not require the higher order thinking skills, the students skills will not develop.

In accordance with the explanation, ideally, the teacher’s higher order thinking skills need to be investigated. Considering that the benefit is to improve physics teaching, the research is focused on the students of physics education, who will be physics teachers in the future. Saido [7] says if the students are accustomed to experiencing learning using higher order thinking skills (HOTS), they will also get used to applying it when they teach later. Thus, it is necessary to develop a physics diagnostic test for students (PhysEDiTHOTS) to test the weakness of their higher order thinking skills. Based on the test result, a teaching model will be developed to improve their higher order thinking skills. This paper is focused on the discussion of higher order thinking skills diagnostic test for students of physics education.

According to Piaget’s theory on development, the formal operational stage begins at the age of eleven. At this stage, children begin to develop the ability to manipulate abstract concepts through propositions and hypothesis [8][9]. College students are generally aged above 18 years old. Therefore, the ability for higher order thinking has already established.

Bloom’s taxonomy has been applied in education, especially in the curriculum and teaching materials [10][11]. Bloom’s revised taxonomy divides the cognitive aspects into lower order thinking skills (LOTS) and higher order thinking skills (HOTS). LOTS includes the ability to remember, to understand, and to apply, while HOTS covers the ability to analyze, to evaluate, and to create [12]. Thus, HOTS in physics teaching includes the ability to analyze, to evaluate and to create.

Brookhart [10] defines higher order thinking skills (HOTS) as follows: (1) higher order thinking is an aspect at the higher stage of Bloom’s cognitive taxonomy, (2) the objective of teaching the taxonomy in the cognitive domain is to teach the students how to transfer knowledge, the ability to think means that the students are able to apply the knowledge and the skills they developed during their learning to new contexts. In this sense, the term “new” is the application of the concept they have never learned before. It does not mean that something new is not something that is “totally new” or something that has never been discovered before. Higher order thinking means that the students have the ability to related the knowledge they have learned earlier and something else they have never learned before.

Nitko & Brookhart [13] explain that the basic principle of evaluating the higher order thinking skills is the ability to combine knowledge and skills on new problems. New phenomena and artificial products are required to access the higher order thinking skills. Boud & Falchikov [14] state that one of the ways is using a set of instruments relevant to the theme or learning context so that the students’ weakness can be tested using a diagnostic test. The diagnostic test can depict the differences between the students’ expectation and their actual ability. This fact can be used to identify the specific problems experienced by the students [15]. Thus, PhysEDiTHOTS are expected to detect students’ weakness in their higher order thinking skills in Physics.

Globalization has resulted in changes in the overall life of society, not least the education and employment sectors. Students and labors mobility between countries presents challenges for universities
to gain recognition from the global community for their educational outcomes. This has led to GATS and AFTA countries including Indonesia compiling a national qualification framework. Qualification framework is an instrument to determine the level of qualification based on description of learning outcome (LO). LO is a statement about what is known, understood and can be done by someone after completing the learning process. The Indonesian Qualification Framework (IQF) is a competency-qualification framework that can match, equate and integrate between the field of education and the field of vocational training and work experience [16]. Thus based on this IQF can be done to rank of labors based on the level of qualification.

The IQF consists of nine levels qualification starting from level 1 to level 9 as the highest level. Graduates of Diploma 4 or Bachelor Degree equivalent to level 6. Knowledge is a mastery of certain systematic concepts, theories, methods, and philosophy gained through reasoning in the learning process, student work experience, research and community service. Graduation level knowledge of 6 undergraduate programs is mastering the theoretical concepts of specific knowledge and skills generally and theoretical concepts of special sections in the field of knowledge and skills in depth. [16]. Students majoring in physics education are students of undergraduate level meaning at the level of qualification 6, are required to master theoretical concepts and skills in physics and theoretical concepts and skills more specifically according to the concentration of expertise in depth.

There are several test types, and among others are multiple choices and essay. Multiple choice tests are still most widely used, and at the same time criticized, assessment and evaluation techniques all around the world [17]. The weakness of the multiple choices test is (1) There is a big chance that correct answers were made through guessing, and the students thinking ability cannot be clearly observed [18]. On the other hand, an essay test can measure, among others, critical thinking ability, originality, and the ability to organize and to synthesize [19]. An essay test can measure lower order up to higher order thinking ability [20][21]. Essay test can elicit students’ higher level cognitive abilities and are beneficial to evaluate complex concepts or skills [22]. Therefore, the advantages of having an essay test are (1) an evaluating instrument urges the students to memorize, to comprehend, and to organize ideas or things they have learned earlier; (2) the students are able to express their ideas in their own words in form of essay; and (3) this instrument can evaluate various competences, for examples the students’ ability to express their opinion, to think critically, and to draw conclusion. Therefore, an essay test is a better alternative.

Grading is done based on the steps done by the test takers. The highest score is given when all the assignments are perfectly done, while the lowest score is given to those who completed only the initial step. The basic assumption of PCM model is to give higher scores to the students with higher ability compared to those with lower ability [23]. Wright & Masters also states that PCM can appropriately analyze responses that measure the skills of critical thinking and conceptual understanding of science [24].

Based on the explanation above, an essay test is a kind of test that is appropriate to measure physics students’ weakness in higher order thinking skills. The effort to detect the students’ weakness is done by using an instrument called Physics Diagnostic Test for Higher Order Thinking Skills (PhysEDiTHOTS). Therefore, it is indeed necessary to develop the instrument. In accordance with that, this research aims at (1) developing PhysEDiTHOTS instrument that detects physics students’ weakness in higher order thinking skills, (2) obtaining the characteristics of PhysEDiTHOTS.

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 is a development research. The instrument development is based on the modified Wilson and Antonio-Oriondo model [25].
The development of test instruments based on modified Wilson and Antonio-Oriondo’s instrument development model is presented in Figure 1. The steps are (1) Determining test objectives, (2) Determining competences tested, (3) Determining test materials, (4) Designing the test blue-print, (5) Writing test items based on the principles of HOTS diagnostic test development, (6) Making scoring guidelines, (7) Content validation, (8) Revising and assembling test items for try-outs, (9) Determining try-out subjects, (10) Try-out implementation, (11) Analyzing try-out result data, and (12) Assembling PhysEDiTHOTS items.

According to experts, IRT analysis requires 200 up to 1000 participants for the samples [26]. Reckase [27] concluded that the samples are the minimum number needed to measure 3 logistic parameter. They are level of difficulty, distinguishing force and possibility of guessing among at least 300 participants [28]. Thus, by involving 271 students as the try-out subjects, this research has fulfilled the requirements to find out the characteristics of the test according to PCM. The students were chosen using proportional random sampling techniques considering students level of study, i.e. first, third and fifth semester.

3. Result and Discussion
This section presents the obtained results and following by discussion

3.1. Test Development Result
PhysEDiTHOTS is essay test developed as a sub-aspects of HOTS and sub-materials of Fundamental Physics based on IQF. HOTS sub-aspects include analyzing, evaluating, and creating. The sub-materials of Fundamental Physics are mechanics, heat, vibration and wave, optic, electricity, and magnet. The
distribution of the test comprising 24 items is presented in Table 1.

Table 1. The distribution of PhysEDiTHOTS items for Fundamental Physics sub-materials.

| Aspect     | Sub-aspect | Mechanics | Heat | Oscillation and Waves | Optics | Electrics | Magnet |
|------------|------------|-----------|------|-----------------------|--------|-----------|--------|
| Analyzing  | Differentiating | 1         | 9    |                       | 13     | 17        | 21     |
|            | Organizing  |           |      |                       |        |           |        |
|            | Attributing |           |      |                       |        |           |        |
| Evaluating | Checking   |           |      |                       | 14     | 23        |        |
|            | Critiquing |           |      |                       |        |           |        |
| Creating   | Generating |           |      |                       | 15     | 24        |        |
|            | Planning   |           |      |                       |        |           |        |
|            | Producing  |           |      |                       | 16     | 20        |        |

3.2. Instrument Validity

The validation of PhysEDiTHOTS contents has been done before the try-out phase according to the expert judgement analyzed using Aiken’s V. V is from is form 0.92 to 1.00. The try-out was done on 271 students of Physics Education, the Faculty of Mathematics and Natural Sciences, Yogyakarta State University. The analysis of the try-out result shows that the empirical validity of PhysEDiTHOTS for each item and for the whole test can be proven.

![Figure 2. Infit MNSQ of items](image_url)

The validity test was done in accordance with Adam Kho’s design [29]. If the Infit MNSQ average and the standard deviation is between 1 and 0.0, it can be concluded that the test fits the PCM. The Infit MNSQ average is 1.01 (about 1) and the standard deviation is 0.1 (about 0.10). Therefore, the instrument generally fits the PCM 1 PL. The item validity and the test takers are valid when it fit the range of Infit MNSQ between 0.77 and 1.30. The Infit MNSQ test is between 0.82 and 1.19 presented in Figure 2. Therefore, the 24 items fit the PCM.

3.3. Difficulty index

The difficulties index of the items is between -0.97 and 1.87 with 0 average and 0.35 standard deviation.
Thus, based on the level of difficulties (-2.0 < b < 2.0), the 24 items are categorized as good items. Figure 3 shows the level of difficulties in each aspect of analyzing, evaluating, and creating.

![Figure 3. Difficulty index of items in each aspect and sub-aspect of the instrument](image)

3.4. Reliability, Information Function, and SEM

The analysis shows that the reliability of the test instrument is about 0.77, a reliability coefficient that fulfills the requirement of a good instrument. From the data analysis, the information function and the standard error of measurement (SEM) are also obtained. The information function and the SEM are presented in Figure 4. This figure shows that PhysEDiTHOTS is appropriate for the students whose ability ($\theta$) is high, between $-1.6 \leq \theta < 3.0$. This is in accordance with the objectives of the instrument development to measure and detect the weakness of higher order thinking skills in Physics.

![Figure 4. Information Function and SEM](image)

3.5. Discussion

The reliability of the PhysEDiTHOTS is within good category and the reliability coefficient is 0.75, according to Mehrens & Lehman [30], who state that an instrument has a good reliability if the
coefficient is higher than 0.65. Besides, based on the information function and the SEM, this test is suitable for students whose ability is high (between -1.60 and 3.0). This means that the instrument has a high reliability based on the information function [31]. Based on the reliability coefficient, the information function of the test, and the estimation parameter, PhysDiHOTS is reliable and stable.

The validity is proven based on the expert judgment that is analyzed using Aiken’s V, while the empirical validity is tested based on PCM. The Aiken’s V of items’ PhysDiTHOTS is form 0.92 to 1.00. The average score and the standard deviation of the Infit MNSQ is 1.00 and 0.10. Thus, it can be concluded that the instrument developed is valid and has fit the PCM. The distribution of Infit MNSQ items is between 0.82 and 1.19. It is within the acceptance boundary, which is between 0.7 and -1.30.

Hambleton & Swaminathan [30] state that the difficulty index of the items is good if it is within the range of -2.00 and 2.00. Items with a difficulty index of 2.00 means that the particular item is very difficult. Therefore, based on the difficulty index of the items in the instrument (between -0.97 and 1.87), it is concluded that PhysEDiTHOTS can be categorized into good items.

The success is the result of the followings, among others are (1) the items which were developed according to the procedure of correct instrument development, and (2) the items which were develop based on the indicator of the aspect of higher order thinking skills for fundamental physics based on IQF, (3) the validity of test content which comprises 24 items and has been proven by the expert judgement, and (4) the implementation of the testing under the supervision of the teacher.

4. Conclusions and Suggestions
This section concludes and gives suggestion.

4.1. Conclusions
Based on the analysis of the data, it can be concluded that
a. PhysEDiTHOTS is developed for the aspects of analyzing, evaluating, and creating, and the sub-materials of Fundamental Physics based on IQF: mechanics, heat, vibration and wave, optic, electricity, and magnet. The form of PhysEDiTHOTS is an essay test consisting 24 items.
b. The characteristics of PhysEDiTHOTS are as follows.
   1) PhysEDiTHOTS is valid based on content validity using Aiken’s V by the expert judgment and empirically, with the criteria that fit PCM.
   2) The difficulty index of all the items in the PhysEDiTHOTS is in the good category with a value between -0.97 and 1.97.
   3) PhysEDiTHOTS is reliable and is based on the information function and the SEM. PhysEDiTHOTS is very appropriate to measure weakness in higher order thinking skills among physics students whose range of ability is from -1.6 to 3.0.

4.2. Suggestions
Based on the analysis of the data, it can be suggested that
a. Teachers are expected to implement PhysEDiTHOTS to diagnose the students’ weakness of higher order thinking skills in Physics for the subject taught. Furthermore, based on the student's weakness can be applied learning model that can improve their HOTS.
b. Further research is expected to conduct analysis testing based on generalized partial credit model (GPCM 2PL).

Acknowledgments
A gratitude is extended to the Directorate General of Higher Education, the Ministry of Research, Technology, and Higher Education for the research grant.
References

[1] Ministry of Education and Culture, (2016). Silabus mata pelajaran sekolah menengah atas/ madrasah aliyyah (SMA/MA) Mata Pelajaran Fisika. Jakarta: Ministry of Education and Culture

[2] Ramos, D.L.S., Dolipas, B. B., & Villamor, B.B. (2013). Higher order thinking skills and academic performance in physics of college students: A regression analysis. International Journal of Innovative Interdisciplinary Research, 48-60.

[3] TIMSS & PIRLS International Study Center. (2012). TIMSS 2011 international results in science. Boston: The TIMSS & PIRLS International Study Center, Boston College. It was accessed 5 January 2013, from http: timss.bc.edu/timss2011/release.html.

[4] Efendi, R. (2010). Kemampuan Fisika Peserta didik Indonesia dalam TIMSS. Prosiding Seminar Nasional Fisika ISBN : 978-979-98010-6-7.

[5] Istiyono, E. (2014). Pengukuran Kemampuan Berpikir Tingkat Tinggi Fisika Peserta Didik SMA di DIY, unpublished. Yogyakarta: Yogyakarta State University.

[6] Madhuri G.V., Kantamreddi, S.S.N., & Goteti L.N.S.P. (2013). Promoting Higher Order Thinking Skills Using Inquiry-Based Learning. European Journal of Engineering Education, 37 (2), 117-123.

[7] Saido, G.M., et al. (2015). Higher Order Thinking Skills Among Secondary School Students in Science Learning. The Malaysian Online Journal of Educational Science, 3(3), 13-20.

[8] Piaget, J. (2005). Benda Kongres. Jakarta: PT Gramedia.

[9] Reedeal, K.E. (2010). Jean Piaget’s Cognitive Development Theory in Mathematics Education. Department of Mathematics and Computer Science–Ripon College, 16-20. http://ripon.edu/mac/summation.

[10] Brookhart, S. M. (2010). How to Assess Higher Order Thinking Skills in Your Classroom. Alexandria: ASCD

[11] Schraw, G., & Robinson, D.H. (2011). Assessment of Higher Order Thinking Skills. New York: Information Age Publishing, Inc

[12] Anderson, L.W., & Krathwohl, D.R. (2001). A Taxonomy of Learning, Teaching, and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives. New York: Longman.

[13] Nitko, A.J & Brookhart, S. M. (2011). Educational assessment of students. (6th ed). Boston: Pearson Education, Inc.

[14] Boud, D., & Falchikov, N. (2007). Aligning assessment with long-term learning. Assessment & Evaluation in Higher Education, 31 (4), 399-413.

[15] Weeden, P., Winter, J., & Broadfoot, P. (2002). Assessment: What in it for School? London: Routledge Fahmer.

[16] Directorate General of Higher Education (2014). Panduan Penyusunan Capaian Pembelajaran Lulusan Program Studi. Jakarta: Ministry of Education and Culture

[17] Bahar, M., Aydin, F., & Karakirik, E. (2009). A Diagnostic Study Of Computer Application of Structural Communication Grid. The Turkish Online Journal of Educational Technology, 8(2), 5-19.

[18] Sudjana, N. (1990). Penilaian Hasil Belajar Mengajar. Bandung: PT Remaja Rosdakarya

[19] Ebel, R.L., (1979). Essentials of educational measurement (3rd ed). London: Prentice-Hall International, Inc

[20] Mardapi, D. (2012). Pengukuran, penilaian, dan evaluasi pendidikan. Yogyakarta: Nuha Litera.

[21] Worthen, B.R., Borg, W.R & White., K. R. (1993). Measurement and Evaluation in the Schools. New York: Longman.

[22] Yang, C.W., Kuo, B.C., & Liao, C.H. (2011). A HO-IRT Based Diagnostic Assessment System With Constructed Response Items. The Turkish Online Journal of Educational Technology, 10(4), 46-51.

[23] Bloom, B.S., et al. (1979). Taxonomy of Educational Objectives: Handbook 1 Cognitive Domain. London:Longmans Group Ltd.

[24] Van der Linden, W. J., & Hambleton, R. K. (1997). Handbook of Modern Item Response Theory. New York: Springer-Verlag New York, Inc

[25] Oriondo, L.L., & Dallo-Antonio, E.M. (1998). Evaluation Educational Outcomes. Manila: Rex Printing Company, inc

[26] Seon, Hi Sin. (2009). How to tread omitted respons in Rasch model based equating. [Versi eletronik] Practical Assessment, Research & Evaluation, 14(1), 1-8. ISSN 1531-7714

[27] Reckase, M.D. (2009). Multidimensional Item Response Theory (Statistics for Social and Behavioral Sciences). London: Springer
[28] Haladyna, T.M. (2004). Developing and Validating Multiple Choice Test Items. New Jersey: Lawrence Erlbaum Associates, Inc.

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

[30] Mehrens, W.A. & Lehmann, I.J. (1991). Measurement and evaluation in education and psychology. (3rd Ed.) New York: Holt, Rinehart and Winston.

[31] Hambleton, R.K., & Swaminathan, H. (1985). Item Response Theory. Boston, MA: Kluwer Inc.