

BUKTI KORESPONDENSI

ARTIKEL JURNAL INTERNASIONAL BEREPUTASI

Judul Artikel : The Development of the HOTS Test of Physics Based on Modern Test Theory: Question Modeling through E-learning of Moodle LMS

Jurnal : International Journal of Instruction, Volume 14 Nomor 4 Tahun 2021

Penulis : Sri Wahyu Widyaningsih, Irfan Yusuf, Zuhdan Kun Prasetyo, Edi Istiyono

Korespondensi : Sri Wahyu Widyaningsih

Konfirmasi Penerimaan Submit Artikel (4 April 2020)

e-**iji**.net'ten Makale gönderildi External Inbox x



e-iji.net <info@e-iji.net>
to me, submit.iji

Sat, Apr 4, 2020, 11:10 AM

Submit Article

First name Sri Wahyu
Last name Widyaningsih
Title (Mrs/Ms/Dr etc) Mrs
Your email: s.widyaningsih@unipa.ac.id
Scope of the Article: educational technology
Article subject Measurement in Physics Learning
Add article (doc, docx) 20200404021109_Sri-Wahyu-Widyaningsih-2-E-**iji**.doc
Confirm (1) 1
Confirm (2) 1

IP: 36.78.212.215



Pemberitahuan Hasil Review oleh Reviewer (22 Agustus 2020)

Amendments Inbox



iji@ogu.edu.tr

to s.widyaningsih, i.yusuf, zuhdan, ed_istiyono, me

Sat, Aug 22, 2020, 1:38 AM



Dear author

You have amendments from reviewers. Could you please amend on attached file "Article 180420_for revision" and send back your revised article and the list of explanations of the revisions done via e-mail (iji@ogu.edu.tr) as an attached file as soon as possible?

Sincerely yours,

Editorial

International Journal of Instruction

International Journal of Instruction

<http://www.e-iji.net>

8 Attachments



Sri Wahyu Widyaningsih <s.widyaningsih@unipa.ac.id>

to iji, s.widyaningsih, i.yusuf, zuhdan, ed_istiyono

Sat, Aug 22, 2020, 1:57 AM



Noted, will do.

Pengiriman Hasil Review (30 November 2020)

Article revision



Sri Wahyu Widyaningsih <s.widyaningsih@unipa.ac.id>
to iji, bcc: Irfan, bcc: -, bcc: -

Mon, Nov 30, 2020, 7:59 PM



Dear

Editor of the International Journal of Instruction

In the following, we send improvements to our article based on feedback from reviewers. We attach articles and certificates that have been proofreading.

Best regards,

Authors

Sri Wahyu Widyaningsih, [et.al](#)

2 Attachments



Pemberitahuan Hasil Review Kedua (21 Desember 2020)

Minor Amendments Inbox x



iji@ogu.edu.tr

to s.widyaningsih, i.yusuf, zuhdan, edi_istiyono, me

Mon, Dec 21, 2020, 6:27 AM



Dear author

You have minor amendments from a reviewer. Could you please amend on attached file "Article 180420_revised_for revision" and send back your revised article and the list of explanations of the revisions done **via e-mail** (iji@ogu.edu.tr) as an attached file as soon as possible?

Sincerely yours,

Editorial

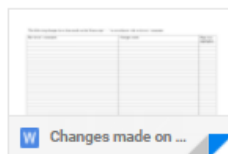
International Journal of Instruction

International Journal of Instruction

<http://www.e-iji.net>

-

4 Attachments



Sri Wahyu Widyaningsih <s.widyaningsih@unipa.ac.id>

to iji, s.widyaningsih, i.yusuf, zuhdan, edi_istiyono

Mon, Dec 21, 2020, 12:43 PM



Noted, will do.



Pengiriman Hasil Revisi Kedua (22 Desember 2020)



Sri Wahyu Widyaningsih <s.widyaningsih@unipa.ac.id>
to iji ▾

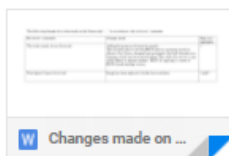
Tue, Dec 22, 2020, 8:41 AM ☆ ↶ ⋮

Dear
Editor of the International Journal of Instruction

In the following, we send the article corrections according to the reviewer's suggestions

Best regards,
Authors
Sri Wahyu Widyaningsih, [et al](#)

2 Attachments



iji@ogu.edu.tr
to s.widyaningsih, i.yusuf, zuhdan, edi_istiyono, me ▾

Thu, Dec 24, 2020, 6:39 AM ☆ ↶ ⋮

Dear author

We received your revised article and sent it to a reviewer. Thank you very much for your interest in **IJI**.

Sincerely yours.



Pemberitahuan Penerimaan Artikel (4 Maret 2021)

Re: Minor Amendments: Acceptance Article 180420 Inbox x



International Journal of Instruction <editor.eiji@gmail.com>
to me, s.widyaningsih, i.yusuf, zuhdan, -

Thu, Mar 4, 2021, 9:01 PM



Dear author

“ Article 180420 ”

This article has been completed the reviewing process and has been accepted for publication. Your manuscript is tentatively scheduled for publication in the October 2021 issue.

Please **contact us** only with this editor.eiji@gmail.com mail for your article from now on. **Others will not be considered.**

We wish you all the best.

Editorial

International Journal of Instruction

International Journal of Instruction

<http://www.e-ijj.net>

<http://www.gateacademy.ch>

From: "s widyaningsih" <s.widyaningsih@unipa.ac.id>

Sent: Thursday, March 4, 2021 1:11:06 PM

Subject: Re: Minor Amendments

Dear Editor

Pemberitahuan Penerbitan OnlineFirst Artikel (8 Juli 2021)



International Journal of Instruction <editor.eiji@gmail.com>

to s.widyaningsih, i.yusuf, zuhdan, -, me ▾

Jul 8, 2021, 10:01 PM



Dear author,

We are happy to announce that the **IJI** is now **Scopus Q1**.

We published your article as OnlineFirst. You can see your article on the web <http://www.e-iji.net/volumes/367-onlinefirst-2>

We will publish as home pages on October 01, 2021.

Note: No changes can be made to the article after 48 hours of publication.

Sincerely yours

International Journal of Instruction

International Journal of Instruction

<http://www.e-iji.net>

<http://www.gateacademy.ch>

We are happy to announce that the **IJI** is now **Scopus Q1**.

The Development of HOTS Test of Physics Based on the Modern Test Theory: Question Modeling through E-learning of Moodle LMS

Sri Wahyu Widyaningsih

Assist. Prof., Faculty of Teacher Training and Education, Universitas Papua, Indonesia, *s.widyaningsih@gmail.com*

Irfan Yusuf

Assist. Prof., Faculty of Teacher Training and Education, Universitas Papua, Indonesia, *i.yusuf@gmail.com*

Zuhdan Kun Prasetyo

Prof., Faculty of Science, Universitas Negeri Yogyakarta, Indonesia, *zuhdan@uny.ac.id*

Edi Istiyono

Prof., Graduate School, Universitas Negeri Yogyakarta, Indonesia, *edi_istiyono@uny.ac.id*

The present study discussed the development of higher-order thinking skills (HOTS) test of physics based on the modern test theory. HOTS questions were designed and presented in the e-learning with the Moodle learning management system (LMS) that could be accessed online. This study employed the ADDIE model with analysis, design, development, implementation, and evaluation stages. The instrument consisted of 24 multiple choice physics questions regarding the direct current circuit topic; the questions were designed by following the aspects and sub-aspects of HOTS and had been validated by the experts of measurement, physics education, physics, and practitioners. Moreover, validity analysis was based on the V Aiken formula, in which every aspect was confirmed valid. The validated instrument was then tried out to all 34 students at the Department of Physics Education, Universitas Papua, who participated in the basic physics subject. Dichotomy data analysis used the Rasch Model (RM) 1-PL through the Quest program, and the test characteristics comprised item fitness, reliability, and difficulty. The trial result obtained the criteria of INFIT MNSQ mean and standard deviation of 1.0 and 0.0, respectively, showing that the items fitted the RM1-PL. In addition, the value of item reliability based on the value summary of the item estimate arrived at 0.66; meanwhile, the case reliability under the summary of the case estimate accounted for 0.85. The reliability value in the range of 0.67- 0.80 was categorized as quite reliable. As based on the criteria of minimum and maximum INFIT MNSQ of 0.77 and 1.30, 24 question items fitted the RM 1-PL model. The result of the Quest output also revealed that the average values of Thresholds and its standard deviation were 0.00 ± 0.71 , or in the acceptance range of -2 to 2. All in all, all 24 question items that had

been tried out had fitted the model with a good category in order that they could be utilized in HOTS measurement.

Keywords: E-learning, HOTS Test, and Modern Test Theory.

INTRODUCTION

Assessment, particularly in the cognitive domain, is central to the learning process and should be carried out accurately and in compliance with the subject to be assessed or measured. Students' cognitive skills in the learning process can be categorized into lower-order thinking (LOT) and higher-order thinking (HOT). The LOTS include remembering, understanding, and applying; the HOTS, on the other hand, encompass analyzing, evaluating, and creating. HOTS are thinking skills that do not only require the remembering skill but also require other higher skills. Indicators to measure HOTS consist of analyzing (C4), evaluating (C5), and creating (C6) skills (Krathwohl & Anderson, 2010).

HOTS also refer to thinking skills when one takes new information, connects it with initial information s/he has, and finally delivers the information to achieve goals or answer questions (Istiyono, Dwandaru, & Muthmainah, 2019). This is in line with skill characteristics in the 21st century published by Partnership of 21st Century Skill stating that 21st century learners should be able to develop competitive skills, such as critical thinking, problem-solving, communication, information and communication technology (ICT) literacy, ICT, information literacy, and media literacy (Brun & Hinostroza, 2014); these focus on HOTS development.

Physics serves as part of science consisting of abstract concepts that are difficult to be directly described. Learning physics is expected to help students develop their thinking skills, in which they are not only demanded to master LOT skills, but also HOTS. Teachers are also urged to deliver learning materials to students, including the HOTS that can be improved by HOTS instrument. A previous study has reported that the majority of teachers find it challenging to develop an assessment instrument of learning outcomes, HOTS questions, in particular (Istiyono, 2018). For this reason, teacher creativity is highly required to measure students' learning outcomes. Today's development of Information and Communication Technology (ICT) can be utilized to design and habituate students to learn anywhere at any time (Yusuf, Widyaningsih, & Sebayang, 2018). Relying on ICT during the learning process is one of the significant innovations, including in the evaluation of students' learning outcomes.

The presentation of evaluation questions can be done in an integrated manner through e-learning programs, one of which is Moodle learning management system (LMS) (Azevedo, 2015; Bogdanović, Barać, Jovanić, Popović, & Radenković, 2014).

The Moodle provides different types of questions, such as multiple choices, true or false, and short answers; these are stored in the taught course database and can be re-used (Limongelli, Sciarrone, & Vaste, 2011). Teachers are also able to give feedback directly to the students and give them correct answers to questions they have worked on (Pandey & Pandey, 2009). One of the advantages of an online evaluation through Moodle LMS is that students can directly figure out their assessment results.

Teachers need to prepare a good test to measure students' learning outcomes. There are two paradigms developed for students' learning outcome assessment through the applied test, i.e., classical and modern approaches. The classical paradigm being utilized is classical test theory or widely known as classical true-score theory, meanwhile, the modern paradigm is item response theory (IRT). The classical test theory is selected due to its ease in the application despite of its limitations in measuring the item difficulty level and discrimination since the calculation of both indicators is based on the test taker's total score. In contrast, the IRT frees up the dependence between the test item and test taker (a concept of parameter invariance); the test taker's response to a test item does not affect another item (a concept of local independence), and; the test item does only measure one measurement dimension (unidimensional concept) (Raykov & Marcoulides, 2015). Therefore, the application answers the needs of modern measurement to date, i.e., a comparison between test taker's skills, question development, and even adaptive test development, so that it is considered able to overcome the classical test theory limitations.

This development study is an initial study with a long-term purpose of developing general physics questions with good quality at the Department of Physics Education, Universitas Papua. As the first stage, this study focuses on students at the department mentioned previously who enroll in General Physics subject taught by the researcher. This study also serves as one of the efforts to expand students' HOTS by applying a variety of HOTS-based learning sources.

METHOD

The ADDIE model, as employed by this study, refers to a general and systematic model of development study with a phased framework, allowing each element to connect with each other (Aldoobie, 2015). The stages of this model used in the development of HOTS instrument are presented in Figure 1.

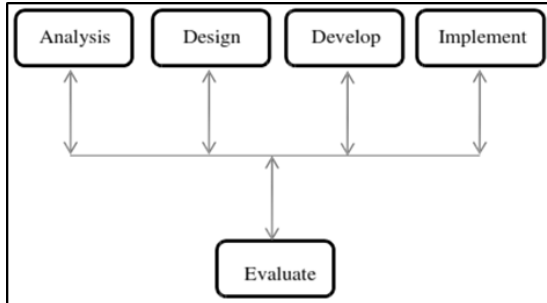


Figure 1
Stages of ADDIE Development Model in Designing Moodle LMS-based HOTS Test.

Analysis

The analysis stage is a process of needs analysis in the form of determining test objectives, identifying problems, analyzing tasks, and determining question formats to be applied. It is revealed that the problems are related to the needs of HOTS instrument design for students at the Department of Physics Education, Universitas Papua.

Design

This stage comprises the process of designing HOTS questions to be used; the design process encompasses creating a question matrix and outline that covers question distribution in every aspect and sub-aspect of HOTS.

Develop

Moreover, every single thing required in the arrangement of HOTS skill questions has been prepared in the next stage. This stage also comprises the process of making the questions regarding HOTS, as well as validating the questions that involve the experts of measurement, physics education, and practitioners. The technique of validity analysis to assess the content validity of the developed questions applies the V Aiken formula (Aiken, 1980, 1985).

$$V = \frac{\sum s}{n(c-1)} \quad (1)$$

“V” refers to the agreement index of validators in regards to item validity; “s” is the assessment score of validators subtracted by the assessment lowest score; “n” refers to the number of validators; “c” is the number of categories that can be chosen by validators. All test items are considered valid if the value of the V Aiken index falls into the range of 0.37 - 1 (Kowsalya, Venkat Lakshmi, & Suresh, 2012). The value of V Aiken of every test item is calculated based on the assessment items of every validator. In this stage, there is also an evaluation process, i.e., revising questions by following validators’ corrections and suggestions.

Implementation

Another stage is applying HOTS questions that have been developed to 34 students in the site area who enroll in general physics subject. This number has been following the sample size for data stability in Rasch Model (RM) 1- PL, which is from 30 to 300, with the limit of INFIT t is from -2 to +2 (Bond & Fox, 2007). Question item analysis is performed based on the raw score of the students by employing the Quest program.

Evaluation

Evaluation is a process of finding out whether or not the developed questions of HOTS have met the expectation. The evaluation stage is carried out in every stage and called a formative evaluation intended for revisions (Lee & Zainal, 2017). For instance, in the design stage, the expert's review is necessary to provide input towards the design. Further, the evaluation stage is undertaken after analyzing empirical questions mathematically by using the Quest software program by referring to the Rasch model. The Quest program is able to do the Rasch measurement, i.e., a comprehensive empirical test of question items. There are three parameters being measured mathematically based on the empirical test of question items.

1. The first parameter is item fitness with the Rasch model by following the value of INFIT MNSQ or INFIT t of the item. The expected values of the unweighted mean square (Outfit MNSQ) in the Quest program and weighted mean square are 1; the variance is 0. On the contrary, the expected value of Mean INFIT t is equal to 0, with the variance equal to 1 (Adams & Khoo, 1996). The provision of INFIT MNSQ for the Rasch Model is shown in Table 1 and Table 2.

Tabel 1

Criteria of Question Item Fitness with the Rasch Model

MNSQ INFIT Value	Criteria
>1,33	Does Not Fit the Rasch Model
0,77 s.d. 1,33	Fits the Rasch Model
<0,77	Does Not Fit the Rasch Model

Tabel 2

The Provision of Outfit t for the Rasch Model.

t OUTFIT Value	Criteria
OUTFIT t \leq 2,00	Fits the Rasch Model
OUTFIT t \geq 2,00	Does Not Fit the Rasch Model

2. The second parameter is reliability. The analysis result of the Quest program also reveals the item and case reliability. The reliability value based on the item estimate is also called as sample reliability; the higher the value, the more the items that fit the tested model. Whereas, the lower the value, the less the items that fit the tested

model, so that it does not give the expected information. The reliability category is provided in Table 3 (Istiyono, 2017).

Tabel 3

Interpretation of Reliability Value

Reliability Value	Criteria
> 0,94	Excellent
0,91 – 0,94	Very Good
0,81 – 0,90	Good
0,67 – 0,80	Acceptable
< 0,67	Poor

3. The third parameter is item difficulty index and respondents' skills presented as difficulty index in the Quest output. Thresholds (THRSHL) show the item difficulty index in the logit scale along with its standard deviation (Hambleton & Rogers, 1989). The provision of the THRSHL value is given in Table 4.

Tabel 4

Criteria of THRSHL Value to Categorize Item Difficulty Level

THRSHL Value	Criteria
$b > 2,00$	Very Difficult
$1,00 < b \leq 2,00$	Difficult
$-1,00 < b \leq 1,00$	Medium
$-1,00 > b \geq 2,00$	Easy
$b < -2,00$	Very Easy

Respondents' skills are shown by the value of the estimate error, in which the criteria of the estimate value of respondents' skills are presented in Table 5.

Tabel 5

Criteria of Estimate Value to Categorize Respondents' Skills

THRSHL Value	Criteria
$b > 2,00$	Very Difficult
$1,00 < b \leq 2,00$	Difficult
$-1,00 < b \leq 1,00$	Medium
$-1,00 > b \geq 2,00$	Easy
$b < -2,00$	Very Easy

The evaluation stage also includes the process of analyzing the HOTS of students on the whole. The level of HOTS is categorized based on the ideal mean and standard deviation. This is applied with the assumption that students' HOTS of physics are normally distributed. The ideal mean (I_m) and ideal standard deviation (I_{sd}) are based on the highest and lowest score of research variables. Table 6 shows the criteria of students' HOTS of physics.

Tabel 6
Criteria of Students' HOTS of Physics

Interval	Criteria
$Im + 1,5 Isb < \theta$	Very high
$Im + 0,5 Isb < \theta \leq Im + 1,5 Isb$	High
$Im - 0,5 Isb < \theta \leq Im + 0,5 Isb$	Medium
$Im - 1,5 Isb < \theta \leq Im - 0,5 Isb$	Low
$0 < Im - 1,5 Isb$	Very Low

Meaning:

Im : ideal mean

Isb : ideal standard deviation

X_{mak} : highest score

X_{min} : lowest score

RESULTS AND DISCUSSION

ADDIE development model can be used for different product developments in education, and one of which is the development of HOT skill questions. This model is simple and systematically structured in its implementation stages. The following is the description of each stage result.

Analysis

Needs analysis is the first stage being done by observation and interview to gather any information needed in the process of physics learning at the Department of Physics Education, Universitas Papua. The researcher's experience indicates that lecturers have applied HOTS learning in the classroom. However, a test to measure students' HOTS has not been conducted. The arrangement of HOTS instrument is required to train and develop students' HOTS. Accordingly, to facilitate the students in accessing other learning sources, this study designs HOT skill questions in an online system through an e-learning program using the Moodle LMS.

Design

In the design stage, the test instrument is designed based on the analysis result in the first stage. Test instrument design in this stage is in the form of question matrix and outline which are adjusted to students' needs and characteristics, and learning sources. The test is a multiple-choice test, in which 24 questions are adjusted to the formulation of a HOTS test that has been created in the test matrix and outline. The question matrix is provided in Table 7.

Tabel 7
The Question Matrix

Aspect	Sub Aspect	Theory		
		Electric current, Ohm's law, and electrical power	Series and parallel circuits of resistor and capacitor	Electric Force, Kirchoff's law, and RC circuit.
Analyze	Differentiating	8	12	21
	Organizing	3	15	20
	Attributing	2	9	23
Evaluate	Checking	4	11	22
	Critiquing	1	16	18
Create	Generating	5	13	19
	Planning	7	14	17
	Producing	6	10	24

Develop

The development of HOTS questions is based on the question matrix and outline that have been designed. Further, the questions are made online through e-learning by utilizing the Moodle LMS. Figure 2 shows all question items in the e-learning program.

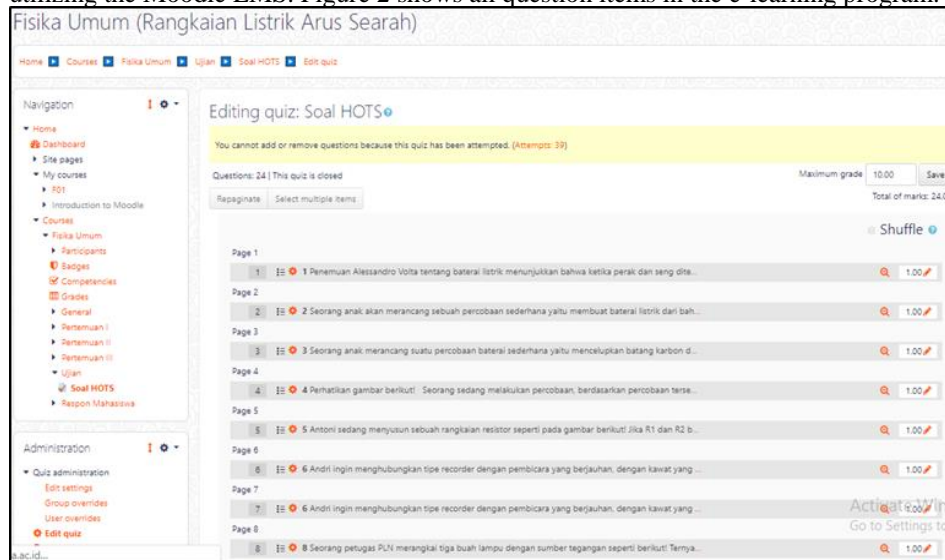


Figure 2
Shows All Question Items in the E-Learning Program

Moodle LMS program presents an interesting display and is easy to access by users (Martín-Blas & Serrano-Fernández, 2009). The questions are displayed interactively, and students can randomly work on the questions. Moodle LMS can present questions

with a picture or other content to make it easier for teachers to design the questions as expected. Figure 3 shows one of the HOTS questions displayed on the e-learning through the Moodle LMS.

The screenshot displays a Moodle LMS quiz interface. At the top, the course title is "Fisika Umum (Rangkaian Listrik Arus Searah)". Below the title, there are navigation links: Home, Courses, Fisika Umum, Ujian, Soal HOTS, and Preview. The main content area is divided into several sections:

- Quiz navigation:** A grid of question numbers from 1 to 24. Question 3 is highlighted in red, indicating it is the current question.
- Question 3:** The question text reads: "Seorang anak merancang suatu percobaan baterai sederhana yaitu mencelupkan batang karbon dan seng pada suatu larutan asam seperti gambar berikut:". Below the text is a diagram of a simple battery. It consists of a beaker labeled "Asam" containing two electrodes: "Elektroda Karbon" (Carbon Electrode) and "Elektroda Seng" (Zinc Electrode). The electrodes are connected to terminals labeled "Terminal" and "Terminal".
- Options:** Below the diagram, there are five multiple-choice options:
 - Semakin banyak ion karbon yang memasuki larutan sehingga beda potensial sulit dipertahankan
 - Larutan asam melarutkan ion karbon lebih banyak
 - Elektron-elektron terlepas dari elektroda seng sehingga elektroda seng bermuatan negatif dan akan habis
 - Semakin banyak ion seng yang terlarung menyebabkan elektroda atau yang lainnya terpolai habis menyebabkan sel mati
 - Elektron-elektron terlepas dari elektroda karbon menyebabkan elektroda seng bermuatan negatif dan lama kelamaan akan netral

Figure 3 Shows of the HOTS Questions Displayed on the E-learning Through the Moodle LMS

The development stage aims to produce a HOTS test instrument that has been validated by experts and practitioners. Product validation is a process of assessing the designed product, or in this case, the test instrument of HOTS in general physics subject in the site area. Product validation is carried out by involving seven validators, i.e., experts of measurement, physics education, physics, and practitioners. The validity test of the instrument includes material, construction, and language. The analysis result of question validity that is assessed by validators obtains the value of V Aiken in the range of 0.76 - 1.00, showing a valid result. The questions validated by experts and practitioners are then revised based on provided corrections and suggestions.

Implementation

The implementation stage in this study is the product trial, in which HOTS questions are tried out to 34 students in the research site. The students work on these questions via online through e-learning by using their own Moodle account upon the completion of all learning stages. Results of the students' learning can be accessed after this process.

Evaluation

Before conducting the estimate analysis of respondents' skills and item difficulty level, the analysis of item fitness is performed by using parameters of INFIT and OUTFIT for mean square and t. The determination of the item fitness with the model is based on the value of INFIT MNSQ and the standard deviation or Infit t (Adams & Khoo, 1996). The fitness of each case is also based on the value of INFIT MNSQ or INFIT t of the item. Table 8 provides the testing result through the Quest program to obtain the values of item estimate and case estimate in the HOTS questions trial.

Tabel 8

Values of Item Estimate and Case Estimate in the HOTS Questions Trial

No	Measurement	Estimates for Items	Estimates for Testi
1.	Average values and standard deviations	$0,00 \pm 0,57$	$0,01 \pm 1,24$
2.	Reliability Estimates	0,66	0,85
3.	The mean value and standard deviation of INFIT MNSQ	$1,00 \pm 0,14$	$0,99 \pm 0,15$
4.	The mean value and standard deviation of OUTFIT MNSQ	$1,09 \pm 0,52$	$1,09 \pm 0,52$
5.	The mean value and standard deviation of INFIT t	$-0,03 \pm 0,81$	$0,00 \pm 0,72$
6.	The mean value and standard deviation of OUTFIT t	$0,21 \pm 0,91$	$0,17 \pm 0,81$

The analysis result reveals that the INFIT MNSQ arrives at the range of 0.86 - 1.14, and INFIT t is -0.28 - 0.72. This result signifies that all 24 questions fit the model as they reach the range of INFIT MNSQ value from 0.77 to 1.30 and use INFIT t with the limit of -2.0 - 2.0 [16]. In addition to testing the fitness, the output of the Quest program also presents the reliability estimate of the test instrument. Table 8 provides the value of item reliability based on the value of summary of item estimate, which is 0.66. On the other hand, the value of person reliability, as based on the summary of case estimate, gets 0.85. These results are in line with the Rasch model, in which the reliability value falls under the range of 0.67 - 0.80 (quite reliable). On that ground, the instrument can be used to measure students' HOTS in the General Physics subject.

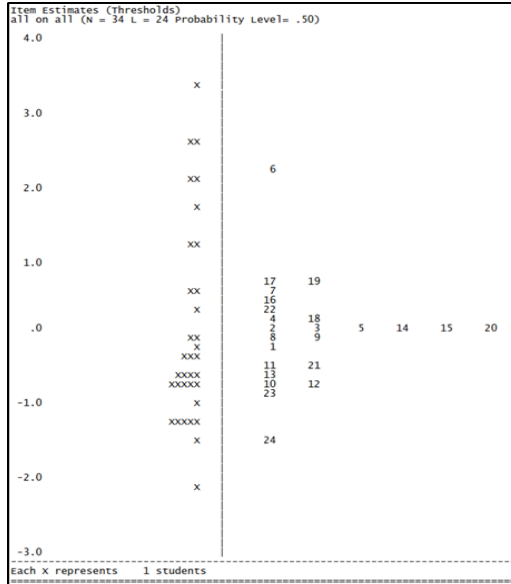


Figure 4
Distribution of Item Difficulty Level and Respondents' Skills

Figure 4 presents the distribution of the respondents according to the difficulty level in the logit scale from -4.0 to +4.0. This map displays the item difficulty level compared to the respondents' skills. Case and item difficulty levels in the Rasch model are expressed in one line in the form of abscissa in the graph with logg-odd unit. The graph of respondents' skills shows a normal curve, meaning that there are only a few respondents with low and high skills; and a lot of respondents with moderate skills. The level of item difficulty of threshold reveals that item 6 is the most difficult question, and item 24 is the easiest one.

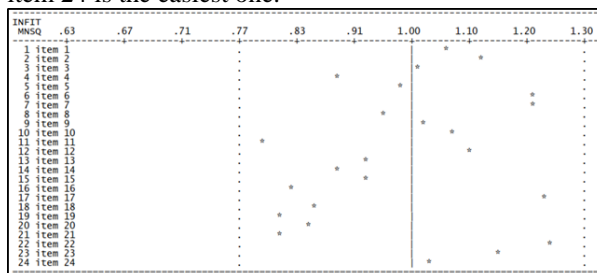


Figure 5
Distribution of INFIT MNSQ Values of Each Question Item of HOTS

Question items that fit the Rasch model are in the range of 0.77 - 1.33. Figure 5 shows that all 24 question items are in the line, implying that they fit the Rasch model.

Item Estimates (Thresholds) In input Order							
all on all (N = 34 L = 24 Probability Level= .50)							
ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1 item 1	18	34	-.26 .39	1.06	1.15	.4	.5
2 item 2	16	34	.04 .40	1.12	1.17	.7	.6
3 item 3	16	34	.04 .40	1.01	.91	.1	-.2
4 item 4	15	34	.19 .40	.88	.93	-.6	-.1
5 item 5	16	34	.04 .40	.98	.89	.0	-.2
6 item 6	5	34	2.27 .57	1.21	2.16	.7	1.4
7 item 7	13	34	.52 .42	1.21	1.27	1.0	.9
8 item 8	17	34	-.11 .40	.96	1.00	-.2	.1
9 item 9	17	34	-.11 .40	1.02	.91	.2	-.2
10 item 10	21	34	-.70 .39	1.07	1.16	.6	.5
11 item 11	19	34	-.41 .39	.79	.66	-1.6	-.9
12 item 12	21	34	-.70 .39	1.10	1.14	.8	.5
13 item 13	20	34	-.55 .39	.93	1.09	-.5	.4
14 item 14	16	34	.04 .40	.88	.78	-.7	-.6
15 item 15	16	34	.04 .40	.93	.82	-.4	-.5
16 item 16	13	33	.47 .42	.82	.69	-.8	-.9
17 item 17	12	34	.69 .43	1.23	1.16	1.0	.6
18 item 18	15	34	.19 .40	.86	.73	-.8	-.8
19 item 19	12	34	.69 .43	.81	.71	-.8	-.8
20 item 20	16	34	.04 .40	.85	.75	-.9	-.7
21 item 21	19	34	-.41 .39	.81	.68	-1.4	-.8
22 item 22	14	34	.35 .41	1.24	1.23	1.2	.8
23 item 23	22	34	-.85 .40	1.15	3.04	1.1	3.1
24 item 24	26	34	-1.50 .43	1.03	1.02	.2	.2
Mean			.00	1.00	1.09	.0	.1
SD			.71	.14	.52	.8	.9

Figure 6
Item Estimates from HOTS Questions

Figure 6 presents the Item Estimate of HOT skill questions based on the trial result. In this figure, there is SCORE-MAXSCR successively showing the score of the respondents who answer correctly, and the number of total respondents. Item 24 is the most correctly-answered, in which 26 out of 34 respondents are able to work on this item. Figure 6 also provides the value of THRSHL that shows the item difficulty index in the logit scale along with its standard deviation. Item 6 has THRSHL or difficulty

index of 2.27 that is greater than 2.0, or in other words, this item is very difficult since only five students can give a correct answer. The average value of THRSHL and its standard deviation accounts for 0.00 ± 0.71 and falls into the range of -2 - 2 (Hambleton & Rogers, 1989). The average value of INFIT MNSQ is 1.00 ± 0.14 and falls under the acceptance range of 0.77 - 1.33; the average value of OUTFIT t arrives at 0.10 ± 0.90 and falls into the acceptance range of ≤ 2.00 . All of these results indicate that all question items that have been developed can be employed to measure students' HOTS.

Case Estimates In Input Order								
all on all (N = 34 L = 24 Probability Level= .50)								
NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFIT MNSQ	OUTFIT MNSQ	INFIT t	OUTFIT t
1 01	12	24	-.02	.43	1.06	1.02	.58	.18
2 02	6	24	-1.21	.49	1.17	1.09	.75	.36
3 03	8	24	-.77	.45	.98	1.01	-.06	-.13
4 04	8	24	-.77	.45	.83	.81	-1.04	-.48
5 05	8	24	-.77	.45	.89	.83	-.59	-.41
6 06	6	24	-1.21	.49	.79	.70	-.84	-.67
7 07	10	24	-.38	.43	.99	.95	-.01	-.09
8 08	6	24	-1.21	.49	1.07	2.30	.36	2.44
9 09	3	24	-2.10	.63	.98	.85	.11	.00
10 10	9	24	-.57	.44	.88	.83	-.80	-.48
11 11	22	24	2.61	.77	.73	.46	-.30	-.56
12 12	5	24	-1.46	.52	.89	.85	-.29	-.18
13 13	20	24	1.75	.57	1.21	1.45	.64	.93
14 14	11	24	-.20	.43	.86	.83	-1.33	-.59
15 15	21	24	2.12	.64	1.18	1.05	.52	.29
16 16	9	24	-.57	.44	1.08	1.06	.59	.28
17 17	7	24	-.98	.47	1.29	2.20	1.38	2.55
18 18	6	24	-1.21	.49	1.23	1.28	.96	.75
19 19	14	24	.35	.43	.92	.87	-.56	-.40
20 20	15	24	.54	.44	.97	1.09	-.13	-.40
21 21	18	24	1.19	.49	.94	.86	-.16	-.25
22 22	21	24	2.12	.64	.93	1.23	-.01	.55
23 23	9	24	-.57	.44	1.07	1.01	.54	.15
24 24	8	24	-.77	.45	1.01	.95	.13	-.03
25 25	10	24	-.38	.43	.87	.82	-1.06	-.57
26 26	15	24	.54	.44	1.05	1.22	.36	.80
27 27	6	24	-1.21	.49	.82	.74	-.69	-.56
28 28	22	24	2.61	.77	.73	.46	-.30	-.56
29 29	12	24	-.02	.43	.92	.88	-.73	-.39
30 30	9	24	-.57	.44	.90	.90	-.64	-.23
31 31	23	24	3.40	1.05	1.18	3.14	.49	1.53
32 32	18	24	1.19	.49	.85	.75	-.53	-.58
33 33	10	23	-.32	.44	1.11	1.11	.97	.45
34 34	8	24	-.77	.45	1.29	1.34	1.61	1.03
Mean			.01		.99	1.09	.00	.17
SD			1.35		.15	.52	.72	.81

Figure 7
Case Estimates from Every Student

Figure 7 serves as the case estimate or the skill level of each student. Information obtained from the case estimate is that the SCORE-MAXSCR shows the score of each respondent from the maximum score sequentially. Respondent 31 answers the most questions (23 out of 24 questions) correctly compared to other respondents. The average estimate value and its standard deviation gets 0.01 ± 1.35 and falls under a moderate category. The analysis result of the case estimate reveals that students' skills are in the moderate category.

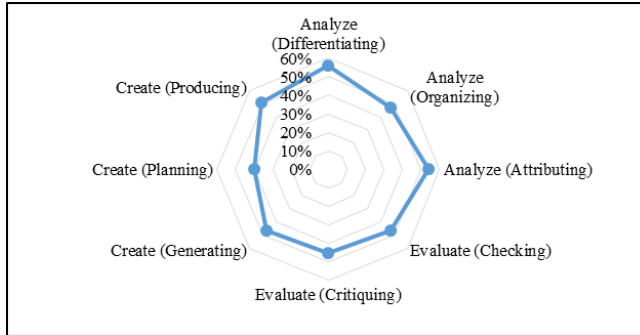


Figure 8
Distribution of Student Answer Percentage HOTS

Figure 8 gives the percentage of students' answers based on the aspects and sub-aspects of HOTS. The analysis result brings out the fact that students tend to find it difficult to answer questions regarding the creating aspect, especially the planning sub-aspect. Creating is the highest level HOTS in Bloom's taxonomy, which therefore, students need to practice developing their creating skills. This figure also signifies that the majority of the students find it easy to answer HOTS questions related to the analysis aspect, differentiating sub-aspect in particular.

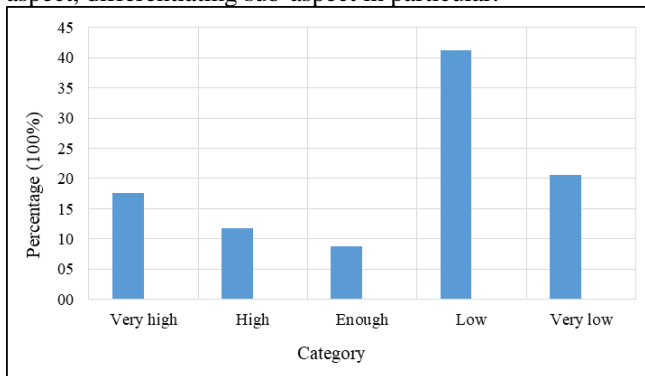


Figure 9
Percentage of Students' HOTS

Figure 9 shows the percentage of students' HOTS. It is seen that most students (41.2%) still have low HOTS; the categories consist of very low (20.6%), moderate (8.8%), high (11.8%), and very high (17.6%). The low category of students' HOTS is influenced by several factors, one of which is that the students are not used to working on HOTS questions (Tanujaya, Mumu, & Margono, 2017; Yusuf & Widyaningsih, 2019). They need to practice developing their HOTS by being exposed to HOTS-based learning sources.

CONCLUSION

Test characteristics comprised item fitness, reliability, and difficulty. Dichotomy data analysis used the Rasch Model through the Quest program. The trial result obtained the criteria of INFIT MNSQ mean and standard deviation of 1.0 and 0.0, respectively, showing that the items fit the RM1-PL. In addition, the value of item reliability based on the value of summary of item estimate arrives at 0.66; meanwhile, the person reliability under the summary of case estimate reaches 0.85, i.e., the reliability value is in the range of 0.67 - 0.80 (quite reliable). As based on the criteria of minimum and maximum INFIT MNSQ of 0.77 and 1.30, 24 question items fit the RM 1-PL model. The result of the Quest output also reveals that the average value of THRSHL and its standard deviation is $0.00 \pm 0,71$, or in the acceptance range of -2 to 2. To sum up, all 24 question items that had been tried out have fit the model with a good category, so that they can be utilized in HOTS measurement.

ACKNOWLEDGMENT

We would like to acknowledge the contribution of the Ministry of Research and Higher Education in funding this study through Inter Higher Education Institution Cooperation scheme with the contract number: 198/SP2H//AMD/LT/DRPM/2020.

REFERENCES

- Adams, R. J., & Khoo, S.-T. (1996). *Quest: the interactive test analysis system*. Camberwell, Vic.: Australian Council for Educational Research.
- Aiken, L. R. (1980). Content Validity and Reliability of Single Items or Questionnaires. *Educational and Psychological Measurement*, 40(4), 955–959.
- Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, 45(1), 131–142.
- Aldoobie, N. (2015). ADDIE Model. *American International Journal of Contemporary Research*, 5(6), 72.
- Azevedo, J. M. (2015). e-Assessment in mathematics courses with multiple-choice questions tests. *CSEDU 2015 - 7th International Conference on Computer Supported Education, Proceedings*, 2, 260–266. <https://doi.org/10.5220/0005452702600266>
- Bogdanović, Z., Barać, D., Jovanić, B., Popović, S., & Radenković, B. (2014). Evaluation of Mobile Assessment in A Learning Management System. *British Journal of Educational Technology*, 45(2), 231–244. <https://doi.org/10.1111/bjet.12015>
- Brun, M., & Hinostroza, J. E. (2014). Learning to become a teacher in the 21st century: ICT integration in Initial Teacher Education in Chile. *Journal of Educational*

- Technology & Society*, 17(3), 222–238. Retrieved from <http://www.jstor.org/stable/jeductechsoci.17.3.222>
- Hambleton, R. K., & Rogers, H. J. (1989). Detecting Potentially Biased Test Items: Comparison of IRT Area and Mantel-Haenszel Methods. *Applied Measurement in Education*, 2(4), 313–334. https://doi.org/10.1207/s15324818ame0204_4
- Istiyono, E. (2017). The Analysis of Senior High School Students' Physics HOTS in Bantul District Measured using PhysReMChoTHOTS. *AIP Conference Proceedings*, 1868(August), 1–7. <https://doi.org/10.1063/1.4995184>
- Istiyono, E. (2018). IT-based HOTS assessment on physics st learning as the 21 century demand at senior high schools: Expectation and reality IT-Based HOTS Assessment on Physics Learning as the 21 st Century Demand at Senior High Schools: Expectation and Reality. *AIP Conference Proceedings*, 2014(020014), 1–6.
- Istiyono, E., Dwandaru, W. S. B., & Muthmainah. (2019). Developing of Bloomian HOTS Physics Test: Content and Construct Validation of The PhysTeBloHOTS Developing of Bloomian HOTS Physics Test: Content and Construct Validation of The PhysTeBloHOTS. *Journal of Physics: Conference Series*, 1397(012017), 1–9. <https://doi.org/10.1088/1742-6596/1397/1/012017>
- Kowsalya, D. N., Venkat Lakshmi, H., & Suresh, K. P. (2012). Development and Validation of a Scale to assess Self-Concept in Mild Intellectually Disabled Children. *International Journal of Social Sciences & Education*, 2(4).
- Krathwohl, D. R., & Anderson, L. W. (2010). Merlin C. Wittrock and the Revision of Bloom's Taxonomy. *Educational Psychologist*, 45(1), 64–65. <https://doi.org/10.1080/00461520903433562>
- Lee, M. F., & Zainal, N. A. (2017). Development of needham model based E-module for electromagnetic field & wave. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 120–124). <https://doi.org/10.1109/IEEM.2017.8289863>
- Limongelli, C., Sciarrone, F., & Vaste, G. (2011). Personalized e-learning in Moodle: the Moodle_LS System. *Journal of E-Learning and Knowledge Society*, 7(1), 49–58. Retrieved from <https://www.learntechlib.org/p/43340>
- Martín-Blas, T., & Serrano-Fernández, A. (2009). The role of new technologies in the learning process: Moodle as a teaching tool in Physics. *Computers & Education*, 52(1), 35–44. <https://doi.org/10.1016/J.COMPEDU.2008.06.005>
- Pandey, S. R., & Pandey, S. (2009). Developing a More Effective and Flexible Learning Management System (LMS) for the Academic Institutions using Moodle. *ICAL 2009 - Technology, Policy and Innovation*, 249–254.
- Raykov, T., & Marcoulides, G. A. (2015). On the Relationship Between Classical Test Theory and Item Response Theory: From One to the Other and Back. *Educational*

Author surnames go here

17

and Psychological Measurement, 76(2), 325–338.
<https://doi.org/10.1177/0013164415576958>

Tanjaya, B., Mumu, J., & Margono, G. (2017). The Relationship between Higher Order Thinking Skills and Academic Performance of Student in Mathematics Instruction. *International Education Studies*, 10(11), 78.
<https://doi.org/10.5539/ies.v10n11p78>

Yusuf, I., & Widyaningsih, S. W. (2019). HOTS profile of physics education students in STEM-based classes using PhET media. *Journal of Physics: Conference Series*, 1157(032021), 1–5.

Yusuf, I., Widyaningsih, S. W., & Sebayang, S. R. B. (2018). Implementation of E-learning based-STEM on Quantum Physics Subject to Student HOTS Ability. *Turkish Science Education*, 15(December), 67–75.



International Journal of Instruction Article Evaluation Form

Mr. /Mrs.

It is to acknowledge you that the Executive Committee of *International Journal of Instruction* has decided that the article mentioned below would be reviewed by you. Thank you very much for your contributions.

Asim ARI
Editor in Chief

Name of the article: The Development of HOTS Test of Physics Based on the Modern Test Theory: Question Modeling through E-learning of Moodle LMS

After reviewing the attached article, please read each item carefully and select the response that best reflects your opinion. To register your response, please **mark** or **type in** the appropriate block.

	Yes	Partially	No
Do you think the title is appropriate?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Does the abstract summarize the article clearly and effectively?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Are the objectives set clearly?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the issue stated clearly?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the literature review adequate?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the design of the research appropriate, and the exemplary, if any, suitable?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the methodology consistent with the practice?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Are the findings expressed clearly?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the presentation of the findings adequate and consistent?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Are the tables, if any, arranged well?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Are the conclusions and generalizations based on the findings?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Are the suggestions meaningful, valid, and based on the findings?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Are the references adequate?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the language clear and understandable?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is cohesion achieved throughout the article?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the work contributing to the field?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

- Evaluation:**
- The article can be published as it is.
 - The article can be published after some revision.
 - The article must undergo a major revision before it can be resubmitted to the journal.
 - The article cannot be published.

Would you like to see the revised article if you have suggested any revisions? Yes No

Please write your report either on this paper or on a spare paper.

REPORT

Section of the Manuscript	Comments and Notes
Title- Abstract-Summary	Abstract is good, but maybe the title could be shorter
Introduction and Literature Review	The introduction and the literature review have good style. It will be good for the article if the literature review has more new references (from 2017 and younger)
Research Methods	Everything is good
Research Findings	Everything is good



Discussion	
Conclusion and Suggestions	<p>After figure 8 author(s) write: "The analysis result brings out the fact that students tend to find it difficult to answer questions regarding the creating aspect, especially the planning sub-aspect.", and "This figure also signifies that the majority of the students find it easy to answer HOTS questions related to the analysis aspect, differentiating sub-aspect in particular." After figure 9 "Figure 9 shows the percentage of students' HOTS. It is seen that most students (41.2%) still have low HOTS; the categories consist of very low (20.6%) The low category of students' HOTS is influenced by several factors, one of which is that the students are not used to working on HOTS questions". At the same time in the conclusions we can find "To sum up, all 24 question items that had been tried out have fit the model with a good category, so that they can be utilized in HOTS measurement." and nothing about student's low HOTS. Maybe the author(s) could add some more information in the conclusions which could show the positive and negative results of this research?</p>
References and Citation	It will be good for the article if the literature review has more new references (from 2017 and younger)
Language	Everything is good
Other issues	Everything is good



International Journal of Instruction Article Evaluation Form

Mr. /Mrs.

It is to acknowledge you that the Executive Committee of *International Journal of Instruction* has decided that the article mentioned below would be reviewed by you. Thank you very much for your contributions.

Asim ARI
Editor in Chief

Name of the article: The Development of HOTS Test of Physics Based on the Modern Test Theory: Question Modeling through E-learning of Moodle LMS

After reviewing the attached article, please read each item carefully and select the response that best reflects your opinion. To register your response, please **mark** or **type in** the appropriate block.

	Yes	Partially	No
Do you think the title is appropriate?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Does the abstract summarize the article clearly and effectively?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Are the objectives set clearly?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Is the issue stated clearly?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the literature review adequate?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Is the design of the research appropriate, and the exemplary, if any, suitable?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the methodology consistent with the practice?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Are the findings expressed clearly?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Is the presentation of the findings adequate and consistent?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Are the tables, if any, arranged well?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Are the conclusions and generalizations based on the findings?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Are the suggestions meaningful, valid, and based on the findings?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Are the references adequate?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Is the language clear and understandable?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is cohesion achieved throughout the article?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Is the work contributing to the field?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

- Evaluation:**
- The article can be published as it is.
 - The article can be published after some revision.
 - The article must undergo a major revision before it can be resubmitted to the journal.
 - The article cannot be published.

Would you like to see the revised article if you have suggested any revisions? Yes No

Please write your report either on this paper or on a spare paper.

REPORT

Section of the Manuscript	Comments and Notes
Title- Abstract-Summary	Please add the conclusion on abstract
Introduction and Literature Review	authors must be explained gap analysis and novelty of the study explicitly (please add 5 recent studies (2015-2020))
Research Methods	Clear
Research Findings	Results should be focus on explanation of main finding, do not written references.



Discussion	discussion must be written in detail. The discussion should explore the significance of the results of the study. The references contained in the introduction should not be re-written in the discussion. A comparison to the previous studies should be presented. The following components should be covered in discussion: How do your results relate to the original question or objectives outlined in the background section (what)? Do you provide interpretation scientifically for each of your results or findings presented (why)? Are your results consistent with what other investigators have reported (what else)? Or are there any differences? Please add 10 pervious studies (2010-2020) to compare your finding.
Conclusion and Suggestions	The conclusion does not contain the repetition of the results and discussion, but rather the summary of the findings as expected in the objectives or hypotheses. If necessary, at the end of the conclusion can also be written the things that will be done related to the next idea of the study. The conclusion is written in the whole paragraph, not the points per point.
References and Citation	add previous studies on introduction and discussion
Language	good
Other issues	-



International Journal of Instruction Article Evaluation Form

Mr. /Mrs.

It is to acknowledge you that the Executive Committee of *International Journal of Instruction* has decided that the article mentioned below would be reviewed by you. Thank you very much for your contributions.

Asim ARI
Editor in Chief

Name of the article: The Development of HOTS Test of Physics Based on the Modern Test Theory: Question Modeling through E-learning of Moodle LMS

After reviewing the attached article, please read each item carefully and select the response that best reflects your opinion. To register your response, please **mark** or **type in** the appropriate block.

	Yes	Partially	No
Do you think the title is appropriate?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Does the abstract summarize the article clearly and effectively?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Are the objectives set clearly?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Is the issue stated clearly?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Is the literature review adequate?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Is the design of the research appropriate, and the exemplary, if any, suitable?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Is the methodology consistent with the practice?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Are the findings expressed clearly?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Is the presentation of the findings adequate and consistent?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Are the tables, if any, arranged well?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Are the conclusions and generalizations based on the findings?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Are the suggestions meaningful, valid, and based on the findings?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Are the references adequate?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the language clear and understandable?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Is cohesion achieved throughout the article?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Is the work contributing to the field?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

- Evaluation:**
- The article can be published as it is.
 - The article can be published after some revision.
 - The article must undergo a major revision before it can be resubmitted to the journal.
 - The article cannot be published.

Would you like to see the revised article if you have suggested any revisions? Yes No

Please write your report either on this paper or on a spare paper.

REPORT

Section of the Manuscript	Comments and Notes
Title- Abstract-Summary	<p>Abstract: This abstract section is almost correct, but it has not written about the background and the aims of the research. The author could re-write background details and the aims of the research by using present tense as a time signal.</p> <p>The abstract should begin by mentioning the topic and its significance in the scientific or academic requirements.</p> <p>The abstract has not explained what HOTS and Question Modeling are. What is the relation both of them? After that, why these new methods are proper? What are the main</p>



	<p>influences?</p> <p>Summary: This part has not specified clearly regarding conclusions and the effect concerning other works (if they exist) or even as scientific data. The author could use example sentences like these to support the conclusion; This indicates that ... or It appears that ... or There might be a necessity for revising the list of criteria within the next ...</p>
Introduction and Literature Review	<p>Introduction: This section has explained background about the meaning of developing Modern HOTS Test.</p> <p>However, in this section cannot explain why HOTS is very important to be implemented and what solutions are applied or essential objectives of this research.</p> <p>What are C4, C5, and C6? Are they codes?</p> <p>This part also has not yet been explained in more detail the art of research (innovation) that will be carried out in the research.</p> <p>Maybe this part can be interpreted from the sources and their relationship with the urgency of the study</p> <p>.</p> <p>For instance, using the phrase "Therefore, the purpose of the HOTS Test is based on Modern theory tests are for ..."</p> <p>So, which one is Modern Test Theory? IRT or e-learning Moodle LMS? And what does their relationship with HOTS?</p> <p>Who are the participants in this research? Basic physics students or General Physics students?</p> <p>Literature Review: Maybe the author could combine previous research with improving HOTS, and what can be developed from the previous studies or previous research? Or is this research still really new? If this part is explained, the author has provided a more precise research basis for the readers.</p>
Research Methods	<p>The next chapter is a research method that explains the technique and the model that was used.</p> <p>It will be better if the author provides information about sampling collection techniques. the readers need to understand the basic method to take the sample.</p> <p>This method part has not described in a specific about time of the research. And for knowing Physics scores online, how long does it take to answer the questions? Or how much time is given to participants to finish the total questions?</p> <p>In the Rasch model that uses the Quest program, what type of data is for input? It must be explained that the reader can be more understands.</p> <p>The picture number and image title should not be separated from entering.</p> <p>The table position should be in the middle of the paragraph.</p> <p>Decrease the dimension of the table cell width to make it more aesthetic</p>
Research Findings	<p>This paper must declare clearly to show what's the research finding going.</p> <p>What is the most powerful impact from HOTS Test of Physics Based on the Modern Test Theory for Physic student to make them understand Physics easily?</p>
Discussion	<p>The next part is the results and discussion, which is the centre part of the article. The first part of the results and the author's discussion explains the ADDIE development model and the description of each stage result. This section should not be involved in the results and discussion section but in the research method part.</p> <p>And, if the test is a multiple-choice question, how is it linked to HOTS? Doesn't HOTS require analysis, evaluation, and creation?</p> <p>What do the figures in Table 7 show? Showing how many values were answered correctly, or what? Or the value of each physical quantity? Then explain the data in the table and how it correlates to the results.</p> <p>What findings were obtained? keep in mind to define the unit used</p> <p>After declaring the research findings, the research findings and relevant theories or hypotheses must be comprehensively discussed.</p> <p>What is the difference between Estimates for Items and Estimates for Testes?</p> <p>What is logg-odd unit?</p> <p>Normal curve? Where is the curve? Why can it explain the skills of students?</p> <p>Explain in more detail in Figure 9. It will illustrate how HOTS each student has. So that they can analyze more deeply whether the applied model can improve HOTS students or not? If yes or no, explain it in more detail and relate it to existing theories or references.</p>

<p>Conclusion and Suggestions</p>	<p>The conclusion must be written concisely (briefly and understandably). It should not re-discuss the results of research. The outcome/conclusion must be able to acknowledge the aim of the research question and may not repeat the abstract or rewrite the results of the experiment. Conclusions part Needs to be improved. What goals will be achieved so that in writing conclusions can be read clearly? This section must reflect the innovation or improvement of existing knowledge. Some suggestions related to results can be computed.</p>
<p>References and Citation</p>	<p>Reerence and citation are desirable.</p>
<p>Language</p>	<p>The sentences are almost perfect, but too many words are often overused. Consider using a more specific synonym to improve the sharpness of witing. There are still many uses of punctuation that is not yet definite, causing meaning in the sentence ineffective. It may be unclear who or what "This" refers to.</p> <p>Consider rewriting the sentence to eliminate the ambiguous reference. Many sentences should be written in active voice than the passive voice. The active voice can provide more clarity, brevity, accountability, or certainty than the passive voice.</p> <p>Inaccurate disposition of adverb of time and place. The picture number and image title should not be separated from entering. Pay attention to equality in the sentence. For example, verbs must be parallel with verbs. the adjective must be parallel with the adjective, etc</p>
<p>Other issues</p>	<p>The manuscript is not enough informing and comprehensive. It also doesn't have enough methodological. Please revise this article with the comments and note that has been mentioned above.</p>



International Journal of Instruction Article Evaluation Form

Mr. /Mrs.

It is to acknowledge you that the Executive Committee of *International Journal of Instruction* has decided that the article mentioned below would be reviewed by you. Thank you very much for your contributions.

Asim ARI
Editor in Chief

Name of the article: The Development of HOTS Test of Physics Based on the Modern Test Theory: Question Modeling through E-learning of Moodle LMS

After reviewing the attached article, please read each item carefully and select the response that best reflects your opinion. To register your response, please **mark** or **type in** the appropriate block.

	Yes	Partially	No
Do you think the title is appropriate?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Does the abstract summarize the article clearly and effectively?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Are the objectives set clearly?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the issue stated clearly?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Is the literature review adequate?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Is the design of the research appropriate, and the exemplary, if any, suitable?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Is the methodology consistent with the practice?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Are the findings expressed clearly?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the presentation of the findings adequate and consistent?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Are the tables, if any, arranged well?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Are the conclusions and generalizations based on the findings?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Are the suggestions meaningful, valid, and based on the findings?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Are the references adequate?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Is the language clear and understandable?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Is cohesion achieved throughout the article?	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Is the work contributing to the field?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

- Evaluation:**
- The article can be published as it is.
 - The article can be published after some revision.
 - The article must undergo a major revision before it can be resubmitted to the journal.
 - The article cannot be published.

Would you like to see the revised article if you have suggested any revisions? Yes No

Please write your report either on this paper or on a spare paper.

REPORT

Section of the Manuscript	Comments and Notes
Title- Abstract- Summary	The abstract is too long and not written effectively. Please write abstract briefly, but consisting the main findings of your research or the contribution of the field. The study have failed to convince that is worthy enough, because there is no state of the art among other related research or the issue in general.
Introduction and Literature Review	
Research Methods	It is just a preliminary study to develop an instrument. However, ADDIE model of instructional design is comprehensive framework to create a new product or methods of



	learning. The research should proof about the effectivity of the instrument by using experiments methods. Comparing the instrument with other tools of measurement which would be contribute to improving students' HOTS.
Research Findings	
Discussion	The discussion should be describe about novelty of the research and how it works.
Conclusion and Suggestions	The Conclusion Section just briefly repeated what have been written on the Findings and Discussion Section.
References and Citation	Not enough references.
Language	There are too many wrong grammatically usage and unsuitable word choices. It should be written correctly.
Other issues	

The Development of HOTS Test of Physics Based on the Modern Test Theory: Question Modeling through E-learning of Moodle LMS

The present study discussed the development of higher-order thinking skills (HOTS) test of physics based on the modern test theory. HOTS questions were designed and presented in the e-learning with the Moodle learning management system (LMS) that could be accessed online. This study employed the ADDIE model with analysis, design, development, implementation, and evaluation stages. The instrument consisted of 24 multiple choice physics questions regarding the direct current circuit topic; the questions were designed by following the aspects and sub-aspects of HOTS and had been validated by the experts of measurement, physics education, physics, and practitioners. Moreover, validity analysis was based on the V Aiken formula, in which every aspect was confirmed valid. The validated instrument was then tried out to all 34 students at the Department of Physics Education, Universitas Papua, who participated in the basic physics subject. Dichotomy data analysis used the Rasch Model (RM) 1-PL through the Quest program, and the test characteristics comprised item fitness, reliability, and difficulty. The trial result obtained the criteria of INFIT MNSQ mean and standard deviation of 1.0 and 0.0, respectively, showing that the items fitted the RM1-PL. In addition, the value of item reliability based on the value summary of the item estimate arrived at 0.66; meanwhile, the case reliability under the summary of the case estimate accounted for 0.85. The reliability value in the range of 0.67- 0.80 was categorized as quite reliable. As based on the criteria of minimum and maximum INFIT MNSQ of 0.77 and 1.30, 24 question items fitted the RM 1-PL model. The result of the Quest output also revealed that the average values of Thresholds and its standard deviation were 0.00 ± 0.71 , or in the acceptance range of -2 to 2. All in all, all 24 question items that had been tried out had fitted the model with a good category in order that they could be utilized in HOTS measurement.

Keywords: E-learning, HOTS Test, and Modern Test Theory.

INTRODUCTION

Assessment, particularly in the cognitive domain, is central to the learning process and should be carried out accurately and in compliance with the subject to be assessed or measured. Students' cognitive skills in the learning process can be categorized into lower-order thinking (LOT) and higher-order thinking (HOT). The LOTS include remembering, understanding, and applying; the HOTS, on the other hand, encompass analyzing, evaluating, and creating. HOTS are thinking skills that do not only require the remembering skill but also require other higher skills. Indicators to measure HOTS consist of analyzing (C4), evaluating (C5), and creating (C6) skills (Krathwohl & Anderson, 2010).

HOTS also refer to thinking skills when one takes new information, connects it with initial information s/he has, and finally delivers the information to achieve goals or answer questions (Istiyono, Dwandaru, & Muthmainah, 2019). This is in line with skill characteristics in the 21st century published by Partnership of 21st Century Skill stating

Commented [p1]: Please add the conclusion after this sentences

that 21st century learners should be able to develop competitive skills, such as critical thinking, problem-solving, communication, information and communication technology (ICT) literacy, ICT, information literacy, and media literacy (Brun & Hinostroza, 2014); these focus on HOTS development.

Physics serves as part of science consisting of abstract concepts that are difficult to be directly described. Learning physics is expected to help students develop their thinking skills, in which they are not only demanded to master LOT skills, but also HOTS. Teachers are also urged to deliver learning materials to students, including the HOTS that can be improved by HOTS instrument. A previous study has reported that the majority of teachers find it challenging to develop an assessment instrument of learning outcomes, HOTS questions, in particular (Istiyono, 2018). For this reason, teacher creativity is highly required to measure students' learning outcomes. Today's development of Information and Communication Technology (ICT) can be utilized to design and habituate students to learn anywhere at any time (Yusuf, Widyaningsih, & Sebayang, 2018). Relying on ICT during the learning process is one of the significant innovations, including in the evaluation of students' learning outcomes.

The presentation of evaluation questions can be done in an integrated manner through e-learning programs, one of which is Moodle learning management system (LMS) (Azevedo, 2015; Bogdanović, Barać, Jovanić, Popović, & Radenković, 2014). The Moodle provides different types of questions, such as multiple choices, true or false, and short answers; these are stored in the taught course database and can be re-used (Limongelli, Sciarone, & Vaste, 2011). Teachers are also able to give feedback directly to the students and give them correct answers to questions they have worked on (Pandey & Pandey, 2009). One of the advantages of an online evaluation through Moodle LMS is that students can directly figure out their assessment results.

Teachers need to prepare a good test to measure students' learning outcomes. There are two paradigms developed for students' learning outcome assessment through the applied test, i.e., classical and modern approaches. The classical paradigm being utilized is classical test theory or widely known as classical true-score theory, meanwhile, the modern paradigm is item response theory (IRT). The classical test theory is selected due to its ease in the application despite of its limitations in measuring the item difficulty level and discrimination since the calculation of both indicators is based on the test taker's total score. In contrast, the IRT frees up the dependence between the test item and test taker (a concept of parameter invariance); the test taker's response to a test item does not affect another item (a concept of local independence), and; the test item does only measure one measurement dimension (unidimensional concept) (Raykov & Marcoulides, 2015). Therefore, the application answers the needs of modern measurement to date, i.e., a comparison between test taker's skills, question development, and even adaptive test development, so that it is considered able to overcome the classical test theory limitations.

This development study is an initial study with a long-term purpose of developing general physics questions with good quality at the Department of Physics Education,

Commented [p2]: After this paragraph, authors must be explained gap analysis and novelty of the study explicitly

Universitas Papua. As the first stage, this study focuses on students at the department mentioned previously who enroll in General Physics subject taught by the researcher. This study also serves as one of the efforts to expand students' HOTS by applying a variety of HOTS-based learning sources.

Commented [p3]: The aim of study must be written after this sentences

METHOD

The ADDIE model, as employed by this study, refers to a general and systematic model of development study with a phased framework, allowing each element to connect with each other (Aldoobie, 2015). The stages of this model used in the development of HOTS instrument are presented in Figure 1.

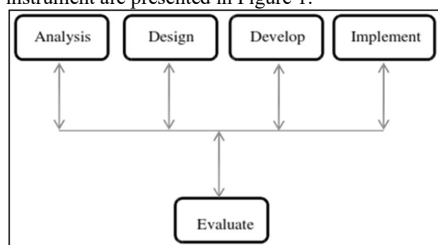


Figure 1
Stages of ADDIE Development Model in Designing Moodle LMS-based HOTS Test.

Analysis

The analysis stage is a process of needs analysis in the form of determining test objectives, identifying problems, analyzing tasks, and determining question formats to be applied. It is revealed that the problems are related to the needs of HOTS instrument design for students at the Department of Physics Education, Universitas Papua.

Design

This stage comprises the process of designing HOTS questions to be used; the design process encompasses creating a question matrix and outline that covers question distribution in every aspect and sub-aspect of HOTS.

Develop

Moreover, every single thing required in the arrangement of HOTS skill questions has been prepared in the next stage. This stage also comprises the process of making the questions regarding HOTS, as well as validating the questions that involve the experts of measurement, physics education, and practitioners. The technique of validity analysis to assess the content validity of the developed questions applies the V Aiken formula (Aiken, 1980, 1985).

$$V = \sum s / n(c-1) \tag{1}$$

“V” refers to the agreement index of validators in regards to item validity; “s” is the assessment score of validators subtracted by the assessment lowest score; “n” refers to the number of validators; “c” is the number of categories that can be chosen by validators. All test items are considered valid if the value of the V Aiken index falls into the range of $0.37 - 1$ (Kowsalya, Venkat Lakshmi, & Suresh, 2012). The value of V Aiken of every test item is calculated based on the assessment items of every validator. In this stage, there is also an evaluation process, i.e., revising questions by following validators’ corrections and suggestions.

Implementation

Another stage is applying HOTS questions that have been developed to 34 students in the site area who enroll in general physics subject. This number has been following the sample size for data stability in Rasch Model (RM) 1- PL, which is from 30 to 300, with the limit of INFIT t is from -2 to +2 (Bond & Fox, 2007). Question item analysis is performed based on the raw score of the students by employing the Quest program.

Evaluation

Evaluation is a process of finding out whether or not the developed questions of HOTS have met the expectation. The evaluation stage is carried out in every stage and called a formative evaluation intended for revisions (Lee & Zainal, 2017). For instance, in the design stage, the expert’s review is necessary to provide input towards the design. Further, the evaluation stage is undertaken after analyzing empirical questions mathematically by using the Quest software program by referring to the Rasch model. The Quest program is able to do the Rasch measurement, i.e., a comprehensive empirical test of question items. There are three parameters being measured mathematically based on the empirical test of question items.

1. The first parameter is item fitness with the Rasch model by following the value of INFIT MNSQ or INFIT t of the item. The expected values of the unweighted mean square (Outfit MNSQ) in the Quest program and weighted mean square are 1; the variance is 0. On the contrary, the expected value of Mean INFIT t is equal to 0, with the variance equal to 1 (Adams & Khoo, 1996). The provision of INFIT MNSQ for the Rasch Model is shown in Table 1 and Table 2.

Tabel 1
Criteria of Question Item Fitness with the Rasch Model

MNSQ INFIT Value	Criteria
>1,33	Does Not Fit the Rasch Model
0,77 s.d. 1,33	Fits the Rasch Model
<0,77	Does Not Fit the Rasch Model

Tabel 2
The Provision of Outfit t for the Rasch Model.

t OUTFIT Value	Criteria
OUTFIT t \leq 2,00	Fits the Rasch Model
OUTFIT t \geq 2,00	Does Not Fit the Rasch Model

2. The second parameter is reliability. The analysis result of the Quest program also reveals the item and case reliability. The reliability value based on the item estimate is also called as sample reliability; the higher the value, the more the items that fit the tested model. Whereas, the lower the value, the less the items that fit the tested model, so that it does not give the expected information. The reliability category is provided in Table 3 (Istiyono, 2017).

Tabel 3
Interpretation of Reliability Value

Reliability Value	Criteria
> 0,94	Excellent
0,91 – 0,94	Very Good
0,81 – 0,90	Good
0,67 – 0,80	Acceptable
< 0,67	Poor

3. The third parameter is item difficulty index and respondents' skills presented as difficulty index in the Quest output. Thresholds (THRSHL) show the item difficulty index in the logit scale along with its standard deviation (Hambleton & Rogers, 1989). The provision of the THRSHL value is given in Table 4.

Tabel 4
Criteria of THRSHL Value to Categorize Item Difficulty Level

THRSHL Value	Criteria
$b > 2,00$	Very Difficult
$1,00 < b \leq 2,00$	Difficult
$-1,00 < b \leq 1,00$	Medium
$-1,00 > b \geq 2,00$	Easy
$b < -2,00$	Very Easy

Respondents' skills are shown by the value of the estimate error, in which the criteria of the estimate value of respondents' skills are presented in Table 5.

Tabel 5
Criteria of Estimate Value to Categorize Respondents' Skills

THRSHL Value	Criteria
$b > 2,00$	Very Difficult
$1,00 < b \leq 2,00$	Difficult
$-1,00 < b \leq 1,00$	Medium
$-1,00 > b \geq 2,00$	Easy
$b < -2,00$	Very Easy

The evaluation stage also includes the process of analyzing the HOTS of students on the whole. The level of HOTS is categorized based on the ideal mean and standard deviation. This is applied with the assumption that students' HOTS of physics are normally distributed. The ideal mean (I_m) and ideal standard deviation (I_{sd}) are based on the highest and lowest score of research variables. Table 6 shows the criteria of students' HOTS of physics.

Tabel 6
Criteria of Students' HOTS of Physics

Interval	Criteria
$Im + 1,5 Isb < 0$	Very high
$Im + 0,5 Isb < 0 \leq Im + 1,5 Isb$	High
$Im - 0,5 Isb < 0 \leq Im + 0,5 Isb$	Medium
$Im - 1,5 Isb < 0 \leq Im - 0,5 Isb$	Low
$0 < Im - 1,5 Isb$	Very Low

Meaning:

Im : ideal mean

Isb : ideal standard deviation

X_{mak} : highest score

X_{min} : lowest score

RESULTS AND DISCUSSION

ADDIE development model can be used for different product developments in education, and one of which is the development of HOT skill questions. This model is simple and systematically structured in its implementation stages. The following is the description of each stage result.

Analysis

Needs analysis is the first stage being done by observation and interview to gather any information needed in the process of physics learning at the Department of Physics Education, Universitas Papua. The researcher's experience indicates that lecturers have applied HOTS learning in the classroom. However, a test to measure students' HOTS has not been conducted. The arrangement of HOTS instrument is required to train and develop students' HOTS. Accordingly, to facilitate the students in accessing other learning sources, this study designs HOT skill questions in an online system through an e-learning program using the Moodle LMS.

Design

In the design stage, the test instrument is designed based on the analysis result in the first stage. Test instrument design in this stage is in the form of question matrix and outline which are adjusted to students' needs and characteristics, and learning sources. The test is a multiple-choice test, in which 24 questions are adjusted to the formulation of a HOTS test that has been created in the test matrix and outline. The question matrix is provided in Table 7.

Tabel 7

The Question Matrix

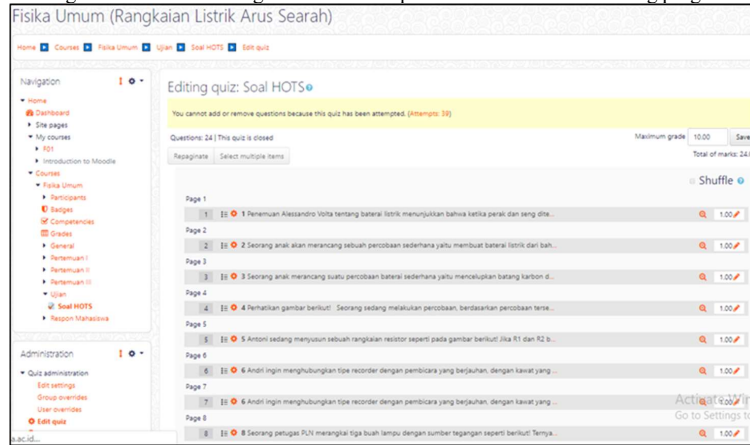
Aspect	Sub Aspect	Theory		
		Electric current, Ohm's law, and electrical power	Series and parallel circuits of resistor and capacitor	Electric Force, Kirchoff's law, and RC circuit.
Analyze	Differentiating	8	12	21

	Organizing	3	15	20
	Attributing	2	9	23
Evaluate	Checking	4	11	22
	Critiquing	1	16	18
Create	Generating	5	13	19
	Planning	7	14	17
	Producing	6	10	24

Commented [p4]: This table must be explained in detail.

Develop

The development of HOTS questions is based on the question matrix and outline that have been designed. Further, the questions are made online through e-learning by utilizing the Moodle LMS. Figure 2 shows all question items in the e-learning program.

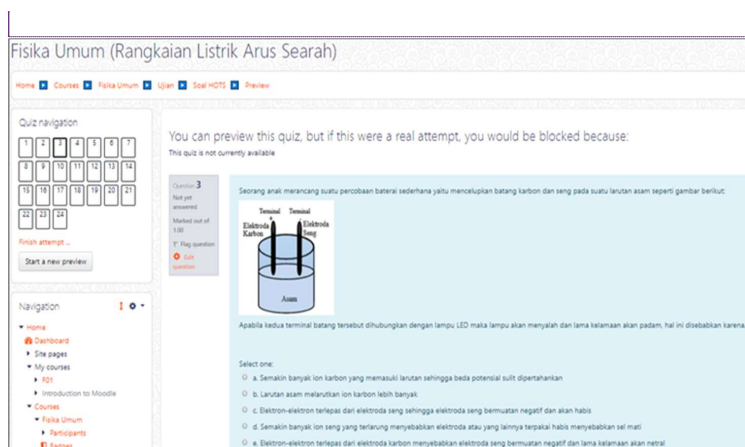


Commented [p5]: Please improve resolution of figure

Figure 2 Shows All Question Items in the E-Learning Program

Moodle LMS program presents an interesting display and is easy to access by users (Martín-Blas & Serrano-Fernández, 2009). The questions are displayed interactively, and students can randomly work on the questions. Moodle LMS can present questions with a picture or other content to make it easier for teachers to design the questions as expected. Figure 3 shows one of the HOTS questions displayed on the e-learning through the Moodle LMS.

Commented [p6]: The references should be written on the discussion.



Commented [p7]: The resolution of figure must be improve

Figure 3
Shows of the HOTS Questions Displayed on the E-learning Through the Moodle LMS

The development stage aims to produce a HOTS test instrument that has been validated by experts and practitioners. Product validation is a process of assessing the designed product, or in this case, the test instrument of HOTS in general physics subject in the site area. Product validation is carried out by involving seven validators, i.e., experts of measurement, physics education, physics, and practitioners. The validity test of the instrument includes material, construction, and language. The analysis result of question validity that is assessed by validators obtains the value of V Aiken in the range of 0.76 - 1.00, showing a valid result. The questions validated by experts and practitioners are then revised based on provided corrections and suggestions.

Implementation

The implementation stage in this study is the product trial, in which HOTS questions are tried out to 34 students in the research site. The students work on these questions via online through e-learning by using their own Moodle account upon the completion of all learning stages. Results of the students' learning can be accessed after this process.

Evaluation

Before conducting the estimate analysis of respondents' skills and item difficulty level, the analysis of item fitness is performed by using parameters of INFIT and OUTFIT for mean square and t . The determination of the item fitness with the model is based on the value of INFIT MNSQ and the standard deviation or Infit t (Adams & Khoo, 1996). The fitness of each case is also based on the value of INFIT MNSQ or INFIT t of the item.

Commented [p8]: Results should be focus on explanation of main finding, do not written references.

Table 8 provides the testing result through the Quest program to obtain the values of item estimate and case estimate in the HOTS questions trial.

Tabel 8
Values of Item Estimate and Case Estimate in the HOTS Questions Trial

No.	Measurement	Estimates for Items	Estimates for Testi
1.	Average values and standard deviations	$0,00 \pm 0,57$	$0,01 \pm 1,24$
2.	Reliability Estimates	0,66	0,85
3.	The mean value and standard deviation of INFIT MNSQ	$1,00 \pm 0,14$	$0,99 \pm 0,15$
4.	The mean value and standard deviation of OUTFIT MNSQ	$1,09 \pm 0,52$	$1,09 \pm 0,52$
5.	The mean value and standard deviation of INFIT t	$-0,03 \pm 0,81$	$0,00 \pm 0,72$
6.	The mean value and standard deviation of OUTFIT t	$0,21 \pm 0,91$	$0,17 \pm 0,81$

The analysis result reveals that the INFIT MNSQ arrives at the range of 0.86 - 1.14, and INFIT t is -0.28 - 0.72. This result signifies that all 24 questions fit the model as they reach the range of INFIT MNSQ value from 0.77 to 1.30 and use INFIT t with the limit of -2.0 - 2.0 [16]. In addition to testing the fitness, the output of the Quest program also presents the reliability estimate of the test instrument. Table 8 provides the value of item reliability based on the value of summary of item estimate, which is 0.66. On the other hand, the value of person reliability, as based on the summary of case estimate, gets 0.85. These results are in line with the Rasch model, in which the reliability value falls under the range of 0.67 - 0.80 (quite reliable). On that ground, the instrument can be used to measure students' HOTS in the General Physics subject.

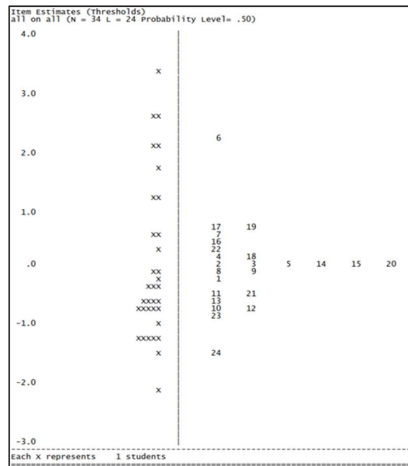


Figure 4
Distribution of Item Difficulty Level and Respondents' Skills

Figure 4 presents the distribution of the respondents according to the difficulty level in the logit scale from -4.0 to +4.0. This map displays the item difficulty level compared to the respondents' skills. Case and item difficulty levels in the Rasch model are expressed in one line in the form of abscissa in the graph with logg-odd unit. The graph of respondents' skills shows a normal curve, meaning that there are only a few respondents with low and high skills; and a lot of respondents with moderate skills. The level of item difficulty of threshold reveals that item 6 is the most difficult question, and item 24 is the easiest one.

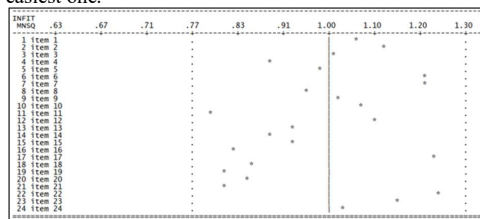


Figure 5
Distribution of INFIT MNSQ Values of Each Question Item of HOTS

Question items that fit the Rasch model are in the range of 0.77 - 1.33. Figure 5 shows that all 24 question items are in the line, implying that they fit the Rasch model.

Author surnames go here

11

Item Estimates (Thresholds) In input Order							
all on all (N = 34 L = 24 Probability Level= .50)							
ITEM NAME	SCORE	MAXSCR	THRSH	INFT	OUTFT	INFT	OUTFT
			1	MNSQ	MNSQ	t	t
1 item 1	18	34	-.26	1.06	1.15	.4	.5
			-.39				
2 item 2	16	34	.04	1.12	1.17	.7	.6
			-.40				
3 item 3	16	34	.04	1.01	.91	.1	-.2
			-.40				
4 item 4	15	34	.19	.88	.93	-.6	-.1
			-.40				
5 item 5	16	34	.04	.98	.89	.0	-.2
			-.40				
6 item 6	5	34	2.27	1.21	2.16	.7	1.4
			-.57				
7 item 7	13	34	.52	1.21	1.27	1.0	.9
			-.42				
8 item 8	17	34	-.11	.96	1.00	-.2	.1
			-.40				
9 item 9	17	34	-.11	1.02	.91	.2	-.2
			-.40				
10 item 10	21	34	-.70	1.07	1.16	.6	.5
			-.39				
11 item 11	19	34	-.41	.79	.66	-1.6	-.9
			-.39				
12 item 12	21	34	-.70	1.10	1.14	.8	.5
			-.39				
13 item 13	20	34	-.55	.93	1.09	-.5	.4
			-.39				
14 item 14	16	34	.04	.88	.78	-.7	-.6
			-.40				
15 item 15	16	34	.04	.93	.82	-.4	-.5
			-.40				
16 item 16	13	33	.47	.82	.69	-.8	-.9
			-.42				
17 item 17	12	34	.69	1.23	1.16	1.0	.6
			-.43				
18 item 18	15	34	.19	.86	.73	-.8	-.8
			-.40				
19 item 19	12	34	.69	.81	.71	-.8	-.8
			-.43				
20 item 20	16	34	.04	.85	.75	-.9	-.7
			-.40				
21 item 21	19	34	-.41	.81	.68	-1.4	-.8
			-.39				
22 item 22	14	34	.35	1.24	1.23	1.2	.8
			-.41				
23 item 23	22	34	-.85	1.15	3.04	1.1	3.1
			-.40				
24 item 24	26	34	-1.50	1.03	1.02	.2	.2
			-.43				
Mean			.00	1.00	1.09	.0	.1
SD			.71	.14	.52	.8	.9

Figure 6
Item Estimates from HOTS Questions

Figure 6 presents the Item Estimate of HOT skill questions based on the trial result. In this figure, there is SCORE-MAXSCR successively showing the score of the respondents who answer correctly, and the number of total respondents. Item 24 is the most correctly-answered, in which 26 out of 34 respondents are able to work on this item. Figure 6 also provides the value of THRSHL that shows the item difficulty index in the logit scale along with its standard deviation. Item 6 has THRSHL or difficulty index of 2.27 that is greater than 2.0, or in other words, this item is very difficult since only

five students can give a correct answer. The average value of THRSHL and its standard deviation accounts for 0.00 ± 0.71 and falls into the range of -2 - 2 (Hambleton & Rogers, 1989). The average value of INFIT MNSQ is 1.00 ± 0.14 and falls under the acceptance range of 0.77 - 1.33; the average value of OUTFIT t arrives at 0.10 ± 0.90 and falls into the acceptance range of ≤ 2.00 . All of these results indicate that all question items that have been developed can be employed to measure students' HOTS.

Case Estimates In Input Order								
all on all (N = 34 L = 24 Probability Level= .50)								
NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFIT MNSQ	OUTFIT MNSQ	INFIT t	OUTFIT t
1 01	12	24	-.02	.43	1.06	1.02	.58	.18
2 02	6	24	-1.21	.49	1.17	1.09	-.75	.36
3 03	8	24	-.77	.45	.98	1.01	-.06	.13
4 04	8	24	-.77	.45	.83	.81	-1.04	-.48
5 05	8	24	-.77	.45	.89	.83	-.59	-.41
6 06	6	24	-1.21	.49	.79	.70	-.84	-.67
7 07	10	24	-.38	.43	.99	.95	-.01	-.09
8 08	6	24	-1.21	.49	1.07	2.30	-.36	2.44
9 09	3	24	-2.10	.63	.98	.85	-.11	.00
10 10	9	24	-.57	.44	.88	.83	-.80	-.48
11 11	22	24	2.61	.77	.73	.46	-.30	-.56
12 12	5	24	-1.46	.52	.89	.85	-.29	-.18
13 13	20	24	1.75	.57	1.21	1.45	.64	.93
14 14	11	24	-.20	.43	.86	.83	-1.33	-.59
15 15	21	24	2.12	.64	1.18	1.05	.52	.29
16 16	9	24	-.57	.44	1.08	1.06	.59	.28
17 17	7	24	-.98	.47	1.29	2.20	1.38	2.55
18 18	6	24	-1.21	.49	1.23	1.28	.96	.75
19 19	14	24	-.35	.43	.92	.87	-.56	-.40
20 20	15	24	.54	.44	.97	1.09	-.13	.40
21 21	18	24	1.19	.49	.94	.86	-.16	-.25
22 22	21	24	2.12	.64	.93	1.23	-.01	.55
23 23	9	24	-.57	.44	1.07	1.01	.54	.15
24 24	8	24	-.77	.45	1.01	.95	-.13	-.03
25 25	10	24	-.38	.43	.87	.82	-1.06	-.57
26 26	15	24	.54	.44	1.05	1.22	-.36	.80
27 27	6	24	-1.21	.49	.82	.74	-.69	-.56
28 28	22	24	2.61	.77	.73	.46	-.30	-.56
29 29	12	24	-.02	.43	.92	.88	-.73	-.39
30 30	9	24	-.57	.44	.90	.90	-.64	-.23
31 31	23	24	3.40	1.05	1.18	3.14	.49	1.53
32 32	18	24	1.19	.49	.85	.75	-.53	-.58
33 33	10	23	-.32	.44	1.11	1.11	.97	.45
34 34	8	24	-.77	.45	1.29	1.34	1.61	1.03
Mean			.01		.99	1.09	.00	.17
SD			1.35		.15	.52	.72	.81

Figure 7
Case Estimates from Every Student

Figure 7 serves as the case estimate or the skill level of each student. Information obtained from the case estimate is that the SCORE-MAXSCR shows the score of each respondent from the maximum score sequentially. Respondent 31 answers the most questions (23 out of 24 questions) correctly compared to other respondents. The average estimate value and its standard deviation gets 0.01 ± 1.35 and falls under a moderate category. The analysis result of the case estimate reveals that students' skills are in the moderate category.

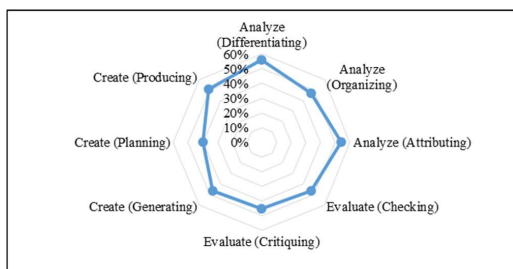


Figure 8
Distribution of Student Answer Percentage HOTS

Figure 8 gives the percentage of students' answers based on the aspects and sub-aspects of HOTS. The analysis result brings out the fact that students tend to find it difficult to answer questions regarding the creating aspect, especially the planning sub-aspect. Creating is the highest level HOTS in Bloom's taxonomy, which therefore, students need to practice developing their creating skills. This figure also signifies that the majority of the students find it easy to answer HOTS questions related to the analysis aspect, differentiating sub-aspect in particular.

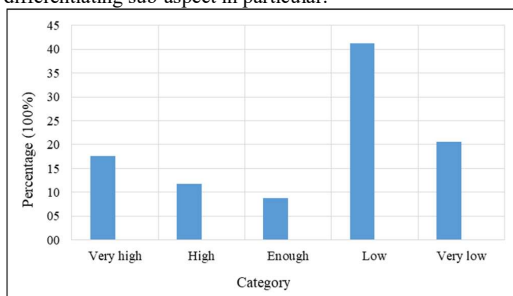


Figure 9
Percentage of Students' HOTS

Figure 9 shows the percentage of students' HOTS. It is seen that most students (41.2%) still have low HOTS; the categories consist of very low (20.6%), moderate (8.8%), high (11.8%), and very high (17.6%). The low category of students' HOTS is influenced by several factors, one of which is that the students are not used to working on HOTS questions (Tanujaya, Mumu, & Margono, 2017; Yusuf & Widyaningsih, 2019). They need to practice developing their HOTS by being exposed to HOTS-based learning sources.

Commented [p9]: After this paragraph, discussion must be written in detail. The discussion should explore the significance of the results of the study. The references contained in the introduction should not be re-written in the discussion. A comparison to the previous studies should be presented. The following components should be covered in discussion: How do your results relate to the original question or objectives outlined in the background section (what)? Do you provide interpretation scientifically for each of your results or findings presented (why)? Are your results consistent with what other investigators have reported (what else)? Or are there any differences?

CONCLUSION

Test characteristics comprised item fitness, reliability, and difficulty. Dichotomy data analysis used the Rasch Model through the Quest program. The trial result obtained the criteria of INFIT MNSQ mean and standard deviation of 1.0 and 0.0, respectively, showing that the items fit the RM1-PL. In addition, the value of item reliability based on the value of summary of item estimate arrives at 0.66; meanwhile, the person reliability under the summary of case estimate reaches 0.85, i.e., the reliability value is in the range of 0.67 - 0.80 (quite reliable). As based on the criteria of minimum and maximum INFIT MNSQ of 0.77 and 1.30, 24 question items fit the RM 1-PL model. The result of the Quest output also reveals that the average value of THRSHL and its standard deviation is $0.00 \pm 0,71$, or in the acceptance range of -2 to 2. To sum up, all 24 question items that had been tried out have fit the model with a good category, so that they can be utilized in HOTS measurement.

ACKNOWLEDGMENT

We would like to acknowledge the contribution of the Ministry of Research and Higher Education in funding this study through Inter Higher Education Institution Cooperation scheme with the contract number: 198/SP2H//AMD/LT/DRPM/2020.

REFERENCES

- Adams, R. J., & Khoo, S.-T. (1996). *Quest: the interactive test analysis system*. Camberwell, Vic.: Australian Council for Educational Research.
- Aiken, L. R. (1980). Content Validity and Reliability of Single Items or Questionnaires. *Educational and Psychological Measurement*, 40(4), 955–959.
- Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, 45(1), 131–142.
- Aldoobie, N. (2015). ADDIE Model. *American International Journal of Contemporary Research*, 5(6), 72.
- Azevedo, J. M. (2015). e-Assessment in mathematics courses with multiple-choice questions tests. *CSSEDU 2015 - 7th International Conference on Computer Supported Education, Proceedings*, 2, 260–266. <https://doi.org/10.5220/0005452702600266>
- Bogdanović, Z., Barać, D., Jovanić, B., Popović, S., & Radenković, B. (2014). Evaluation of Mobile Assessment in A Learning Management System. *British Journal of Educational Technology*, 45(2), 231–244. <https://doi.org/10.1111/bjet.12015>
- Brun, M., & Hinostroza, J. E. (2014). Learning to become a teacher in the 21st century: ICT integration in Initial Teacher Education in Chile. *Journal of Educational Technology & Society*, 17(3), 222–238. Retrieved from <http://www.jstor.org/stable/jeductechsoci.17.3.222>
- Hambleton, R. K., & Rogers, H. J. (1989). Detecting Potentially Biased Test Items: Comparison of IRT Area and Mantel-Haenszel Methods. *Applied Measurement in*

Commented [p10]: The conclusion does not contain the repetition of the results and discussion, but rather the summary of the findings as expected in the objectives or hypotheses. If necessary, at the end of the conclusion can also be written the things that will be done related to the next idea of the study. The conclusion is written in the whole paragraph, not the points per point.

Author surnames go here

15

Education, 2(4), 313–334. https://doi.org/10.1207/s15324818ame0204_4

Istiyono, E. (2017). The Analysis of Senior High School Students' Physics HOTS in Bantul District Measured using PhysReMChoTHOTS. *AIP Conference Proceedings*, 1868(August), 1–7. <https://doi.org/10.1063/1.4995184>

Istiyono, E. (2018). IT-based HOTS assessment on physics st learning as the 21 century demand at senior high schools : Expectation and reality IT-Based HOTS Assessment on Physics Learning as the 21 st Century Demand at Senior High Schools : Expectation and Reality. *AIP Conference Proceedings*, 2014(020014), 1–6.

Istiyono, E., Dwandaru, W. S. B., & Muthmainah. (2019). Developing of Bloomian HOTS Physics Test : Content and Construct Validation of The PhysTeBloHOTS Developing of Bloomian HOTS Physics Test : Content and Construct Validation of The PhysTeBloHOTS. *Journal of Physics: Conference Series*, 1397(012017), 1–9. <https://doi.org/10.1088/1742-6596/1397/1/012017>

Kowsalya, D. N., Venkat Lakshmi, H., & Suresh, K. P. (2012). Development and Validation of a Scale to assess Self-Concept in Mild Intellectually Disabled Children. *International Journal of Social Sciences & Education*, 2(4).

Krathwohl, D. R., & Anderson, L. W. (2010). Merlin C. Wittrock and the Revision of Bloom's Taxonomy. *Educational Psychologist*, 45(1), 64–65. <https://doi.org/10.1080/00461520903433562>

Lee, M. F., & Zainal, N. A. (2017). Development of needham model based E-module for electromagnetic field & wave. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 120–124). <https://doi.org/10.1109/IEEM.2017.8289863>

Limongelli, C., Sciarrone, F., & Vaste, G. (2011). Personalized e-learning in Moodle: the Moodle_LS System. *Journal of E-Learning and Knowledge Society*, 7(1), 49–58. Retrieved from <https://www.learntechlib.org/p/43340>

Martín-Blas, T., & Serrano-Fernández, A. (2009). The role of new technologies in the learning process: Moodle as a teaching tool in Physics. *Computers & Education*, 52(1), 35–44. <https://doi.org/10.1016/J.COMPEDU.2008.06.005>

Pandey, S. R., & Pandey, S. (2009). Developing a More Effective and Flexible Learning Management System (LMS) for the Academic Institutions using Moodle. *ICAL 2009 - Technology, Policy and Innovation*, 249–254.

Raykov, T., & Marcoulides, G. A. (2015). On the Relationship Between Classical Test Theory and Item Response Theory: From One to the Other and Back. *Educational and Psychological Measurement*, 76(2), 325–338. <https://doi.org/10.1177/0013164415576958>

Tanujaya, B., Mumu, J., & Margono, G. (2017). The Relationship between Higher Order Thinking Skills and Academic Performance of Student in Mathematics

Instruction. *International Education Studies*, 10(11), 78.
<https://doi.org/10.5539/ies.v10n11p78>

Yusuf, I., & Widyaningsih, S. W. (2019). HOTS profile of physics education students in STEM-based classes using PhET media. *Journal of Physics: Conference Series*, 1157(032021), 1–5.

Yusuf, I., Widyaningsih, S. W., & Sebayang, S. R. B. (2018). Implementation of E-learning based-STEM on Quantum Physics Subject to Student HOTS Ability. *Turkish Science Education*, 15(December), 67–75.

The Development of HOTS Test of Physics Based on the Modern Test Theory: Question Modeling through E-learning of Moodle LMS

The present study discussed the development of higher-order thinking skills (HOTS) test of physics based on the modern test theory. HOTS questions were designed and presented in the e-learning with the Moodle learning management system (LMS) that could be accessed online. This study employed the ADDIE model with analysis, design, development, implementation, and evaluation stages. The instrument consisted of 24 multiple choice physics questions regarding the direct current circuit topic; the questions were designed by following the aspects and sub-aspects of HOTS and had been validated by the experts of measurement, physics education, physics, and practitioners. Moreover, validity analysis was based on the V Aiken formula, in which every aspect was confirmed valid. The validated instrument was then tried out to all 34 students at the Department of Physics Education, Universitas Papua, who participated in the basic physics subject. Dichotomy data analysis used the Rasch Model (RM) 1-PL through the Quest program, and the test characteristics comprised item fitness, reliability, and difficulty. The trial result obtained the criteria of INFIT MNSQ mean and standard deviation of 1.0 and 0.0, respectively, showing that the items fitted the RM1-PL. In addition, the value of item reliability based on the value summary of the item estimate arrived at 0.66; meanwhile, the case reliability under the summary of the case estimate accounted for 0.85. The reliability value in the range of 0.67- 0.80 was categorized as quite reliable. As based on the criteria of minimum and maximum INFIT MNSQ of 0.77 and 1.30, 24 question items fitted the RM 1-PL model. The result of the Quest output also revealed that the average values of Thresholds and its standard deviation were 0.00 ± 0.71 , or in the acceptance range of -2 to 2. All in all, all 24 question items that had been tried out had fitted the model with a good category in order that they could be utilized in HOTS measurement.

Keywords: E-learning, HOTS Test, and Modern Test Theory.

INTRODUCTION

Assessment, particularly in the cognitive domain, is central to the learning process and should be carried out accurately and in compliance with the subject to be assessed or measured. Students' cognitive skills in the learning process can be categorized into lower-order thinking (LOT) and higher-order thinking (HOT). The LOTS include remembering, understanding, and applying; the HOTS, on the other hand, encompass analyzing, evaluating, and creating. HOTS are thinking skills that do not only require the remembering skill but also require other higher skills. Indicators to measure HOTS consist of analyzing (C4), evaluating (C5), and creating (C6) skills (Krathwohl & Anderson, 2010).

HOTS also refer to thinking skills when one takes new information, connects it with initial information s/he has, and finally delivers the information to achieve goals or answer questions (Istiyono, Dwandaru, & Muthmainah, 2019). This is in line with skill characteristics in the 21st century published by Partnership of 21st Century Skill stating

Commented [u1]: Wrong words choices (This study discusses about the development.....)

Commented [u2]: Expert judgement of physics

Commented [u3]: Physicist

Commented [u4]: Teacher or lecturer

Commented [u5]: Wrong word choices and unsuitable grammatical..

Commented [u6]: "then" is preposition word. It can be shorted, (e.g. the instrument has been tested to 34 students of Physics Education)

Commented [u7]: You can specifically explained it on methods section.

Commented [u8]: It conclude that all of questions are valid and reliable to use. Is that any question which not accurate or has not valid based on the RM1-PL?

Commented [u9]: The abstract is too long. Please write your abstract briefly, but consisting the main findings of your research.

Commented [u10]: You have failed to convince that your study is worthy enough. There is no state of the art of your study among other related research or the issue in general.

Commented [u11]: ??

that 21st century learners should be able to develop competitive skills, such as critical thinking, problem-solving, communication, information and communication technology (ICT) literacy, ICT, information literacy, and media literacy (Brun & Hinostroza, 2014); these focus on HOTS development.

Physics serves as part of science consisting of abstract concepts that are difficult to be directly described. Learning physics is expected to help students develop their thinking skills, in which they are not only demanded to master LOT skills, but also HOTS. Teachers are also urged to deliver learning materials to students, including the HOTS that can be improved by HOTS instrument. A previous study has reported that the majority of teachers find it challenging to develop an assessment instrument of learning outcomes, HOTS questions, in particular (Istiyono, 2018). For this reason, teacher creativity is highly required to measure students' learning outcomes. Today's development of Information and Communication Technology (ICT) can be utilized to design and habituate students to learn anywhere at any time (Yusuf, Widyaningsih, & Sebayang, 2018). Relying on ICT during the learning process is one of the significant innovations, including in the evaluation of students' learning outcomes.

The presentation of evaluation questions can be done in an integrated manner through e-learning programs, one of which is Moodle learning management system (LMS) (Azevedo, 2015; Bogdanović, Barać, Jovanić, Popović, & Radenković, 2014). The Moodle provides different types of questions, such as multiple choices, true or false, and short answers; these are stored in the taught course database and can be re-used (Limongelli, Sciarrone, & Vaste, 2011). Teachers are also able to give feedback directly to the students and give them correct answers to questions they have worked on (Pandey & Pandey, 2009). One of the advantages of an online evaluation through Moodle LMS is that students can directly figure out their assessment results.

Teachers need to prepare a good test to measure students' learning outcomes. There are two paradigms developed for students' learning outcome assessment through the applied test, i.e., classical and modern approaches. The classical paradigm being utilized is classical test theory or widely known as classical true-score theory, meanwhile, the modern paradigm is item response theory (IRT). The classical test theory is selected due to its ease in the application despite of its limitations in measuring the item difficulty level and discrimination since the calculation of both indicators is based on the test taker's total score. In contrast, the IRT frees up the dependence between the test item and test taker (a concept of parameter invariance); the test taker's response to a test item does not affect another item (a concept of local independence), and; the test item does only measure one measurement dimension (unidimensional concept) (Raykov & Marcoulides, 2015). Therefore, the application answers the needs of modern measurement to date, i.e., a comparison between test taker's skills, question development, and even adaptive test development, so that it is considered able to overcome the classical test theory limitations.

This development study is an initial study with a long-term purpose of developing general physics questions with good quality at the Department of Physics Education, Universitas Papua. As the first stage, this study focuses on students at the department

Commented [u12]: This is not proper words choices. The sentence has gramaticall error

Commented [u13]: What is it?

Commented [u14]: Too much conjunction

Commented [u15]: Considerable

Commented [u16]: Preliminary study

mentioned previously who enroll in General Physics subject taught by the researcher. This study also serves as one of the efforts to expand students' HOTS by applying a variety of HOTS-based learning sources.

METHOD

The ADDIE model, as employed by this study, refers to a general and systematic model of development study with a phased framework, allowing each element to connect with each other (Aldoobie, 2015). The stages of this model used in the development of HOTS instrument are presented in Figure 1.

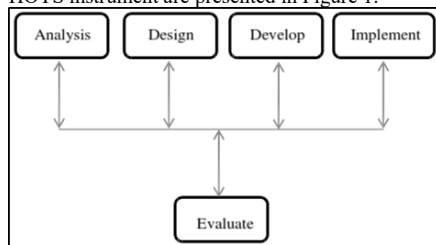


Figure 1
Stages of ADDIE Development Model in Designing Moodle LMS-based HOTS Test.

Analysis

The analysis stage is a process of needs analysis in the form of determining test objectives, identifying problems, analyzing tasks, and determining question formats to be applied. It is revealed that the problems are related to the needs of HOTS instrument design for students at the Department of Physics Education, Universitas Papua.

Design

This stage comprises the process of designing HOTS questions to be used; the design process encompasses creating a question matrix and outline that covers question distribution in every aspect and sub-aspect of HOTS.

Develop

Moreover, every single thing required in the arrangement of HOTS skill questions has been prepared in the next stage. This stage also comprises the process of making the questions regarding HOTS, as well as validating the questions that involve the experts of measurement, physics education, and practitioners. The technique of validity analysis to assess the content validity of the developed questions applies the V Aiken formula (Aiken, 1980, 1985).

$$V = \frac{\sum s}{n(c-1)} \quad (1)$$

“V” refers to the agreement index of validators in regards to item validity; “s” is the assessment score of validators subtracted by the assessment lowest score; “n” refers to

the number of validators; “c” is the number of categories that can be chosen by validators. All test items are considered valid if the value of the V Aiken index falls into the range of 0.37 - 1 (Kowsalya, Venkat Lakshmi, & Suresh, 2012). The value of V Aiken of every test item is calculated based on the assessment items of every validator. In this stage, there is also an evaluation process, i.e., revising questions by following validators’ corrections and suggestions.

Implementation

Another stage is applying HOTS questions that have been developed to 34 students in the site area who enroll in general physics subject. This number has been following the sample size for data stability in Rasch Model (RM) 1- PL, which is from 30 to 300, with the limit of INFIT t is from -2 to +2 (Bond & Fox, 2007). Question item analysis is performed based on the raw score of the students by employing the Quest program.

Evaluation

Evaluation is a process of finding out whether or not the developed questions of HOTS have met the expectation. The evaluation stage is carried out in every stage and called a formative evaluation intended for revisions (Lee & Zainal, 2017). For instance, in the design stage, the expert’s review is necessary to provide input towards the design. Further, the evaluation stage is undertaken after analyzing empirical questions mathematically by using the Quest software program by referring to the Rasch model. The Quest program is able to do the Rasch measurement, i.e., a comprehensive empirical test of question items. There are three parameters being measured mathematically based on the empirical test of question items.

1. The first parameter is item fitness with the Rasch model by following the value of INFIT MNSQ or INFIT t of the item. The expected values of the unweighted mean square (Outfit MNSQ) in the Quest program and weighted mean square are 1; the variance is 0. On the contrary, the expected value of Mean INFIT t is equal to 0, with the variance equal to 1 (Adams & Khoo, 1996). The provision of INFIT MNSQ for the Rasch Model is shown in Table 1 and Table 2.

Tabel 1

Criteria of Question Item Fitness with the Rasch Model

MNSQ INFIT Value	Criteria
>1,33	Does Not Fit the Rasch Model
0,77 s.d. 1,33	Fits the Rasch Model
<0,77	Does Not Fit the Rasch Model

Tabel 2

The Provision of Outfit t for the Rasch Model.

t OUTFIT Value	Criteria
OUTFIT $t \leq 2,00$	Fits the Rasch Model
OUTFIT $t \geq 2,00$	Does Not Fit the Rasch Model

2. The second parameter is reliability. The analysis result of the Quest program also reveals the item and case reliability. The reliability value based on the item estimate is

also called as sample reliability; the higher the value, the more the items that fit the tested model. Whereas, the lower the value, the less the items that fit the tested model, so that it does not give the expected information. The reliability category is provided in Table 3 (Istiyono, 2017).

Tabel 3
Interpretation of Reliability Value

Reliability Value	Criteria
$> 0,94$	Excellent
$0,91 - 0,94$	Very Good
$0,81 - 0,90$	Good
$0,67 - 0,80$	Acceptable
$< 0,67$	Poor

3. The third parameter is item difficulty index and respondents' skills presented as difficulty index in the Quest output. Thresholds (THRSHL) show the item difficulty index in the logit scale along with its standard deviation (Hambleton & Rogers, 1989). The provision of the THRSHL value is given in Table 4.

Tabel 4
Criteria of THRSHL Value to Categorize Item Difficulty Level

THRSHL Value	Criteria
$b > 2,00$	Very Difficult
$1,00 < b \leq 2,00$	Difficult
$-1,00 < b \leq 1,00$	Medium
$-1,00 > b \geq 2,00$	Easy
$b < -2,00$	Very Easy

Respondents' skills are shown by the value of the estimate error, in which the criteria of the estimate value of respondents' skills are presented in Table 5.

Tabel 5
Criteria of Estimate Value to Categorize Respondents' Skills

THRSHL Value	Criteria
$b > 2,00$	Very Difficult
$1,00 < b \leq 2,00$	Difficult
$-1,00 < b \leq 1,00$	Medium
$-1,00 > b \geq 2,00$	Easy
$b < -2,00$	Very Easy

The evaluation stage also includes the process of analyzing the HOTS of students on the whole. The level of HOTS is categorized based on the ideal mean and standard deviation. This is applied with the assumption that students' HOTS of physics are normally distributed. The ideal mean (I_m) and ideal standard deviation (I_{sd}) are based on the highest and lowest score of research variables. Table 6 shows the criteria of students' HOTS of physics.

Tabel 6
Criteria of Students' HOTS of Physics

Interval	Criteria
$Im + 1,5 Isb < \theta$	Very high
$Im + 0,5 Isb < \theta \leq Im + 1,5 Isb$	High
$Im - 0,5 Isb < \theta \leq Im + 0,5 Isb$	Medium
$Im - 1,5 Isb < \theta \leq Im - 0,5 Isb$	Low
$0 < Im - 1,5 Isb$	Very Low

Meaning:

Im : ideal mean

Isb : ideal standard deviation

X_{mak} : highest score

X_{min} : lowest score

RESULTS AND DISCUSSION

ADDIE development model can be used for different product developments in education, and one of which is the development of HOT skill questions. This model is simple and systematically structured in its implementation stages. The following is the description of each stage result.

Analysis

Needs analysis is the first stage being done by observation and interview to gather any information needed in the process of physics learning at the Department of Physics Education, Universitas Papua. The researcher's experience indicates that lecturers have applied HOTS learning in the classroom. However, a test to measure students' HOTS has not been conducted. The arrangement of HOTS instrument is required to train and develop students' HOTS. Accordingly, to facilitate the students in accessing other learning sources, this study designs HOT skill questions in an online system through an e-learning program using the Moodle LMS.

Design

In the design stage, the test instrument is designed based on the analysis result in the first stage. Test instrument design in this stage is in the form of question matrix and outline which are adjusted to students' needs and characteristics, and learning sources. The test is a multiple-choice test, in which 24 questions are adjusted to the formulation of a HOTS test that has been created in the test matrix and outline. The question matrix is provided in Table 7.

Tabel 7

The Question Matrix

Aspect	Sub Aspect	Theory		
		Electric current, Ohm's law, and electrical power	Series and parallel circuits of resistor and capacitor	Electric Force, Kirchoff's law, and RC circuit.
Analyze	Differentiating	8	12	21
	Organizing	3	15	20
	Attributing	2	9	23
Evaluate	Checking	4	11	22

Commented [u17]: You should explained how your instruments has improved students High Order Thinking Skills. Not just describe what you have done to develop the instrument.

	Critiquing	1	16	18
Create	Generating	5	13	19
	Planning	7	14	17
	Producing	6	10	24

Develop

The development of HOTS questions is based on the question matrix and outline that have been designed. Further, the questions are made online through e-learning by utilizing the Moodle LMS. Figure 2 shows all question items in the e-learning program.

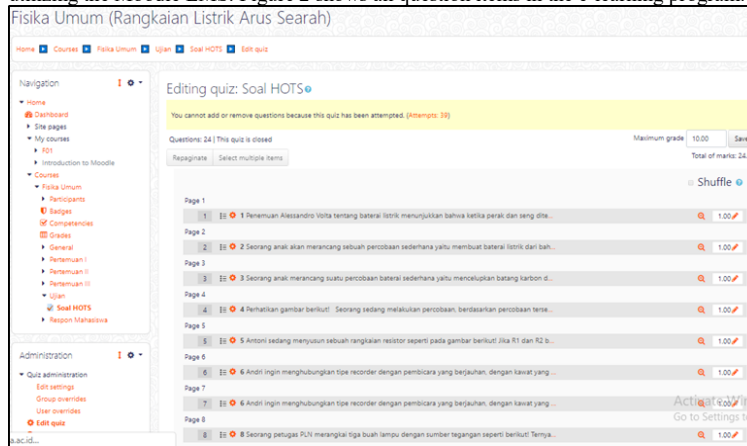


Figure 2 Shows All Question Items in the E-Learning Program

Moodle LMS program presents an interesting display and is easy to access by users (Martín-Blas & Serrano-Fernández, 2009). The questions are displayed interactively, and students can randomly work on the questions. Moodle LMS can present questions with a picture or other content to make it easier for teachers to design the questions as expected. Figure 3 shows one of the HOTS questions displayed on the e-learning through the Moodle LMS.

Figure 3
Shows of the HOTS Questions Displayed on the E-learning Through the Moodle LMS

The development stage aims to produce a HOTS test instrument that has been validated by experts and practitioners. Product validation is a process of assessing the designed product, or in this case, the test instrument of HOTS in general physics subject in the site area. Product validation is carried out by involving seven validators, i.e., experts of measurement, physics education, physics, and practitioners. The validity test of the instrument includes material, construction, and language. The analysis result of question validity that is assessed by validators obtains the value of V Aiken in the range of 0.76 - 1.00, showing a valid result. The questions validated by experts and practitioners are then revised based on provided corrections and suggestions.

Implementation

The implementation stage in this study is the product trial, in which HOTS questions are tried out to 34 students in the research site. The students work on these questions via online through e-learning by using their own Moodle account upon the completion of all learning stages. Results of the students' learning can be accessed after this process.

Evaluation

Before conducting the estimate analysis of respondents' skills and item difficulty level, the analysis of item fitness is performed by using parameters of INFIT and OUTFIT for mean square and t . The determination of the item fitness with the model is based on the value of INFIT MNSQ and the standard deviation or Infit t (Adams & Khoo, 1996). The fitness of each case is also based on the value of INFIT MNSQ or INFIT t of the item. Table 8 provides the testing result through the Quest program to obtain the values of item estimate and case estimate in the HOTS questions trial.

Tabel 8
Values of Item Estimate and Case Estimate in the HOTS Questions Trial

No	Measurement	Estimates for Items	Estimates for Testi
1.	Average values and standard deviations	0,00 ± 0,57	0,01 ± 1,24
2.	Reliability Estimates	0,66	0,85
3.	The mean value and standard deviation of INFIT MNSQ	1,00 ± 0,14	0,99 ± 0,15
4.	The mean value and standard deviation of OUTFIT MNSQ	1,09 ± 0,52	1,09 ± 0,52
5.	The mean value and standard deviation of INFIT t	-0,03 ± 0,81	0,00 ± 0,72
6.	The mean value and standard deviation of OUTFIT t	0,21 ± 0,91	0,17 ± 0,81

The analysis result reveals that the INFIT MNSQ arrives at the range of 0.86 - 1.14, and INFIT t is -0.28 - 0.72. This result signifies that all 24 questions fit the model as they reach the range of INFIT MNSQ value from 0.77 to 1.30 and use INFIT t with the limit of -2.0 - 2.0 [16]. In addition to testing the fitness, the output of the Quest program also presents the reliability estimate of the test instrument. Table 8 provides the value of item reliability based on the value of summary of item estimate, which is 0.66. On the other hand, the value of person reliability, as based on the summary of case estimate, gets 0.85. These results are in line with the Rasch model, in which the reliability value falls under the range of 0.67 - 0.80 (quite reliable). On that ground, the instrument can be used to measure students' HOTS in the General Physics subject.

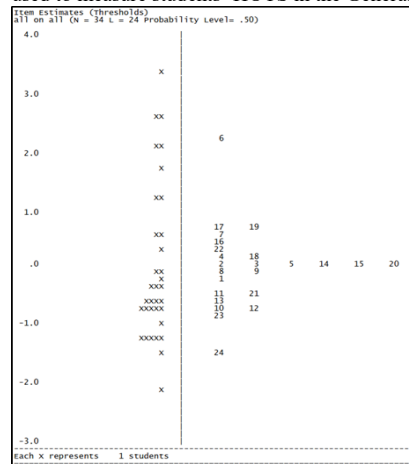


Figure 4
Distribution of Item Difficulty Level and Respondents' Skills

Figure 4 presents the distribution of the respondents according to the difficulty level in the logit scale from -4.0 to +4.0. This map displays the item difficulty level compared to the respondents' skills. Case and item difficulty levels in the Rasch model are expressed in one line in the form of abscissa in the graph with logg-odd unit. The graph of respondents' skills shows a normal curve, meaning that there are only a few respondents with low and high skills; and a lot of respondents with moderate skills. The level of item difficulty of threshold reveals that item 6 is the most difficult question, and item 24 is the easiest one.

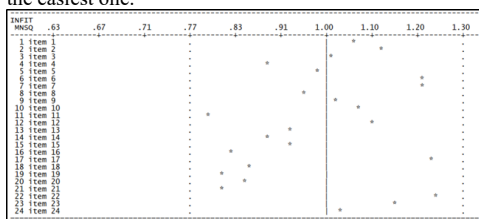


Figure 5
Distribution of INFIT MNSQ Values of Each Question Item of HOTS

Question items that fit the Rasch model are in the range of 0.77 - 1.33. Figure 5 shows that all 24 question items are in the line, implying that they fit the Rasch model.

Author surnames go here

11

Item Estimates (Thresholds) In input Order							
all on all (N = 34 L = 24 Probability Level= .50)							
ITEM NAME	SCORE	MAXSCR	THRSH	INFT	OUTFT	INFT	OUTFT
			1	MNSQ	MNSQ	t	t
1 item 1	18	34	-.26 .39	1.06	1.15	.4	.5
2 item 2	16	34	.04 .40	1.12	1.17	.7	.6
3 item 3	16	34	.04 .40	1.01	.91	.1	-.2
4 item 4	15	34	.19 .40	.88	.93	-.6	-.1
5 item 5	16	34	.04 .40	.98	.89	.0	-.2
6 item 6	5	34	2.27 .37	1.21	2.16	.7	1.4
7 item 7	13	34	.52 .42	1.21	1.27	1.0	.9
8 item 8	17	34	-.11 .40	.96	1.00	-.2	.1
9 item 9	17	34	-.11 .40	1.02	.91	.2	-.2
10 item 10	21	34	-.70 .39	1.07	1.16	.6	.5
11 item 11	19	34	-.41 .39	.79	.66	-1.6	-.9
12 item 12	21	34	-.70 .39	1.10	1.14	.8	.5
13 item 13	20	34	-.55 .39	.93	1.09	-.5	.4
14 item 14	16	34	.04 .40	.88	.78	-.7	-.6
15 item 15	16	34	.04 .40	.93	.82	-.4	-.5
16 item 16	13	33	.47 .42	.82	.69	-.8	-.9
17 item 17	12	34	.69 .43	1.23	1.16	1.0	.6
18 item 18	15	34	.19 .40	.86	.73	-.8	-.8
19 item 19	12	34	.69 .43	.81	.71	-.8	-.8
20 item 20	16	34	.04 .40	.85	.75	-.9	-.7
21 item 21	19	34	-.41 .39	.81	.68	-1.4	-.8
22 item 22	14	34	.35 .41	1.24	1.23	1.2	.8
23 item 23	22	34	-.85 .40	1.15	3.04	1.1	3.1
24 item 24	26	34	-1.50 .43	1.03	1.02	.2	.2
Mean			.00	1.00	1.09	.0	.1
SD			.71	.14	.52	.8	.9

Figure 6
Item Estimates from HOTS Questions

Figure 6 presents the Item Estimate of HOT skill questions based on the trial result. In this figure, there is SCORE-MAXSCR successively showing the score of the respondents who answer correctly, and the number of total respondents. Item 24 is the most correctly-answered, in which 26 out of 34 respondents are able to work on this item. Figure 6 also provides the value of THRSHL that shows the item difficulty index in the logit scale along with its standard deviation. Item 6 has THRSHL or difficulty index of 2.27 that is greater than 2.0, or in other words, this item is very difficult since

only five students can give a correct answer. The average value of THRSHL and its standard deviation accounts for 0.00 ± 0.71 and falls into the range of -2 - 2 (Hambleton & Rogers, 1989). The average value of INFIT MNSQ is 1.00 ± 0.14 and falls under the acceptance range of 0.77 - 1.33; the average value of OUTFIT t arrives at 0.10 ± 0.90 and falls into the acceptance range of ≤ 2.00 . All of these results indicate that all question items that have been developed can be employed to measure students' HOTS.

Case Estimates in input order
all on all (N = 34 L = 24 Probability Level= .50)

NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFIT MNSQ	OUTFIT MNSQ	INFIT t	OUTFIT t
1 01	12	24	-.02	.43	1.06	1.02	.58	.18
2 02	6	24	-1.21	.49	1.17	1.09	.75	.36
3 03	8	24	-.77	.45	.98	1.01	-.06	.13
4 04	8	24	-.77	.45	.83	.81	-1.04	-.48
5 05	8	24	-.77	.45	.89	.83	-.59	-.41
6 06	6	24	-1.21	.49	.79	.70	-.84	-.67
7 07	10	24	-.38	.43	.99	.95	-.01	-.09
8 08	6	24	-1.21	.49	1.07	2.30	.36	2.44
9 09	3	24	-2.10	.63	.98	.85	.11	.00
10 10	9	24	-.57	.44	.88	.83	-.80	-.48
11 11	22	24	2.61	.77	.73	.46	-.30	-.56
12 12	5	24	-1.46	.52	.89	.85	-.29	-.18
13 13	20	24	1.75	.57	1.21	1.45	.64	.93
14 14	11	24	-.20	.43	.86	.83	-1.33	-.59
15 15	21	24	2.12	.64	1.18	1.05	.52	.29
16 16	9	24	-.57	.44	1.08	1.06	.59	.28
17 17	7	24	-.98	.47	1.29	2.20	1.38	2.55
18 18	6	24	-1.21	.49	1.23	1.28	.96	.75
19 19	14	24	.35	.43	.92	.87	-.56	-.40
20 20	15	24	.54	.44	.97	1.09	-.13	.40
21 21	18	24	1.19	.49	.94	.86	-.16	-.25
22 22	21	24	2.12	.64	.93	1.23	-.01	.53
23 23	9	24	-.57	.44	1.07	1.01	.54	.15
24 24	8	24	-.77	.45	1.01	.95	.13	-.03
25 25	10	24	-.38	.43	.87	.82	-1.06	-.57
26 26	15	24	.54	.44	1.05	1.22	.36	.80
27 27	6	24	-1.21	.49	.82	.74	-.69	-.56
28 28	22	24	2.61	.77	.73	.46	-.30	-.56
29 29	12	24	-.02	.43	.92	.88	-.73	-.39
30 30	9	24	-.57	.44	.90	.90	-.64	-.23
31 31	23	24	3.40	1.05	1.18	3.14	.49	1.53
32 32	18	24	1.19	.49	.85	.75	-.53	-.58
33 33	10	23	-.32	.44	1.11	1.11	.97	.45
34 34	8	24	-.77	.45	1.29	1.34	1.61	1.03
Mean			.01		.99	1.09	.00	.17
SD			1.35		.15	.52	.72	.81

Figure 7
Case Estimates from Every Student

Figure 7 serves as the case estimate or the skill level of each student. Information obtained from the case estimate is that the SCORE-MAXSCR shows the score of each respondent from the maximum score sequentially. Respondent 31 answers the most questions (23 out of 24 questions) correctly compared to other respondents. The average estimate value and its standard deviation gets 0.01 ± 1.35 and falls under a moderate category. The analysis result of the case estimate reveals that students' skills are in the moderate category.

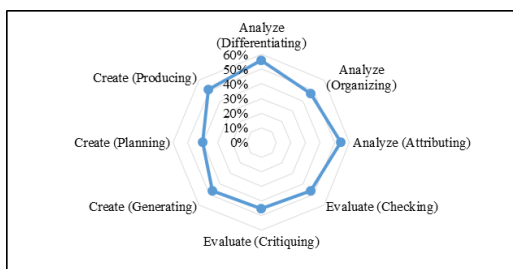


Figure 8
Distribution of Student Answer Percentage HOTS

Figure 8 gives the percentage of students' answers based on the aspects and sub-aspects of HOTS. The analysis result brings out the fact that students tend to find it difficult to answer questions regarding the creating aspect, especially the planning sub-aspect. Creating is the highest level HOTS in Bloom's taxonomy, which therefore, students need to practice developing their creating skills. This figure also signifies that the majority of the students find it easy to answer HOTS questions related to the analysis aspect, differentiating sub-aspect in particular.

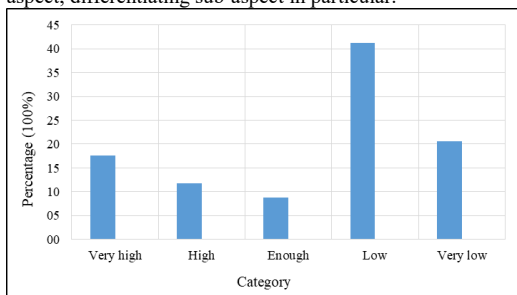


Figure 9
Percentage of Students' HOTS

Figure 9 shows the percentage of students' HOTS. It is seen that most students (41.2%) still have low HOTS; the categories consist of very low (20.6%), moderate (8.8%), high (11.8%), and very high (17.6%). The low category of students' HOTS is influenced by several factors, one of which is that the students are not used to working on HOTS questions (Tanujaya, Mumu, & Margono, 2017; Yusuf & Widyaningsih, 2019). They need to practice developing their HOTS by being exposed to HOTS-based learning sources.

CONCLUSION

Test characteristics comprised item fitness, reliability, and difficulty. Dichotomy data analysis used the Rasch Model through the Quest program. The trial result obtained the criteria of INFIT MNSQ mean and standard deviation of 1.0 and 0.0, respectively, showing that the items fit the RM1-PL. In addition, the value of item reliability based on the value of summary of item estimate arrives at 0.66; meanwhile, the person reliability under the summary of case estimate reaches 0.85, i.e., the reliability value is in the range of 0.67 - 0.80 (quite reliable). As based on the criteria of minimum and maximum INFIT MNSQ of 0.77 and 1.30, 24 question items fit the RM 1-PL model. The result of the Quest output also reveals that the average value of THRSHL and its standard deviation is 0.00 ± 0.71 , or in the acceptance range of -2 to 2. To sum up, all 24 question items that had been tried out have fit the model with a good category, so that they can be utilized in HOTS measurement.

ACKNOWLEDGMENT

We would like to acknowledge the contribution of the Ministry of Research and Higher Education in funding this study through Inter Higher Education Institution Cooperation scheme with the contract number: 198/SP2H//AMD/LT/DRPM/2020.

REFERENCES

- Adams, R. J., & Khoo, S.-T. (1996). *Quest: the interactive test analysis system*. Camberwell, Vic.: Australian Council for Educational Research.
- Aiken, L. R. (1980). Content Validity and Reliability of Single Items or Questionnaires. *Educational and Psychological Measurement*, 40(4), 955–959.
- Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, 45(1), 131–142.
- Aldoobie, N. (2015). ADDIE Model. *American International Journal of Contemporary Research*, 5(6), 72.
- Azevedo, J. M. (2015). e-Assessment in mathematics courses with multiple-choice questions tests. *CSEDU 2015 - 7th International Conference on Computer Supported Education, Proceedings*, 2, 260–266. <https://doi.org/10.5220/0005452702600266>
- Bogdanović, Z., Barać, D., Jovanić, B., Popović, S., & Radenković, B. (2014). Evaluation of Mobile Assessment in A Learning Management System. *British Journal of Educational Technology*, 45(2), 231–244. <https://doi.org/10.1111/bjet.12015>
- Brun, M., & Hinostroza, J. E. (2014). Learning to become a teacher in the 21st century: ICT integration in Initial Teacher Education in Chile. *Journal of Educational Technology & Society*, 17(3), 222–238. Retrieved from <http://www.jstor.org/stable/jeductechsoci.17.3.222>
- Hambleton, R. K., & Rogers, H. J. (1989). Detecting Potentially Biased Test Items: Comparison of IRT Area and Mantel-Haenszel Methods. *Applied Measurement in*

Author surnames go here

15

Education, 2(4), 313–334. https://doi.org/10.1207/s15324818ame0204_4

Istiyono, E. (2017). The Analysis of Senior High School Students' Physics HOTS in Bantul District Measured using PhysReMChoTHOTS. *AIP Conference Proceedings*, 1868(August), 1–7. <https://doi.org/10.1063/1.4995184>

Istiyono, E. (2018). IT-based HOTS assessment on physics st learning as the 21 century demand at senior high schools : Expectation and reality IT-Based HOTS Assessment on Physics Learning as the 21 st Century Demand at Senior High Schools : Expectation and Reality. *AIP Conference Proceedings*, 2014(020014), 1–6.

Istiyono, E., Dwardaru, W. S. B., & Muthmainah. (2019). Developing of Bloomian HOTS Physics Test : Content and Construct Validation of The PhysTeBloHOTS Developing of Bloomian HOTS Physics Test : Content and Construct Validation of The PhysTeBloHOTS. *Journal of Physics: Conference Series*, 1397(012017), 1–9. <https://doi.org/10.1088/1742-6596/1397/1/012017>

Kowsalya, D. N., Venkat Lakshmi, H., & Suresh, K. P. (2012). Development and Validation of a Scale to assess Self-Concept in Mild Intellectually Disabled Children. *International Journal of Social Sciences & Education*, 2(4).

Krathwohl, D. R., & Anderson, L. W. (2010). Merlin C. Wittrock and the Revision of Bloom's Taxonomy. *Educational Psychologist*, 45(1), 64–65. <https://doi.org/10.1080/00461520903433562>

Lee, M. F., & Zainal, N. A. (2017). Development of needham model based E-module for electromagnetic field & wave. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 120–124). <https://doi.org/10.1109/IEEM.2017.8289863>

Limongelli, C., Sciarrone, F., & Vaste, G. (2011). Personalized e-learning in Moodle: the Moodle_LS System. *Journal of E-Learning and Knowledge Society*, 7(1), 49–58. Retrieved from <https://www.learntechlib.org/p/43340>

Martín-Blas, T., & Serrano-Fernández, A. (2009). The role of new technologies in the learning process: Moodle as a teaching tool in Physics. *Computers & Education*, 52(1), 35–44. <https://doi.org/10.1016/J.COMPEDU.2008.06.005>

Pandey, S. R., & Pandey, S. (2009). Developing a More Effective and Flexible Learning Management System (LMS) for the Academic Institutions using Moodle. *ICAL 2009 - Technology, Policy and Innovation*, 249–254.

Raykov, T., & Marcoulides, G. A. (2015). On the Relationship Between Classical Test Theory and Item Response Theory: From One to the Other and Back. *Educational and Psychological Measurement*, 76(2), 325–338. <https://doi.org/10.1177/0013164415576958>

Tanujaya, B., Mumu, J., & Margono, G. (2017). The Relationship between Higher Order Thinking Skills and Academic Performance of Student in Mathematics Instruction. *International Education Studies*, 10(11), 78.

<https://doi.org/10.5539/ies.v10n11p78>

Yusuf, I., & Widyaningsih, S. W. (2019). HOTS profile of physics education students in STEM-based classes using PhET media. *Journal of Physics: Conference Series*, 1157(032021), 1–5.

Yusuf, I., Widyaningsih, S. W., & Sebayang, S. R. B. (2018). Implementation of E-learning based-STEM on Quantum Physics Subject to Student HOTS Ability. *Turkish Science Education*, 15(December), 67–75.

The Development of the HOTS Test of Physics Based on Modern Test Theory: Question Modeling through E-learning of Moodle LMS

Sri Wahyu Widyaningsih

Assist. Prof., Faculty of Teacher Training and Education, Universitas Papua, Indonesia, s.widyaningsih@unipa.ac.id

Irfan Yusuf

Assist. Prof., Faculty of Teacher Training and Education, Universitas Papua, Indonesia, i.yusuf@unipa.ac.id

Zuhdan Kun Prasetyo

Prof., Faculty of Science, Universitas Negeri Yogyakarta, Indonesia, zuhdan@uny.ac.id

Edi Istiyono

Prof., Graduate School, Universitas Negeri Yogyakarta, Indonesia, edi_istiyono@uny.ac.id

The present study discussed the development of the HOTS test of physics based on modern test theory. HOTS questions were designed and presented in the e-learning. Further, this research employed the ADDIE model with analysis, design, development, implementation, and evaluation stages. The instrument consisted of 24 multiple-choice physics questions; the questions were designed by following the aspects and sub-aspects of HOTS and validated by the assessment of physics experts, physicists, and lecturers. Moreover, the validity analysis was based on Aiken's V formula, in which every aspect was confirmed to be valid. The instrument had been tested on 34 students of the Physics Education Department, Universitas Papua. Dichotomy data analysis used the Rasch Model (RM) 1-PL through the Quest program, and the test characteristics comprised item fitness, reliability, and difficulty. The trial result obtained the criteria of INFIT MNSQ mean and standard deviation of 1.0 and 0.0, respectively, indicating that the items fitted the RM1-PL. In addition, the value of item reliability based on the item estimate summary arrived at 0.66; meanwhile, the case reliability under the summary of the case estimate accounted for 0.85. The reliability value in the range of 0.67- 0.80 was categorized as quite reliable. Drawing upon the criteria of minimum and maximum INFIT MNSQ of 0.77 and 1.30, 24 question items fitted the RM 1-PL model. The Quest output result also suggested that the average values of Thresholds and its standard deviation were 0.00 ± 0.71 , or in the acceptance range of -2 to 2. Overall, all 24 question items that had been tested have fitted the model with a good category. They can be used in the HOTS measurement and can increase students' HOTS.

Keywords: E-learning, HOTS Test, and Modern Test Theory.

INTRODUCTION

Assessment, especially in the cognitive domain, is central to the learning process and should be carried out accurately and in compliance with the subject to be assessed or measured. Students' cognitive skills in the learning process can be categorized into Lower-Order Thinking Skills (LOTS) and Higher-Order Thinking Skills (HOTS). The LOTS includes remembering, understanding, and applying; the HOTS, on the other hand, consists of analyzing, evaluating, and creating. HOTS is thinking skills that require not only the remembering skill but also other higher skills. Indicators to measure HOTS encompass analyzing (C4), evaluating (C5), and creating (C6) skills (Krathwohl & Anderson, 2010).

HOTS also refers to thinking skills when one takes new information, connects it with initial information s/he has, and finally delivers the information to achieve goals or answer questions (Istiyono, Dwandaru, & Muthmainah, 2019). This is in line with skill characteristics in the 21st century published by Partnership of 21st Century Skill stating that 21st-century learners should be able to develop competitive skills, such as critical thinking, problem-solving, communication, information and communication technology (ICT) literacy, ICT, information literacy, and media literacy (Brun & Hinostroza, 2014); these focus on HOTS development.

Physics serves as part of science, comprising abstract concepts that are difficult to be directly described. Learning physics is expected to help students develop their thinking skills, in which they are not only demanded to master LOTS, but also HOTS. Teachers are also urged to deliver learning materials to students, including the HOTS, that can be improved by the HOTS instrument. A previous study has reported that the majority of teachers find it challenging to formulate an assessment instrument of learning outcomes, HOTS questions, in particular (Istiyono, 2018). For this reason, teachers' creativity is highly required to measure student learning outcomes. Today's development of ICT can be utilized to design and habituate students to learn anywhere at any time (Yusuf, Widyaningsih, & Sebayang, 2018). Relying on ICT during the learning process is one of the significant innovations, including the evaluation of student learning outcomes.

Evaluation questions can be posed in an integrated manner through e-learning systems, such as Moodle Learning Management System (LMS) (Azevedo, 2015; Bogdanović, Barać, Jovanić, Popović, & Radenković, 2014). The Moodle provides different types of questions, namely multiple choices, true or false, and short answers; these are stored in the taught course database and can be reapplied (Limongelli, Sciarrone, & Vaste, 2011). Teachers are also able to offer feedback directly to the students and give them correct answers to questions they have worked on (Pandey &

Pandey, 2009). One of the advantages of an online evaluation through Moodle LMS is that students can figure out their assessment results right away.

Teachers need to prepare a good test to measure student learning outcomes. There are two paradigms developed to assess student learning outcomes through the used test, i.e., classical and modern approaches. The classical paradigm being utilized is classical test theory or widely known as classical true-score theory; meanwhile, the modern paradigm is item response theory (IRT). The classical test theory is selected due to its ease in the application despite its limitations in measuring the item difficulty level and discrimination since both indicators' calculation is based on the test taker's total score. In contrast, the IRT frees up the dependence between the test item and the test taker (a concept of parameter invariance); the test taker's response to a test item does not affect another item (a concept of local independence), and; the test item does only measure one measurement dimension (Raykov & Marcoulides, 2015). Therefore, the application answers the needs of modern measurement to date, i.e., comparing test taker's skills, question development, and even adaptive test development. It is considered able to overcome the limitations of the classical test theory.

On account of the simplicity of the analysis, most teachers have analyzed assessment tools using classical analysis techniques. The use of classical analytical techniques features some limitations, including the difficulty of defining individual learners' skills. The calculated error of measurement does not include persons but groups together. This is because each test taker's response to the questions cannot be clarified by classical test theory. Efforts are thereby required to free the measuring tool from attachment to the sample (sample-free) employing the IRT.

This is a preliminary study with a long-term purpose of developing general physics questions with good quality at the Department of Physics Education, Universitas Papua. As the first stage, this study focuses on students at the department mentioned previously who enroll in General Physics subject taught by the researcher. This study also serves as one of the efforts to expand students' HOTS by applying a variety of HOTS-based learning sources. This research aims to develop HOTS physics questions based on IRT designed and presented with LMS Moodle on e-learning, which can be accessed online.

METHOD

As employed by this study, the ADDIE model refers to a general and systematic model of development study with a phased framework, allowing each element to connect (Aldoobie, 2015). The stages of this model used in the development of the HOTS instrument are presented in Figure 1.

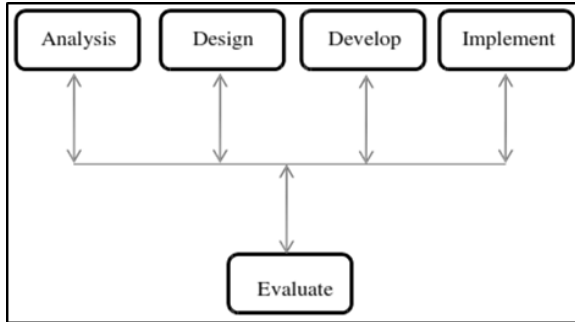


Figure 1
Stages of ADDIE Development Model in Designing Moodle LMS-based HOTS Test

Analysis

The analysis stage was a process of needs analysis to determine test objectives, identify problems, analyze tasks, and determine question formats to be applied. It was shown that the problems were related to the needs of HOTS instrument design for students at the Department of Physics Education, Universitas Papua.

Design

This stage comprised the process of designing HOTS questions to be used; the design process encompassed creating a question matrix and outline that covered question distribution in every aspect and sub-aspect of HOTS.

Develop

Every single thing required in the arrangement of HOTS questions has been prepared in the next stage. This stage also covered the process of making the questions regarding HOTS, as well as validating the questions that involved the experts of measurement, physics education, and practitioners. The validity analysis technique to assess the content validity of the developed questions relied on the Aiken's V formula (Aiken, 1980, 1985).

$$V = \frac{\sum s}{n(c-1)} \quad (1)$$

"V" refers to the agreement index of validators in regards to item validity; "s" is the assessment score of validators subtracted by the assessment lowest score; "n" refers to the number of validators; "c" is the number of categories that can be chosen by validators. All test items are considered valid if the value of the Aiken's V index falls under the range of 0.37 to 1.00 (Kowsalya, Venkat Lakshmi, & Suresh, 2012). The value of Aiken's V of every test item was calculated based on the assessment items of

every validator. In this stage, there was also an evaluation process, i.e., revising questions by following validators' corrections and suggestions.

Implementation

Another stage was applying HOTS questions that had been developed to 34 students in the site area who enrolled in general physics subject. This number followed the sample size for data stability in Rasch Model (RM) 1- PL, which is from 30 to 300, with the limit of INFIT t is from -2 to +2 (Bond, Yan, & Heene, 2020). Question item analysis was performed based on the raw score of the students by employing the Quest program.

Evaluation

The evaluation was a process of finding out whether HOTS's developed questions had met the expectation. The evaluation stage is carried out in every stage and is called a formative evaluation intended for revisions (Lee & Zainal, 2017). For instance, in the design stage, the expert's review is necessary to provide input towards the design. Besides, the evaluation stage was undertaken after analyzing empirical questions mathematically by using the Quest software program by referring to the Rasch model. The Quest program can do the Rasch measurement, i.e., a comprehensive empirical test of question items. There were three parameters being measured mathematically based on the empirical test of question items, as follows.

1. The first parameter is item fitness with the Rasch model by following the value of INFIT MNSQ or INFIT t of the item. The expected values of the unweighted mean square (Outfit MNSQ) in the Quest program and weighted mean square are 1; the variance is 0. On the contrary, the expected value of Mean INFIT t is equal to 0, with the variance equal to 1 (Adams & Khoo, 1996). The provision of INFIT MNSQ for the Rasch Model is presented in Table 1 and Table 2 below.

Table 1

Criteria of Question Item Fitness with the Rasch Model

MNSQ INFIT Value	Criteria
>1.33	Does Not Fit the Rasch Model
0.77 to 1.33	Fits the Rasch Model
<0.77	Does Not Fit the Rasch Model

Table 2

The Provision of Outfit t for the Rasch Model.

t OUTFIT Value	Criteria
OUTFIT t ≤ 2.00	Fits the Rasch Model
OUTFIT t ≥ 2.00	Does Not Fit the Rasch Model

2. The second parameter is reliability. The analysis result of the Quest program also showed the item and case reliability. The reliability value based on the item estimate is also called sample reliability; the higher the value, the more the items that fit the tested model. Whereas, the lower the value, the less the items that fit the tested model, so that it does not give the expected information. The reliability category is provided in the following table (Istiyono, 2017).

Table 3

Interpretation of Reliability Value

Reliability Value	Criteria
> 0.94	Excellent
0.91 – 0.94	Very Good
0.81 – 0.90	Good
0.67 – 0.80	Fair
< 0.67	Poor

3. The third parameter is the item difficulty index and respondents' skills presented as difficulty index in the Quest output. Thresholds (THRSHL) show the item difficulty index in the logit scale along with its standard deviation (Hambleton & Rogers, 1989). The provision of the THRSHL value is in Table 4.

Table 4

Criteria of THRSHL Value to Categorize Item Difficulty Level

THRSHL Value	Criteria
$b > 2.00$	Very Difficult
$1.00 < b \leq 2.00$	Difficult
$-1.00 < b \leq 1.00$	Medium
$-1.00 > b \geq 2.00$	Easy
$b < -2.00$	Very Easy

Respondents' skills were shown by the value of the estimate error, in which the criteria of the estimate value of respondents' skills are given in Table 5.

Table 5

Criteria of Estimate Value to Categorize Respondents' Skills

THRSHL Value	Criteria
$b > 2.00$	Very Difficult
$1.00 < b \leq 2.00$	Difficult
$-1,00 < b \leq 1.00$	Medium
$-1.00 > b \geq 2.00$	Easy
$b < -2.00$	Very Easy

The evaluation stage also included the process of analyzing the HOTS of students on the whole. The level of HOTS is categorized based on the ideal mean and standard deviation. This was applied with the assumption that students' HOTS of physics were normally distributed. The ideal mean (I_m) and ideal standard deviation (I_{sd}) are based on the highest and lowest score of research variables. Table 6 shows the criteria of students' HOTS of physics.

Table 6
Criteria of Students' HOTS of Physics

Interval	Criteria
$Im + 1.5 Isb < \theta$	Very high
$Im + 0.5 Isb < \theta \leq Im + 1.5 Isb$	High
$Im - 0.5 Isb < \theta \leq Im + 0.5 Isb$	Moderate
$Im - 1.5 Isb < \theta \leq Im - 0.5 Isb$	Low
$0 < Im - 1.5 Isb$	Very Low

Meaning:

Im : ideal mean

Isb : ideal standard deviation

X_{mak} : highest score

X_{min} : lowest score

RESULTS

The ADDIE development model can be used for different product developments in education, and one of which is the development of HOTS questions. This model is simple and systematically structured in its implementation stages. The following is a description of each stage result.

Analysis

A needs analysis was the first stage being done by observation and interview to gather any information required in physics learning at the Department of Physics Education, Universitas Papua. The researchers' experience indicated that the lecturers had applied HOTS learning in the classroom. However, a test to measure students' HOTS has not been conducted. The arrangement of HOTS instrument is required to train and develop students' HOTS. Accordingly, to facilitate the students in accessing other learning sources, this study designed HOTS questions in an online system through an e-learning program using the Moodle LMS.

Design

In the design stage, the test instrument was designed based on the analysis result in the first stage. The test instrument design was in the form of a question matrix and outline adjusted to students' needs and characteristics and learning sources. The test was in a multiple-choice format, in which 24 questions were adjusted to the formulation of a HOTS test that had been created in the test matrix and outline. The question matrix is provided in Table 7.

Table 7
The Question Matrix

Aspects	Sub Aspects	Theories		
		Electric current, Ohm's law, and electrical power	Series and parallel circuits of resistor and capacitor	Electric Force, Kirchoff's law, and RC circuit.
Analyze	Differentiating	8	12	21
	Organizing	3	15	20
	Attributing	2	9	23
Evaluate	Checking	4	11	22
	Critiquing	1	16	18
	Generating	5	13	19
Create	Planning	7	14	17
	Producing	6	10	24

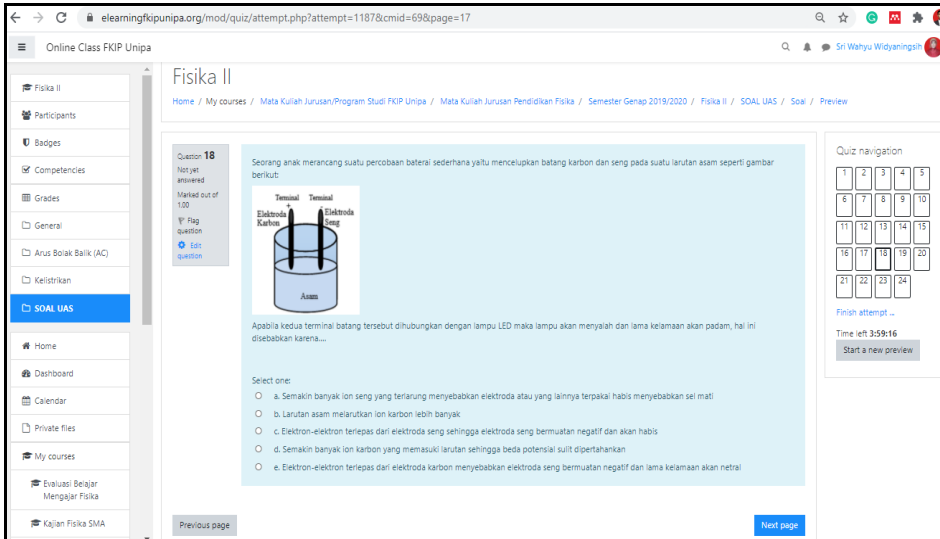
Develop

The development of HOTS questions was based on the question matrix and outline that had been designed. In addition, the questions were formulated online through e-learning by utilizing the Moodle LMS. Figure 2 below shows all question items in the e-learning program.

The screenshot displays the Moodle LMS interface for an online class. The left sidebar shows navigation options like 'Fisika II', 'Participants', 'Badges', 'Competencies', 'Grades', 'General', 'Arus Bolak Balik (AC)', 'Kelistrikan', 'SOAL UAS', 'Home', 'Dashboard', 'Calendar', 'Private files', 'My courses', 'Evaluasi Belajar Mengajar Fisika', and 'Kajian Fisika SMA'. The main content area shows a list of 20 questions across 13 pages. Each question is a multiple-choice item with a score of 1.00. The questions are in Indonesian and cover various physics topics related to electricity and circuits.

Figure 2
All Question Items in the E-Learning Program

The questions are displayed interactively, and students can randomly work on the questions. Moodle LMS can present questions with a picture or other contents to make it easier for teachers to design the questions as expected. Figure 3 illustrates one of the HOTS questions displayed on the e-learning through the Moodle LMS.



The screenshot shows a Moodle LMS quiz question interface. The question is titled "Question 18" and is a "Not yet answered" type. It asks about a simple battery experiment involving carbon and zinc electrodes in an acid solution. A diagram shows a beaker with "Asam" (acid) and two electrodes: "Terminal Karbon" (Carbon) and "Terminal Elektroda Seng" (Zinc Electrode). The question asks what happens when the terminals are connected to an LED lamp. Five multiple-choice options are provided, ranging from "many ions" to "electrons leaving the carbon electrode".

Figure 3
HOTS Questions Displayed on the E-learning Through the Moodle LMS

The development stage aims to produce a HOTS test instrument that has been validated by experts and practitioners. Product validation is a process of assessing the designed product, or in this case, the test instrument of HOTS in general physics subject in the site area. Product validation was carried out by involving seven validators, i.e., experts of measurement, physics education, physics, and practitioners. The validity test of the instrument included material, construction, and language. The analysis result of the question validity assessed by validators obtained the value of Aiken's V in the range of 0.76 to 1.00, showing a valid result. The questions validated by experts and practitioners were then revised following the provided corrections and suggestions.

Implementation

The implementation stage in this study was the product trial, in which HOTS questions were tried out to 34 students in the research site. The students worked on these questions online through e-learning by using their own Moodle account upon

completing all learning stages. Results of the students' learning can be accessed after this process.

Evaluation

Before conducting the estimate analysis of respondents' skills and item difficulty level, the analysis of item fitness was performed using INFIT and OUTFIT for mean square and t. The determination of the item fitness with the model is based on the value of INFIT MNSQ and the standard deviation or Infit t (Adams & Khoo, 1996). The fitness of each case is also based on the value of INFIT MNSQ or INFIT t of the item. Table 8 provides the testing result through the Quest program to obtain the values of item estimate and case estimate in the HOTS questions trial.

Table 8

Values of Item Estimate and Case Estimate in the HOTS Questions Trial

No	Measurement	Estimates for Items	Estimates for Testing
1.	Average values and standard deviations	0.00 ± 0.57	0.01 ± 1.24
2.	Reliability Estimates	0.66	0.85
3.	The mean and standard deviation of INFIT MNSQ	1.00 ± 0.14	0.99 ± 0.15
4.	The mean and standard deviation of OUTFIT MNSQ	1.09 ± 0.52	1.09 ± 0.52
5.	The mean and standard deviation of INFIT t	-0.03 ± 0.81	0.00 ± 0.72
6.	The mean and standard deviation of OUTFIT t	0.21 ± 0.91	0.17 ± 0.81

The analysis result suggested that the INFIT MNSQ got the range of 0.86 to 1.14, and INFIT t is -0.28 to 0.72. This signified that all 24 questions fit the model as they reached the range of INFIT MNSQ value from 0.77 to 1.30 and used INFIT t with the limit of -2.0 to 2.0. In addition to testing the fitness, the Quest program's output also presented the reliability estimate of the test instrument. The above table shows the value of item reliability based on the value of the item estimate summary, which is 0.66. On the other hand, the value of person reliability, as based on the case estimate summary, gets 0.85. These results were in line with the Rasch model, in which the reliability value fell under the range of 0.67 to 0.80 (quite reliable). On that ground, the instrument can be employed to measure students' HOTS in the General Physics subject.

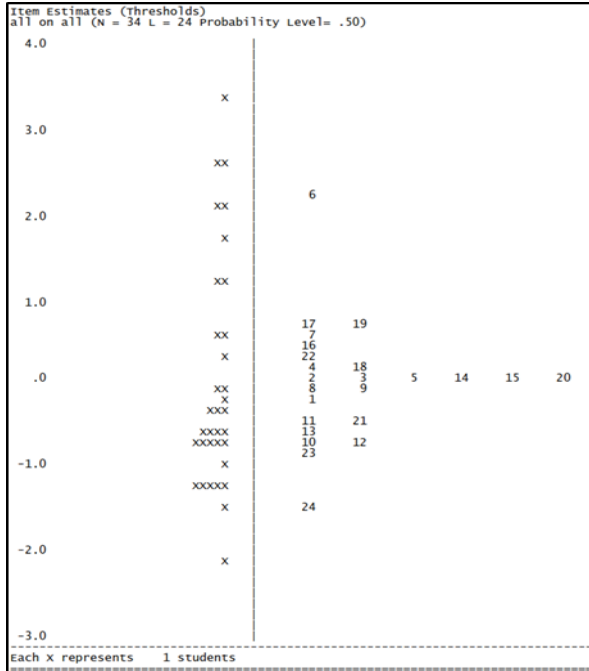


Figure 4
Distribution of Item Difficulty Level and Respondents' Skills

Figure 4 presents the distribution of the respondents according to the difficulty level in the logit scale from -4.0 to +4.0. This map displays the item difficulty level compared to the respondents' skills. Case and item difficulty levels in the Rasch model are expressed in one line in the form of abscissa in the graph with a log-odd unit. The graph of respondents' skills shows a normal curve, meaning that there are only a few respondents with low and high skills; and many respondents with moderate skills. The level of item difficulty of threshold revealed that item 6 was the most difficult question, and item 24 was the easiest one.

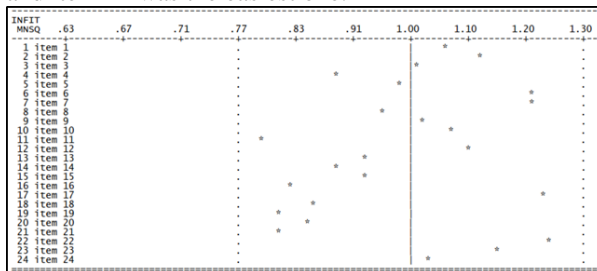


Figure 5
Distribution of INFIT MNSQ Values of Each Question Item of HOTS

Question items that fit the Rasch model are in the range of 0.77 to 1.33. By referring to Figure 5, we can see that all 24 question items are in the line, implying that they fit the Rasch model.

Item Estimates (Thresholds) In input order all on all (N = 34 L = 24 Probability Level= .50)							
ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1 item 1	18	34	-.26 .39	1.06	1.15	.4	.5
2 item 2	16	34	.04 .40	1.12	1.17	.7	.6
3 item 3	16	34	.04 .40	1.01	.91	.1	-.2
4 item 4	15	34	.19 .40	.88	.93	-.6	-.1
5 item 5	16	34	.04 .40	.98	.89	.0	-.2
6 item 6	5	34	2.27 .57	1.21	2.16	.7	1.4
7 item 7	13	34	.52 .42	1.21	1.27	1.0	.9
8 item 8	17	34	-.11 .40	.96	1.00	-.2	.1
9 item 9	17	34	-.11 .40	1.02	.91	.2	-.2
10 item 10	21	34	-.70 .39	1.07	1.16	.6	.5
11 item 11	19	34	-.41 .39	.79	.66	-1.6	-.9
12 item 12	21	34	-.70 .39	1.10	1.14	.8	.5
13 item 13	20	34	-.55 .39	.93	1.09	-.5	.4
14 item 14	16	34	.04 .40	.88	.78	-.7	-.6
15 item 15	16	34	.04 .40	.93	.82	-.4	-.5
16 item 16	13	33	.47 .42	.82	.69	-.8	-.9
17 item 17	12	34	.69 .43	1.23	1.16	1.0	.6
18 item 18	15	34	.19 .40	.86	.73	-.8	-.8
19 item 19	12	34	.69 .43	.81	.71	-.8	-.8
20 item 20	16	34	.04 .40	.85	.75	-.9	-.7
21 item 21	19	34	-.41 .39	.81	.68	-1.4	-.8
22 item 22	14	34	.35 .41	1.24	1.23	1.2	.8
23 item 23	22	34	-.85 .40	1.15	3.04	1.1	3.1
24 item 24	26	34	-1.50 .43	1.03	1.02	.2	.2
Mean			.00	1.00	1.09	.0	.1
SD			.71	.14	.52	.8	.9

Figure 6
Item Estimates of HOTS Questions

The previous figure presents the Item Estimate of HOTS questions based on the trial result. In this figure, there is SCORE-MAXSCR successively showing the respondents who answer correctly and the number of total respondents. Item 24 was the most

correctly-answered, in which 26 out of 34 respondents could work on this item. Figure 6 also provides the value of THRSHL that shows the item difficulty index in the logit scale along with its standard deviation. Item 6 got a THRSHL or difficulty index of 2.27 that was greater than 2.0, or in other words, this item was very difficult since only five students could give a correct answer. Also, the average value of THRSHL and its standard deviation accounted for 0.00 ± 0.71 and fell under the range of -2 to 2 (Hambleton & Rogers, 1989). The average value of INFIT MNSQ was 1.00 ± 0.14 and achieved the acceptance range of 0.77 to 1.33; the average value of OUTFIT t arrived at 0.10 ± 0.90 and was included in the acceptance range of ≤ 2.00 . Accordingly, these results indicate that all question items being developed can be utilized to measure students' HOTS.

Case Estimates In input order all on all (N = 34 L = 24 Probability Level= .50)								
NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFIT MNSQ	OUTFIT MNSQ	INFIT t	OUTFIT t
1 01	12	24	-.02	.43	1.06	1.02	.58	.18
2 02	6	24	-1.21	.49	1.17	1.09	.75	.36
3 03	8	24	-.77	.45	.98	1.01	-.06	.13
4 04	8	24	-.77	.45	.83	.81	-1.04	-.48
5 05	8	24	-.77	.45	.89	.83	-.59	-.41
6 06	6	24	-1.21	.49	.79	.70	-.84	-.67
7 07	10	24	-.38	.43	.99	.95	-.01	-.09
8 08	6	24	-1.21	.49	1.07	2.30	.36	2.44
9 09	3	24	-2.10	.63	.98	.85	.11	.00
10 10	9	24	-.57	.44	.88	.83	-.80	-.48
11 11	22	24	2.61	.77	.73	.46	-.30	-.56
12 12	5	24	-1.46	.52	.89	.85	-.29	-.18
13 13	20	24	1.75	.57	1.21	1.45	.64	.93
14 14	11	24	-.20	.43	.86	.83	-1.33	-.59
15 15	21	24	2.12	.64	1.18	1.05	.52	.29
16 16	9	24	-.57	.44	1.08	1.06	.59	.28
17 17	7	24	-.98	.47	1.29	2.20	1.38	2.55
18 18	6	24	-1.21	.49	1.23	1.28	.96	.75
19 19	14	24	.35	.43	.92	.87	-.56	-.40
20 20	15	24	.54	.44	.97	1.09	-.13	-.40
21 21	18	24	1.19	.49	.94	.86	-.16	-.55
22 22	21	24	2.12	.64	.93	1.23	-.01	-.55
23 23	9	24	-.57	.44	1.07	1.01	.54	.15
24 24	8	24	-.77	.45	1.01	.95	.13	-.03
25 25	10	24	-.38	.43	.87	.82	-1.06	-.57
26 26	15	24	.54	.44	1.05	1.22	.36	.80
27 27	6	24	-1.21	.49	.82	.74	-.69	-.56
28 28	22	24	2.61	.77	.73	.46	-.30	-.56
29 29	12	24	-.02	.43	.92	.88	-.73	-.39
30 30	9	24	-.57	.44	.90	.90	-.64	-.23
31 31	23	24	3.40	1.05	1.18	3.14	.49	1.53
32 32	18	24	1.19	.49	.85	.75	-.53	-.58
33 33	10	23	-.32	.44	1.11	1.11	.97	.45
34 34	8	24	-.77	.45	1.29	1.34	1.61	1.03
Mean			.01		.99	1.09	.00	.17
SD			1.35		.15	.52	.72	.81

Figure 7
Case Estimates of Every Student

Figure 7 serves as the case estimate or the skill level of each student. Information obtained from the case estimate is that the SCORE-MAXSCR shows each respondent's score from the maximum score sequentially. Respondent 31 answered the majority of the questions (23 out of 24 questions) correctly compared to other respondents. The average estimate value and its standard deviation got 0.01 ± 1.35 and were in a moderate category. The analysis result of the case estimate revealed that students' skills were in the moderate category.

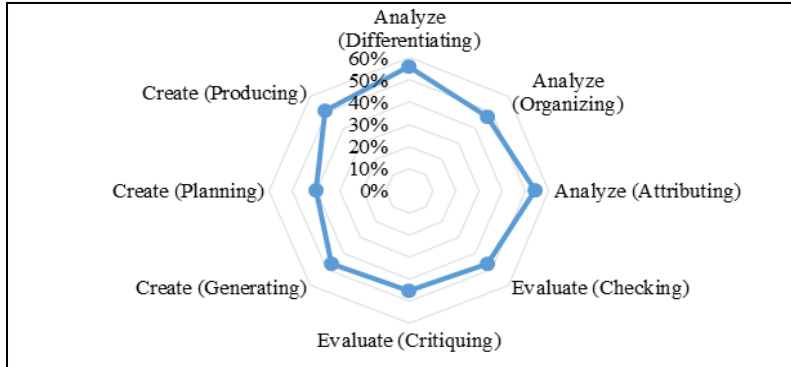


Figure 8
Distribution of Students' Answer Percentage HOTS

Figure 8 provides the percentage of students' answers based on the aspects and sub-aspects of HOTS. The analysis result pointed out that students tended to find it difficult to answer questions regarding the creating aspect, specifically the planning sub-aspect. Creating is the highest level of HOTS in Bloom's taxonomy; therefore, students need to practice developing their creating skills. This figure also signifies that most students find it easy to answer HOTS questions related to the analysis aspect, differentiating sub-aspect in particular.

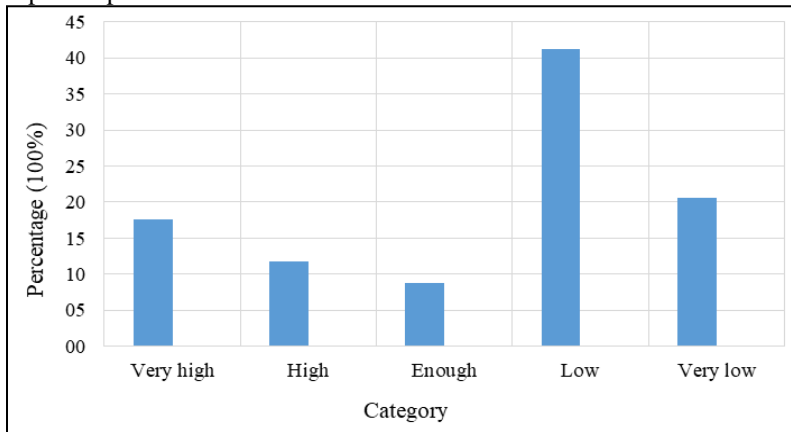


Figure 9
Percentage of Students' HOTS

The above figure shows the percentage of students' HOTS. It is seen that most students (41.2%) still have low HOTS; the categories consist of very low (20.6%), moderate (8.8%), high (11.8%), and very high (17.6%).

DISCUSSION

This study aims to produce the HOTS instrument presented in e-learning using Moodle LMS and determine the number of HOTS after using the instrument. The findings were valid and useable. The HOTS instrument validity was seen from the construct validity and face validity. Construct validity intends to investigate the HOTS instrument's accuracy and collect responses from experts and practitioners. Based on validator evaluation, the Aiken's V value was obtained from 0.76 to 1.00, suggesting a valid result. This result indicated that the HOTS instrument featured good material, design, and language aspects. The material aspect relates to the question items according to the indicators; has only one correct answer key; contents follow the calculation goal and the education level; the item distractors work properly. The construction feature of the HOTS instrument associates with the subject matter; has clearly-formulated answer choices; the subject matter does not lead to a correct answer; no multiple negative shapes; has homogeneous answer choices; has a similar length of answer choices; the items do not depend on each other; and the options are type. Next, it relates to the formulation of communicative language, grammatical sentences, non-multi-significant sentences, and standard/general/neutral vocabulary in the language aspect. Using Moodle LMS as a medium to serve HOTS instruments will promote the access of the students to online questions. E-learning using LMS Moodle is equipped with various facilities supporting online learning implementation that allows students to learn independently (Martín-Blas & Serrano-Fernández, 2009; Yildiz, Tezer, & Uzunboylu, 2018). Moodle LMS program presents an interesting display and is user-friendly (Martín-Blas & Serrano-Fernández, 2009). Students can work on the questions interactively and see the results directly.

Face validity in this analysis was obtained and evaluated based on students' HOTS instrument tests. Analyzing the HOTS instrument used IRT analysis methodology. It was suggested that all 24 items were fit as they reached the range of 0.77 to 1.30 in the MNSQ INFIT value, and -2.0 to 2.0 in the INFIT t. The item reliability value following the item estimate value summary measured at 0.66; meanwhile, the person's reliability based on the case estimate summary was 0.85 or very accurate (0.67 to 0.80). Thus, the instrument produced is appropriate for measuring students' HOTS as it has met the requirements according to the IRT analysis result.

The analysis result of students' HOTS obtained the average approximate value or skill level of each student, along with the standard deviation of 0.01 ± 1.35 (moderate category). The case estimate result indicated that the HOTS skills of the students were in the moderate category. The low category of students' HOTS was influenced by several factors, one of which was that the students were not used to working on HOTS questions (Tanujaya, Mumu, & Margono, 2017; Yusuf & Widyaningsih, 2019). They needed to

practice developing their HOTS by being exposed to HOTS-based learning sources. To realize HOTS, students are required to be more active in learning (Winarti, Cari, Widha, & Istiyono, 2015; Yusuf & Widyaningsih, 2019). Lecturers are also expected to act as facilitators who provide various learning resources and provide feedback on the students' tasks (Masruroh & Prasetyo, 2018). The use of e-learning allows students to access different learning resources in the form of texts, animations, simulations, multimedia, or virtual laboratories that can be accessed directly (Skultety, Gonzalez, & Vargas, 2017; Tee, Siti, Tengku, & Zainudin, 2013). It is expected that these e-learning facilities can facilitate students in learning so that their HOTS can be developed. Students' HOTS can also be improved through assignments and exercises in the learning process (Istiyono, Dwandaru, Megawati, & Ermansah, 2018; Yusuf & Widyaningsih, 2018). On this ground, it is of major importance to train the students' HOTS by applying learning technologies and quality instrument presentations through the IRT analysis.

CONCLUSION

The HOTS instrument presented by Moodle LMS in e-learning obtains a good performance. The IRT analysis, including item fit, reliability, and difficulty, acquires the mean and standard deviation parameters for INFIT MNSQ of 1.0 and 0.0; the items have proven to fit RM 1-PL. Additionally, test characteristics comprised item fitness, reliability, and difficulty. The trial result obtains the criteria of INFIT MNSQ mean and standard deviation of 1.0 and 0.0, respectively, implying that the items fit the RM1-PL. In addition, the value of item reliability based on the value of item estimate summary arrives at 0.66; meanwhile, the person reliability under the case estimate summary reaches 0.85, i.e., the reliability value is in the range of 0.67 - 0.80 (quite reliable). As based on the criteria of minimum and maximum INFIT MNSQ of 0.77 and 1.30, 24 question items fit the RM 1-PL model. The Quest output result also reveals that the average values of THRSHL and its standard deviation are 0.00 ± 0.71 , or in the acceptance range of -2 to 2. To sum up, all 24 question items that had been tried out have fit the model with a good category, so that they can be used in the HOTS measurement. Every student's average estimate or skill level along with the standard deviation is 0.01 ± 1.35 or in the moderate category. Students' HOTS must be practiced by providing HOTS-based learning resources.

ACKNOWLEDGMENT

We would like to acknowledge the contribution of the Ministry of Research and Higher Education in funding this study through the Inter-University Cooperation scheme with the contract number: 198/SP2H/AMD/LT/DRPM/2020.

Author surnames go here

17

REFERENCES

- Adams, R. J., & Khoo, S.-T. (1996). *Quest: the interactive test analysis system*. Camberwell, Vic.: Australian Council for Educational Research.
- Aiken, L. R. (1980). Content Validity and Reliability of Single Items or Questionnaires. *Educational and Psychological Measurement*, 40(4), 955–959.
- Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, 45(1), 131–142.
- Aldoobie, N. (2015). ADDIE Model. *American International Journal of Contemporary Research*, 5(6), 72.
- Azevedo, J. M. (2015). e-Assessment in mathematics courses with multiple-choice questions tests. *CSEDU 2015 - 7th International Conference on Computer Supported Education, Proceedings*, 2, 260–266. <https://doi.org/10.5220/0005452702600266>
- Bogdanović, Z., Barać, D., Jovanić, B., Popović, S., & Radenković, B. (2014). Evaluation of Mobile Assessment in A Learning Management System. *British Journal of Educational Technology*, 45(2), 231–244. <https://doi.org/10.1111/bjet.12015>
- Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York: Routledge.
- Brun, M., & Hinostroza, J. E. (2014). Learning to become a teacher in the 21st century: ICT integration in Initial Teacher Education in Chile. *Journal of Educational Technology & Society*, 17(3), 222–238.
- Hambleton, R. K., & Rogers, H. J. (1989). Detecting Potentially Biased Test Items: Comparison of IRT Area and Mantel-Haenszel Methods. *Applied Measurement in Education*, 2(4), 313–334. https://doi.org/10.1207/s15324818ame0204_4
- Istiyono, E. (2017). The Analysis of Senior High School Students' Physics HOTS in Bantul District Measured using PhysReMChoTHOTS. *AIP Conference Proceedings*, 1868(August), 1–7. <https://doi.org/10.1063/1.4995184>
- Istiyono, E. (2018). IT-based HOTS assessment on physics st learning as the 21 century demand at senior high schools: Expectation and reality IT-Based HOTS Assessment on Physics Learning as the 21 st Century Demand at Senior High Schools: Expectation and Reality. *AIP Conference Proceedings*, 2014(020014), 1–6.
- Istiyono, E., Dwandaru, W. S. B., Megawati, I., & Ermansah. (2018). Application of Bloomian and Marzanoian Higher Order Thinking Skills in the Physics Learning Assessment: an Inevitability. *Advances in Social Science, Education and Humanities Research*, 164(ICLI 2017), 136–142. <https://doi.org/10.2991/icli-17.2018.26>

- Istiyono, E., Dwandaru, W. S. B., & Muthmainah. (2019). Developing of Bloomian HOTS Physics Test: Content and Construct Validation of The PhysTeBloHOTS. *Journal of Physics: Conference Series*, 1397(012017), 1–9.
- Kowsalya, D. N., Venkat Lakshmi, H., & Suresh, K. P. (2012). Development and Validation of a Scale to assess Self-Concept in Mild Intellectually Disabled Children. *International Journal of Social Sciences & Education*, 2(4).
- Krathwohl, D. R., & Anderson, L. W. (2010). Merlin C. Wittrock and the Revision of Bloom's Taxonomy. *Educational Psychologist*, 45(1), 64–65. <https://doi.org/10.1080/00461520903433562>
- Lee, M. F., & Zainal, N. A. (2017). Development of needham model based E-module for electromagnetic field & wave. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 120–124). <https://doi.org/10.1109/IEEM.2017.8289863>
- Limongelli, C., Sciarrone, F., & Vaste, G. (2011). Personalized e-learning in Moodle: the Moodle_LS System. *Journal of E-Learning and Knowledge Society*, 7(1), 49–58. Retrieved from <https://www.learntechlib.org/p/43340>
- Martín-Blas, T., & Serrano-Fernández, A. (2009). The role of new technologies in the learning process: Moodle as a teaching tool in Physics. *Computers & Education*, 52(1), 35–44. <https://doi.org/10.1016/J.COMPEDU.2008.06.005>
- Masruroh, A. N., & Prasetyo, Z. K. (2018). Effect of E-Module with Guided Inquiry Approach Containing Nature of Science to Student's Science Literacy. *E-Journal Pend. IPA*, 7(3), 165–171.
- Pandey, S. R., & Pandey, S. (2009). Developing a More Effective and Flexible Learning Management System (LMS) for the Academic Institutions using Moodle. *ICAL 2009 - Technology, Policy and Innovation*, 249–254.
- Raykov, T., & Marcoulides, G. A. (2015). On the Relationship Between Classical Test Theory and Item Response Theory: From One to the Other and Back. *Educational and Psychological Measurement*, 76(2), 325–338. <https://doi.org/10.1177/0013164415576958>
- Skultety, L., Gonzalez, G., & Vargas, G. (2017). Using Technology to Support Teachers' Lesson Adaptations during Lesson Study. *Journal of Technology and Teacher Education*, 25(2), 185–213. Retrieved from <https://www.learntechlib.org/p/172139>
- Tanujaya, B., Mumu, J., & Margono, G. (2017). The Relationship between Higher Order Thinking Skills and Academic Performance of Student in Mathematics Instruction. *International Education Studies*, 10(11), 78–85.
- Tee, S. S., Siti, T., Tengku, M., & Zainudin, S. (2013). User Testing for Moodle Application. *International Journal of Software Engineering and Its Applications*, 7(5), 243–252.

Author surnames go here

19

- Winarti, Cari, Widha, S., & Istiyono, E. (2015). Analysis of Higher Order Thinking Skills Content of Physics Examinations In Madrasah Aliyah. In *International Conference on Mathematics, Science, and Education 2015 (ICMSE 2015)* (Vol. 2015, pp. 32–38).
- Yildiz, E. P., Tezer, M., & Uzunboylu, H. (2018). Student Opinion Scale Related to Moodle LMS in an Online Learning Environment: Validity and Reliability Study. *International Journal of Interactive Mobile Technologies (IJIM)*, 12(4), 97–108.
- Yusuf, I., & Widyaningsih, S. W. (2018). Profil Kemampuan Mahasiswa dalam Menyelesaikan Soal HOTS di Jurusan Pendidikan Fisika Universitas Papua. *Jurnal Komunikasi Pendidikan*, 2(14), 42–49.
- Yusuf, I., & Widyaningsih, S. W. (2019). HOTS profile of physics education students in STEM-based classes using PhET media. *Journal of Physics: Conference Series*, 1157(032021), 1–5.
- Yusuf, I., Widyaningsih, S. W., & Sebayang, S. R. B. (2018). Implementation of E-learning based-STEM on Quantum Physics Subject to Student HOTS Ability. *Turkish Science Education*, 15(December), 67–75.



CERTIFICATE OF PROFESSIONAL PROOFREADING AND EDITING

To Whom It May Concern:

This is to certify that the document titled “*The Development of the HOTS Test of Physics Based on Modern Test Theory: Question Modeling through E-learning of Moodle LMS*”. Commissioned to us by Sri Wahyu Widyaningsih, Irfan Yusuf, Zuhdan Kun Prasetyo, Edi Istiyono has been proofread and edited for English grammar, punctuation and spelling by Transbahasa Professional Translation and Language services.

Director,



Novriyanto Napu, M.AppLing., PhD

Disclaimer: The author is free to accept or reject our changes in the document after our edit. However, we do not bear responsibility for revisions made to the document after our edit on 05/10/2020.

TRANSBAHASA

Professional Translation and Language Services

SK Menteri Hukum dan HAM RI No. AHU-0009641.AH.01.07.2017

Jl. Ir. Hi. Joesoef Dalie 34 Kota Gorontalo Email. transbahasa.go@gmail.com / Phone. +62 853 9862 5876

www.transbahasa.co.id



International Journal of Instruction Article Evaluation Form

Mr. /Mrs.

It is to acknowledge you that the Executive Committee of *International Journal of Instruction* has decided that the article mentioned below would be reviewed by you. Thank you very much for your contributions.

Asim ARI
Editor in Chief

Name of the article: The Development of HOTS Test of Physics Based on the Modern Test Theory: Question Modeling through E-learning of Moodle LMS

After reviewing the attached article, please read each item carefully and select the response that best reflects your opinion. To register your response, please **mark** or **type in** the appropriate block.

	Yes	Partially	No
Do you think the title is appropriate?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Does the abstract summarize the article clearly and effectively?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Are the objectives set clearly?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the issue stated clearly?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the literature review adequate?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the design of the research appropriate, and the exemplary, if any, suitable?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the methodology consistent with the practice?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Are the findings expressed clearly?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the presentation of the findings adequate and consistent?	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Are the tables, if any, arranged well?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Are the conclusions and generalizations based on the findings?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Are the suggestions meaningful, valid, and based on the findings?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Are the references adequate?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the language clear and understandable?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is cohesion achieved throughout the article?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Is the work contributing to the field?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

- Evaluation:**
- The article can be published as it is.
 - The article can be published after some revision.
 - The article must undergo a major revision before it can be resubmitted to the journal.
 - The article cannot be published.

Would you like to see the revised article if you have suggested any revisions? Yes No

Please write your report either on this paper or on a spare paper.

REPORT

Section of the Manuscript	Comments and Notes
Title- Abstract-Summary	The title has been showed the content of article, but abstract must be revised in the purpose section
Introduction and Literature Review	good
Research Methods	good
Research Findings	Some figures are poor and not clear. Therefore, it must be revised.



Discussion	good
Conclusion and Suggestions	good
References and Citation	good
Language	good
Other issues	-

The Development of the HOTS Test of Physics Based on Modern Test Theory: Question Modeling through E-learning of Moodle LMS

The present study discussed the development of the HOTS test of physics based on modern test theory. HOTS questions were designed and presented in the e-learning. Further, this research employed the ADDIE model with analysis, design, development, implementation, and evaluation stages. The instrument consisted of 24 multiple-choice physics questions; the questions were designed by following the aspects and sub-aspects of HOTS and validated by the assessment of physics experts, physicists, and lecturers. Moreover, the validity analysis was based on Aiken's V formula, in which every aspect was confirmed to be valid. The instrument had been tested on 34 students of the Physics Education Department, Universitas Papua. Dichotomy data analysis used the Rasch Model (RM) 1-PL through the Quest program, and the test characteristics comprised item fitness, reliability, and difficulty. The trial result obtained the criteria of INFIT MNSQ mean and standard deviation of 1.0 and 0.0, respectively, indicating that the items fitted the RM1-PL. In addition, the value of item reliability based on the item estimate summary arrived at 0.66; meanwhile, the case reliability under the summary of the case estimate accounted for 0.85. The reliability value in the range of 0.67- 0.80 was categorized as quite reliable. Drawing upon the criteria of minimum and maximum INFIT MNSQ of 0.77 and 1.30, 24 question items fitted the RM 1-PL model. The Quest output result also suggested that the average values of Thresholds and its standard deviation were 0.00 ± 0.71 , or in the acceptance range of -2 to 2. Overall, all 24 question items that had been tested have fitted the model with a good category. They can be used in the HOTS measurement and can increase students' HOTS.

Keywords: E-learning, HOTS Test, and Modern Test Theory.

INTRODUCTION

Assessment, especially in the cognitive domain, is central to the learning process and should be carried out accurately and in compliance with the subject to be assessed or measured. Students' cognitive skills in the learning process can be categorized into Lower-Order Thinking Skills (LOTS) and Higher-Order Thinking Skills (HOTS). The LOTS includes remembering, understanding, and applying; the HOTS, on the other hand, consists of analyzing, evaluating, and creating. HOTS is thinking skills that require not only the remembering skill but also other higher skills. Indicators to measure HOTS encompass analyzing (C4), evaluating (C5), and creating (C6) skills (Krathwohl & Anderson, 2010).

HOTS also refers to thinking skills when one takes new information, connects it with initial information s/he has, and finally delivers the information to achieve goals or answer questions (Istiyono, Dwandaru, & Muthmainah, 2019). This is in line with skill characteristics in the 21st century published by Partnership of 21st Century Skill stating that 21st-century learners should be able to develop competitive skills, such as critical thinking, problem-solving, communication, information and communication technology

Commented [p1]: The study aimed, do not discussed

(ICT) literacy, ICT, information literacy, and media literacy (Brun & Hinostrroza, 2014); these focus on HOTS development.

Physics serves as part of science, comprising abstract concepts that are difficult to be directly described. Learning physics is expected to help students develop their thinking skills, in which they are not only demanded to master LOTS, but also HOTS. Teachers are also urged to deliver learning materials to students, including the HOTS, that can be improved by the HOTS instrument. A previous study has reported that the majority of teachers find it challenging to formulate an assessment instrument of learning outcomes, HOTS questions, in particular (Istiyono, 2018). For this reason, teachers' creativity is highly required to measure student learning outcomes. Today's development of ICT can be utilized to design and habituate students to learn anywhere at any time (Yusuf, Widyaningsih, & Sebayang, 2018). Relying on ICT during the learning process is one of the significant innovations, including the evaluation of student learning outcomes.

Evaluation questions can be posed in an integrated manner through e-learning systems, such as Moodle Learning Management System (LMS) (Azevedo, 2015; Bogdanović, Barać, Jovanić, Popović, & Radenković, 2014). The Moodle provides different types of questions, namely multiple choices, true or false, and short answers; these are stored in the taught course database and can be reapplied (Limongelli, Sciarrone, & Vaste, 2011). Teachers are also able to offer feedback directly to the students and give them correct answers to questions they have worked on (Pandey & Pandey, 2009). One of the advantages of an online evaluation through Moodle LMS is that students can figure out their assessment results right away.

Teachers need to prepare a good test to measure student learning outcomes. There are two paradigms developed to assess student learning outcomes through the used test, i.e., classical and modern approaches. The classical paradigm being utilized is classical test theory or widely known as classical true-score theory; meanwhile, the modern paradigm is item response theory (IRT). The classical test theory is selected due to its ease in the application despite its limitations in measuring the item difficulty level and discrimination since both indicators' calculation is based on the test taker's total score. In contrast, the IRT frees up the dependence between the test item and the test taker (a concept of parameter invariance); the test taker's response to a test item does not affect another item (a concept of local independence), and; the test item does only measure one measurement dimension (Raykov & Marcoulides, 2015). Therefore, the application answers the needs of modern measurement to date, i.e., comparing test taker's skills, question development, and even adaptive test development. It is considered able to overcome the limitations of the classical test theory.

On account of the simplicity of the analysis, most teachers have analyzed assessment tools using classical analysis techniques. The use of classical analytical techniques features some limitations, including the difficulty of defining individual learners' skills. The calculated error of measurement does not include persons but groups together. This is because each test taker's response to the questions cannot be clarified by classical test

theory. Efforts are thereby required to free the measuring tool from attachment to the sample (sample-free) employing the IRT.

This is a preliminary study with a long-term purpose of developing general physics questions with good quality at the Department of Physics Education, Universitas Papua. As the first stage, this study focuses on students at the department mentioned previously who enroll in General Physics subject taught by the researcher. This study also serves as one of the efforts to expand students' HOTS by applying a variety of HOTS-based learning sources. This research aims to develop HOTS physics questions based on IRT designed and presented with LMS Moodle on e-learning, which can be accessed online.

METHOD

As employed by this study, the ADDIE model refers to a general and systematic model of development study with a phased framework, allowing each element to connect (Aldoobie, 2015). The stages of this model used in the development of the HOTS instrument are presented in Figure 1.

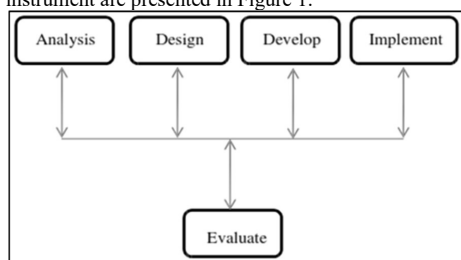


Figure 1
Stages of ADDIE Development Model in Designing Moodle LMS-based HOTS Test

Analysis

The analysis stage was a process of needs analysis to determine test objectives, identify problems, analyze tasks, and determine question formats to be applied. It was shown that the problems were related to the needs of HOTS instrument design for students at the Department of Physics Education, Universitas Papua.

Design

This stage comprised the process of designing HOTS questions to be used; the design process encompassed creating a question matrix and outline that covered question distribution in every aspect and sub-aspect of HOTS.

Develop

Every single thing required in the arrangement of HOTS questions has been prepared in the next stage. This stage also covered the process of making the questions regarding HOTS, as well as validating the questions that involved the experts of measurement,

physics education, and practitioners. The validity analysis technique to assess the content validity of the developed questions relied on the Aiken's V formula (Aiken, 1980, 1985).

$$V = \frac{\sum s}{n(c-1)} \quad (1)$$

"V" refers to the agreement index of validators in regards to item validity; "s" is the assessment score of validators subtracted by the assessment lowest score; "n" refers to the number of validators; "c" is the number of categories that can be chosen by validators. All test items are considered valid if the value of the Aiken's V index falls under the range of 0.37 to 1.00 (Kowsalya, Venkat Lakshmi, & Suresh, 2012). The value of Aiken's V of every test item was calculated based on the assessment items of every validator. In this stage, there was also an evaluation process, i.e., revising questions by following validators' corrections and suggestions.

Implementation

Another stage was applying HOTS questions that had been developed to 34 students in the site area who enrolled in general physics subject. This number followed the sample size for data stability in Rasch Model (RM) 1- PL, which is from 30 to 300, with the limit of INFIT t is from -2 to +2 (Bond, Yan, & Heene, 2020). Question item analysis was performed based on the raw score of the students by employing the Quest program.

Evaluation

The evaluation was a process of finding out whether HOTS's developed questions had met the expectation. The evaluation stage is carried out in every stage and is called a formative evaluation intended for revisions (Lee & Zainal, 2017). For instance, in the design stage, the expert's review is necessary to provide input towards the design. Besides, the evaluation stage was undertaken after analyzing empirical questions mathematically by using the Quest software program by referring to the Rasch model. The Quest program can do the Rasch measurement, i.e., a comprehensive empirical test of question items. There were three parameters being measured mathematically based on the empirical test of question items, as follows.

1. The first parameter is item fitness with the Rasch model by following the value of INFIT MNSQ or INFIT t of the item. The expected values of the unweighted mean square (Outfit MNSQ) in the Quest program and weighted mean square are 1; the variance is 0. On the contrary, the expected value of Mean INFIT t is equal to 0, with the variance equal to 1 (Adams & Khoo, 1996). The provision of INFIT MNSQ for the Rasch Model is presented in Table 1 and Table 2 below.

Table 1

Criteria of Question Item Fitness with the Rasch Model

MNSQ INFIT Value	Criteria
>1.33	Does Not Fit the Rasch Model
0.77 to 1.33	Fits the Rasch Model
<0.77	Does Not Fit the Rasch Model

Table 2

The Provision of Outfit t for the Rasch Model.

t OUTFIT Value	Criteria
OUTFIT $t \leq 2.00$	Fits the Rasch Model
OUTFIT $t \geq 2.00$	Does Not Fit the Rasch Model

2. The second parameter is reliability. The analysis result of the Quest program also showed the item and case reliability. The reliability value based on the item estimate is also called sample reliability; the higher the value, the more the items that fit the tested model. Whereas, the lower the value, the less the items that fit the tested model, so that it does not give the expected information. The reliability category is provided in the following table (Istiyono, 2017).

Table 3

Interpretation of Reliability Value

Reliability Value	Criteria
> 0.94	Excellent
$0.91 - 0.94$	Very Good
$0.81 - 0.90$	Good
$0.67 - 0.80$	Fair
< 0.67	Poor

3. The third parameter is the item difficulty index and respondents' skills presented as difficulty index in the Quest output. Thresholds (THRSHL) show the item difficulty index in the logit scale along with its standard deviation (Hambleton & Rogers, 1989). The provision of the THRSHL value is in Table 4.

Table 4

Criteria of THRSHL Value to Categorize Item Difficulty Level

THRSHL Value	Criteria
$b > 2.00$	Very Difficult
$1.00 < b \leq 2.00$	Difficult
$-1.00 < b \leq 1.00$	Medium
$-1.00 > b \geq 2.00$	Easy
$b < -2.00$	Very Easy

Respondents' skills were shown by the value of the estimate error, in which the criteria of the estimate value of respondents' skills are given in Table 5.

Table 5

Criteria of Estimate Value to Categorize Respondents' Skills

THRSHL Value	Criteria
$b > 2.00$	Very Difficult
$1.00 < b \leq 2.00$	Difficult
$-1.00 < b \leq 1.00$	Medium
$-1.00 > b \geq 2.00$	Easy
$b < -2.00$	Very Easy

The evaluation stage also included the process of analyzing the HOTS of students on the whole. The level of HOTS is categorized based on the ideal mean and standard deviation. This was applied with the assumption that students' HOTS of physics were normally

distributed. The ideal mean (I_m) and ideal standard deviation (I_{sd}) are based on the highest and lowest score of research variables. Table 6 shows the criteria of students' HOTS of physics.

Table 6
Criteria of Students' HOTS of Physics

Interval	Criteria
$I_m + 1.5 I_{sb} < \theta$	Very high
$I_m + 0.5 I_{sb} < \theta \leq I_m + 1.5 I_{sb}$	High
$I_m - 0.5 I_{sb} < \theta \leq I_m + 0.5 I_{sb}$	Moderate
$I_m - 1.5 I_{sb} < \theta \leq I_m - 0.5 I_{sb}$	Low
$0 < I_m - 1.5 I_{sb}$	Very Low

Meaning:

I_m : ideal mean

I_{sb} : ideal standard deviation

X_{mak} : highest score

X_{min} : lowest score

RESULTS

The ADDIE development model can be used for different product developments in education, and one of which is the development of HOTS questions. This model is simple and systematically structured in its implementation stages. The following is a description of each stage result.

Analysis

A needs analysis was the first stage being done by observation and interview to gather any information required in physics learning at the Department of Physics Education, Universitas Papua. The researchers' experience indicated that the lecturers had applied HOTS learning in the classroom. However, a test to measure students' HOTS has not been conducted. The arrangement of HOTS instrument is required to train and develop students' HOTS. Accordingly, to facilitate the students in accessing other learning sources, this study designed HOTS questions in an online system through an e-learning program using the Moodle LMS.

Design

In the design stage, the test instrument was designed based on the analysis result in the first stage. The test instrument design was in the form of a question matrix and outline adjusted to students' needs and characteristics and learning sources. The test was in a multiple-choice format, in which 24 questions were adjusted to the formulation of a HOTS test that had been created in the test matrix and outline. The question matrix is provided in Table 7.

Table 7
The Question Matrix

Aspects	Sub Aspects	Theories		
		Electric current, Ohm's law, and electrical power	Series and parallel circuits of resistor and capacitor	Electric Force, Kirchoff's law, and RC circuit.
Analyze	Differentiating	8	12	21
	Organizing	3	15	20
	Attributing	2	9	23
Evaluate	Checking	4	11	22
	Critiquing	1	16	18
Create	Generating	5	13	19
	Planning	7	14	17
	Producing	6	10	24

Develop

The development of HOTS questions was based on the question matrix and outline that had been designed. In addition, the questions were formulated online through e-learning by utilizing the Moodle LMS. Figure 2 below shows all question items in the e-learning program.

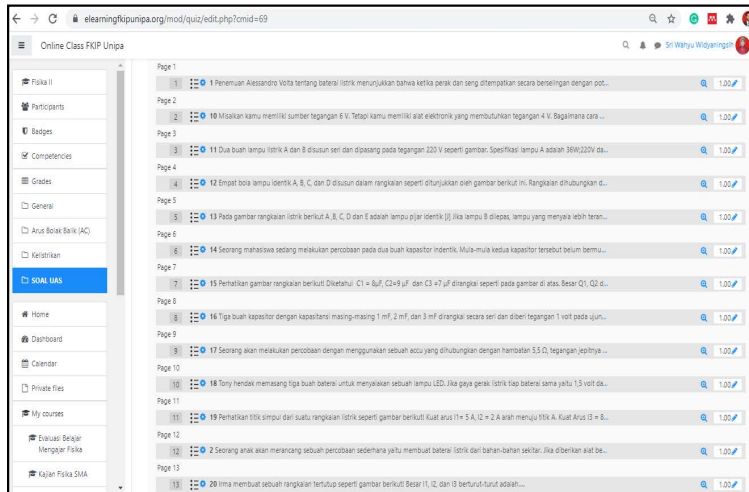
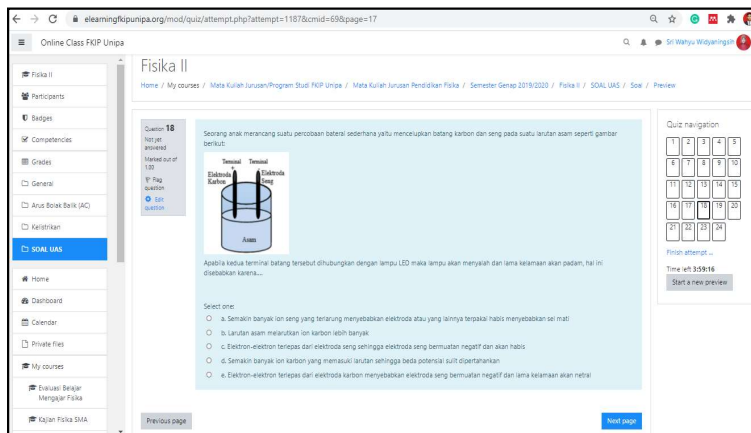


Figure 2
All Question Items in the E-Learning Program

The questions are displayed interactively, and students can randomly work on the questions. Moodle LMS can present questions with a picture or other contents to make it easier for teachers to design the questions as expected. Figure 3 illustrates one of the HOTS questions displayed on the e-learning through the Moodle LMS.

Commented [p2]: Poor figure. It must be revised



Commented [p3]: Poor figure. It must be revised

Figure 3
HOTS Questions Displayed on the E-learning Through the Moodle LMS

The development stage aims to produce a HOTS test instrument that has been validated by experts and practitioners. Product validation is a process of assessing the designed product, or in this case, the test instrument of HOTS in general physics subject in the site area. Product validation was carried out by involving seven validators, i.e., experts of measurement, physics education, physics, and practitioners. The validity test of the instrument included material, construction, and language. The analysis result of the question validity assessed by validators obtained the value of Aiken's V in the range of 0.76 to 1.00, showing a valid result. The questions validated by experts and practitioners were then revised following the provided corrections and suggestions.

Implementation

The implementation stage in this study was the product trial, in which HOTS questions were tried out to 34 students in the research site. The students worked on these questions online through e-learning by using their own Moodle account upon completing all learning stages. Results of the students' learning can be accessed after this process.

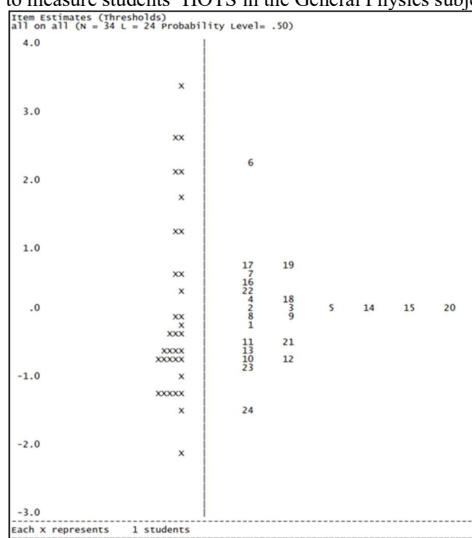
Evaluation

Before conducting the estimate analysis of respondents' skills and item difficulty level, the analysis of item fitness was performed using INFIT and OUTFIT for mean square and t. The determination of the item fitness with the model is based on the value of INFIT MNSQ and the standard deviation or Infit t (Adams & Khoo, 1996). The fitness of each case is also based on the value of INFIT MNSQ or INFIT t of the item. Table 8 provides the testing result through the Quest program to obtain the values of item estimate and case estimate in the HOTS questions trial.

Table 8
Values of Item Estimate and Case Estimate in the HOTS Questions Trial

No.	Measurement	Estimates for Items	Estimates for Testing
1.	Average values and standard deviations	0.00 ± 0.57	0.01 ± 1.24
2.	Reliability Estimates	0.66	0.85
3.	The mean and standard deviation of INFIT MNSQ	1.00 ± 0.14	0.99 ± 0.15
4.	The mean and standard deviation of OUTFIT MNSQ	1.09 ± 0.52	1.09 ± 0.52
5.	The mean and standard deviation of INFIT t	-0.03 ± 0.81	0.00 ± 0.72
6.	The mean and standard deviation of OUTFIT t	0.21 ± 0.91	0.17 ± 0.81

The analysis result suggested that the INFIT MNSQ got the range of 0.86 to 1.14, and INFIT t is -0.28 to 0.72. This signified that all 24 questions fit the model as they reached the range of INFIT MNSQ value from 0.77 to 1.30 and used INFIT t with the limit of -2.0 to 2.0. In addition to testing the fitness, the Quest program's output also presented the reliability estimate of the test instrument. The above table shows the value of item reliability based on the value of the item estimate summary, which is 0.66. On the other hand, the value of person reliability, as based on the case estimate summary, gets 0.85. These results were in line with the Rasch model, in which the reliability value fell under the range of 0.67 to 0.80 (quite reliable). On that ground, the instrument can be employed to measure students' HOTS in the General Physics subject.



Author surnames go here

11

Figure 4
Distribution of Item Difficulty Level and Respondents' Skills

Figure 4 presents the distribution of the respondents according to the difficulty level in the logit scale from -4.0 to +4.0. This map displays the item difficulty level compared to the respondents' skills. Case and item difficulty levels in the Rasch model are expressed in one line in the form of abscissa in the graph with a log-odd unit. The graph of respondents' skills shows a normal curve, meaning that there are only a few respondents with low and high skills; and many respondents with moderate skills. The level of item difficulty of threshold revealed that item 6 was the most difficult question, and item 24 was the easiest one.

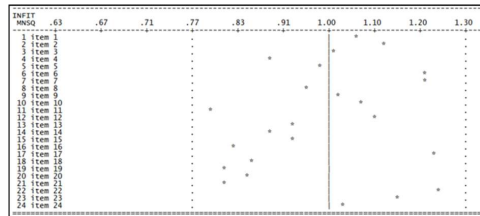


Figure 5
Distribution of INFIT MNSQ Values of Each Question Item of HOTS

Question items that fit the Rasch model are in the range of 0.77 to 1.33. By referring to Figure 5, we can see that all 24 question items are in the line, implying that they fit the Rasch model.

Item Estimates (Thresholds) In input Order							
all on all (N = 34 L = 24 Probability Level = .50)							
ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1 item 1	18	34	-.26 .39	1.06	1.15	.4	.5
2 item 2	16	34	.04 .40	1.12	1.17	.7	.6
3 item 3	16	34	.04 .40	1.01	.91	.1	-.2
4 item 4	15	34	.19 .40	.88	.93	-.6	-.1
5 item 5	16	34	.04 .40	.98	.89	.0	-.2
6 item 6	5	34	2.27 .57	1.21	2.16	.7	1.4
7 item 7	13	34	.52 .42	1.21	1.27	1.0	.9
8 item 8	17	34	-.11 .40	.96	1.00	-.2	.1
9 item 9	17	34	-.11 .40	1.02	.91	.2	-.2
10 item 10	21	34	-.70 .39	1.07	1.16	.6	.5
11 item 11	19	34	-.41 .39	.79	.66	-1.6	-.9
12 item 12	21	34	-.70 .39	1.10	1.14	.8	.5
13 item 13	20	34	-.55 .39	.93	1.09	-.5	.4
14 item 14	16	34	.04 .40	.88	.78	-.7	-.6
15 item 15	16	34	.04 .40	.93	.82	-.4	-.5
16 item 16	13	33	.47 .42	.82	.69	-.8	-.9
17 item 17	12	34	.69 .43	1.23	1.16	1.0	.6
18 item 18	15	34	.19 .40	.86	.73	-.8	-.8
19 item 19	12	34	.69 .43	.81	.71	-.8	-.8
20 item 20	16	34	.04 .40	.85	.75	-.9	-.7
21 item 21	19	34	-.41 .39	.81	.68	-1.4	-.8
22 item 22	14	34	.35 .41	1.24	1.23	1.2	.8
23 item 23	22	34	-.85 .40	1.15	3.04	1.1	3.1
24 item 24	26	34	-1.50 .43	1.03	1.02	.2	.2
Mean			.00	1.00	1.09	.0	.1
SD			.71	.14	.52	.8	.9

Figure 6
Item Estimates of HOTS Questions

The previous figure presents the Item Estimate of HOTS questions based on the trial result. In this figure, there is SCORE-MAXSCR successively showing the respondents who answer correctly and the number of total respondents. Item 24 was the most correctly-answered, in which 26 out of 34 respondents could work on this item. Figure 6 also provides the value of THRSHL that shows the item difficulty index in the logit scale along with its standard deviation. Item 6 got a THRSHL or difficulty index of 2.27 that was greater than 2.0, or in other words, this item was very difficult since only five

Author surnames go here

13

students could give a correct answer. Also, the average value of THRSHL and its standard deviation accounted for 0.00 ± 0.71 and fell under the range of -2 to 2 (Hambleton & Rogers, 1989). The average value of INFIT MNSQ was 1.00 ± 0.14 and achieved the acceptance range of 0.77 to 1.33; the average value of OUTFIT t arrived at 0.10 ± 0.90 and was included in the acceptance range of ≤ 2.00 . Accordingly, these results indicate that all question items being developed can be utilized to measure students' HOTS.

Case Estimates in Input Order
all on all (N = 34 L = 24 Probability Level= .50)

NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFIT MNSQ	OUTFIT MNSQ	INFIT t	OUTFIT t
1 01	12	24	-.02	.43	1.06	1.02	-.58	.18
2 02	6	24	-1.21	.49	1.17	1.09	-.75	.36
3 03	8	24	-.77	.45	.98	1.01	-.06	-.13
4 04	8	24	-.77	.45	.83	.81	-1.04	-.48
5 05	8	24	-.77	.45	.89	.83	-.59	-.41
6 06	6	24	-1.21	.49	.79	.70	-.84	-.67
7 07	10	24	-.38	.43	.99	.95	-.01	-.09
8 08	6	24	-1.21	.49	1.07	2.30	-.36	2.44
9 09	3	24	-2.10	.63	.98	.85	-.11	.00
10 10	9	24	-.57	.44	.88	.83	-.80	-.48
11 11	22	24	2.61	.77	.73	.46	-.20	-.56
12 12	5	24	-1.46	.52	.89	.85	-.29	-.18
13 13	20	24	1.75	.57	1.21	1.45	.64	.93
14 14	11	24	-.20	.43	.86	.83	-1.33	-.59
15 15	21	24	2.12	.64	1.18	1.05	.52	.29
16 16	9	24	-.57	.44	1.08	1.06	.59	.28
17 17	7	24	-.98	.47	1.29	2.20	1.38	2.55
18 18	6	24	-1.21	.49	1.23	1.28	.96	.75
19 19	14	24	.35	.43	.92	.87	-.56	-.40
20 20	15	24	.54	.44	.97	1.09	-.13	.40
21 21	18	24	1.19	.49	.94	.86	-.16	-.25
22 22	21	24	2.12	.64	.93	1.23	-.01	.55
23 23	9	24	-.57	.44	1.07	1.01	.54	.15
24 24	8	24	-.77	.45	1.01	.95	.13	-.03
25 25	10	24	-.38	.43	.87	.82	-1.06	-.57
26 26	15	24	.54	.44	1.05	1.22	.36	.80
27 27	6	24	-1.21	.49	.82	.74	-.69	-.56
28 28	22	24	2.61	.77	.73	.46	-.30	-.56
29 29	12	24	-.02	.43	.92	.88	-.73	-.39
30 30	9	24	-.57	.44	.90	.90	-.64	-.23
31 31	23	24	3.40	1.05	1.18	3.14	.49	1.53
32 32	18	24	1.19	.49	.85	.75	-.53	-.58
33 33	10	23	-.32	.44	1.11	1.11	-.97	-.45
34 34	8	24	-.77	.45	1.29	1.34	1.61	1.03
Mean			.01		.99	1.09	.00	.17
SD			1.35		.15	.52	.72	.81

Figure 7
Case Estimates of Every Student

Figure 7 serves as the case estimate or the skill level of each student. Information obtained from the case estimate is that the SCORE-MAXSCR shows each respondent's score from the maximum score sequentially. Respondent 31 answered the majority of the questions (23 out of 24 questions) correctly compared to other respondents. The average estimate value and its standard deviation got 0.01 ± 1.35 and were in a moderate category. The analysis result of the case estimate revealed that students' skills were in the moderate category.

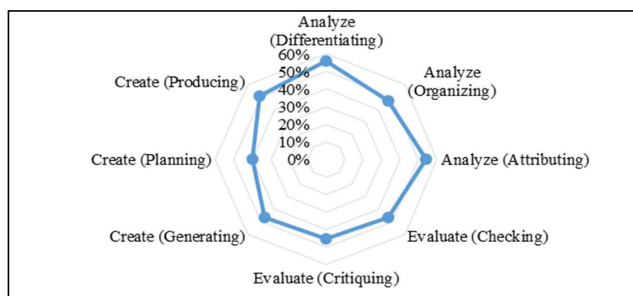


Figure 8
Distribution of Students' Answer Percentage HOTS

Figure 8 provides the percentage of students' answers based on the aspects and sub-aspects of HOTS. The analysis result pointed out that students tended to find it difficult to answer questions regarding the creating aspect, specifically the planning sub-aspect. Creating is the highest level of HOTS in Bloom's taxonomy; therefore, students need to practice developing their creating skills. This figure also signifies that most students find it easy to answer HOTS questions related to the analysis aspect, differentiating sub-aspect in particular.

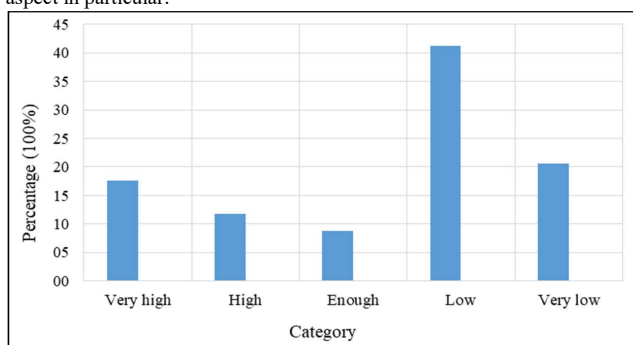


Figure 9
Percentage of Students' HOTS

The above figure shows the percentage of students' HOTS. It is seen that most students (41.2%) still have low HOTS; the categories consist of very low (20.6%), moderate (8.8%), high (11.8%), and very high (17.6%).

Commented [p4]: Please do not repeat the content of figure. However, the figure must be explained in detail about the tren or major finding

Author surnames go here

15

DISCUSSION

This study aims to produce the HOTS instrument presented in e-learning using Moodle LMS and determine the number of HOTS after using the instrument. The findings were valid and useable. The HOTS instrument validity was seen from the construct validity and face validity. Construct validity intends to investigate the HOTS instrument's accuracy and collect responses from experts and practitioners. Based on validator evaluation, the Aiken's V value was obtained from 0.76 to 1.00, suggesting a valid result. This result indicated that the HOTS instrument featured good material, design, and language aspects. The material aspect relates to the question items according to the indicators; has only one correct answer key; contents follow the calculation goal and the education level; the item distractors work properly. The construction feature of the HOTS instrument associates with the subject matter; has clearly-formulated answer choices; the subject matter does not lead to a correct answer; no multiple negative shapes; has homogeneous answer choices; has a similar length of answer choices; the items do not depend on each other; and the options are type. Next, it relates to the formulation of communicative language, grammatical sentences, non-multi-significant sentences, and standard/general/neutral vocabulary in the language aspect. Using Moodle LMS as a medium to serve HOTS instruments will promote the access of the students to online questions. E-learning using LMS Moodle is equipped with various facilities supporting online learning implementation that allows students to learn independently (Martín-Blas & Serrano-Fernández, 2009; Yildiz, Tezer, & Uzunboylu, 2018). Moodle LMS program presents an interesting display and is user-friendly (Martín-Blas & Serrano-Fernández, 2009). Students can work on the questions interactively and see the results directly.

Face validity in this analysis was obtained and evaluated based on students' HOTS instrument tests. Analyzing the HOTS instrument used IRT analysis methodology. It was suggested that all 24 items were fit as they reached the range of 0.77 to 1.30 in the MNSQ INFIT value, and -2.0 to 2.0 in the INFIT t. The item reliability value following the item estimate value summary measured at 0.66; meanwhile, the person's reliability based on the case estimate summary was 0.85 or very accurate (0.67 to 0.80). Thus, the instrument produced is appropriate for measuring students' HOTS as it has met the requirements according to the IRT analysis result.

The analysis result of students' HOTS obtained the average approximate value or skill level of each student, along with the standard deviation of 0.01 ± 1.35 (moderate category). The case estimate result indicated that the HOTS skills of the students were in the moderate category. The low category of students' HOTS was influenced by several factors, one of which was that the students were not used to working on HOTS questions (Tanujaya, Mumu, & Margono, 2017; Yusuf & Widyaningsih, 2019). They needed to practice developing their HOTS by being exposed to HOTS-based learning sources. To realize HOTS, students are required to be more active in learning (Winarti, Cari, Widha, & Istiyono, 2015; Yusuf & Widyaningsih, 2019). Lecturers are also expected to act as facilitators who provide various learning resources and provide feedback on the students'

tasks (Masruroh & Prasetyo, 2018). The use of e-learning allows students to access different learning resources in the form of texts, animations, simulations, multimedia, or virtual laboratories that can be accessed directly (Skultety, Gonzalez, & Vargas, 2017; Tee, Siti, Tengku, & Zainudin, 2013). It is expected that these e-learning facilities can facilitate students in learning so that their HOTS can be developed. Students' HOTS can also be improved through assignments and exercises in the learning process (Istiyono, Dwandaru, Megawati, & Ermansah, 2018; Yusuf & Widyaningsih, 2018). On this ground, it is of major importance to train the students' HOTS by applying learning technologies and quality instrument presentations through the IRT analysis.

CONCLUSION

The HOTS instrument presented by Moodle LMS in e-learning obtains a good performance. The IRT analysis, including item fit, reliability, and difficulty, acquires the mean and standard deviation parameters for INFIT MNSQ of 1.0 and 0.0; the items have proven to fit RM 1-PL. Additionally, test characteristics comprised item fitness, reliability, and difficulty. The trial result obtains the criteria of INFIT MNSQ mean and standard deviation of 1.0 and 0.0, respectively, implying that the items fit the RM1-PL. In addition, the value of item reliability based on the value of item estimate summary arrives at 0.66; meanwhile, the person reliability under the case estimate summary reaches 0.85, i.e., the reliability value is in the range of 0.67 - 0.80 (quite reliable). As based on the criteria of minimum and maximum INFIT MNSQ of 0.77 and 1.30, 24 question items fit the RM 1-PL model. The Quest output result also reveals that the average values of THRSHL and its standard deviation are 0.00 ± 0.71 , or in the acceptance range of -2 to 2. To sum up, all 24 question items that had been tried out have fit the model with a good category, so that they can be used in the HOTS measurement. Every student's average estimate or skill level along with the standard deviation is 0.01 ± 1.35 or in the moderate category. Students' HOTS must be practiced by providing HOTS-based learning resources.

ACKNOWLEDGMENT

We would like to acknowledge the contribution of the Ministry of Research and Higher Education in funding this study through the Inter-University Cooperation scheme with the contract number: 198/SP2H/AMD/LT/DRPM/2020.

REFERENCES

- Adams, R. J., & Khoo, S.-T. (1996). *Quest: the interactive test analysis system*. Camberwell, Vic.: Australian Council for Educational Research.
- Aiken, L. R. (1980). Content Validity and Reliability of Single Items or Questionnaires. *Educational and Psychological Measurement*, 40(4), 955–959.
- Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, 45(1), 131–142.
- Aldoobie, N. (2015). ADDIE Model. *American International Journal of Contemporary*

Author surnames go here

17

Research, 5(6), 72.

Azevedo, J. M. (2015). e-Assessment in mathematics courses with multiple-choice questions tests. *CSEDU 2015 - 7th International Conference on Computer Supported Education, Proceedings*, 2, 260–266. <https://doi.org/10.5220/0005452702600266>

Bogdanović, Z., Barać, D., Jovanić, B., Popović, S., & Radenković, B. (2014). Evaluation of Mobile Assessment in A Learning Management System. *British Journal of Educational Technology*, 45(2), 231–244. <https://doi.org/10.1111/bjet.12015>

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York: Routledge.

Brun, M., & Hinostroza, J. E. (2014). Learning to become a teacher in the 21st century: ICT integration in Initial Teacher Education in Chile. *Journal of Educational Technology & Society*, 17(3), 222–238.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting Potentially Biased Test Items: Comparison of IRT Area and Mantel-Haenszel Methods. *Applied Measurement in Education*, 2(4), 313–334. https://doi.org/10.1207/s15324818ame0204_4

Istiyono, E. (2017). The Analysis of Senior High School Students' Physics HOTS in Bantul District Measured using PhysReMChoTHOTS. *AIP Conference Proceedings*, 1868(August), 1–7. <https://doi.org/10.1063/1.4995184>

Istiyono, E. (2018). IT-based HOTS assessment on physics st learning as the 21 century demand at senior high schools: Expectation and reality IT-Based HOTS Assessment on Physics Learning as the 21 st Century Demand at Senior High Schools : Expectation and Reality. *AIP Conference Proceedings*, 2014(020014), 1–6.

Istiyono, E., Dwandaru, W. S. B., Megawati, I., & Ermansah. (2018). Application of Bloomian and Marzanoian Higher Order Thinking Skills in the Physics Learning Assessment: an Inevitability. *Advances in Social Science, Education and Humanities Research*, 164(ICLI 2017), 136–142. <https://doi.org/10.2991/icli-17.2018.26>

Istiyono, E., Dwandaru, W. S. B., & Muthmainah. (2019). Developing of Bloomian HOTS Physics Test: Content and Construct Validation of The PhysTeBloHOTS. *Journal of Physics: Conference Series*, 1397(012017), 1–9.

Kowsalya, D. N., Venkat Lakshmi, H., & Suresh, K. P. (2012). Development and Validation of a Scale to assess Self-Concept in Mild Intellectually Disabled Children. *International Journal of Social Sciences & Education*, 2(4).

Krathwohl, D. R., & Anderson, L. W. (2010). Merlin C. Wittrock and the Revision of Bloom's Taxonomy. *Educational Psychologist*, 45(1), 64–65. <https://doi.org/10.1080/00461520903433562>

Lee, M. F., & Zainal, N. A. (2017). Development of needham model based E-module for electromagnetic field & wave. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 120–124).

<https://doi.org/10.1109/IEEM.2017.8289863>

Limongelli, C., Sciarrone, F., & Vaste, G. (2011). Personalized e-learning in Moodle: the Moodle_LS System. *Journal of E-Learning and Knowledge Society*, 7(1), 49–58. Retrieved from <https://www.learntechlib.org/p/43340>

Martín-Blas, T., & Serrano-Fernández, A. (2009). The role of new technologies in the learning process: Moodle as a teaching tool in Physics. *Computers & Education*, 52(1), 35–44. <https://doi.org/10.1016/J.COMPEDU.2008.06.005>

Masruroh, A. N., & Prasetyo, Z. K. (2018). Effect of E-Module with Guided Inquiry Approach Containing Nature of Science to Student's Science Literacy. *E-Journal Pend. IPA*, 7(3), 165–171.

Pandey, S. R., & Pandey, S. (2009). Developing a More Effective and Flexible Learning Management System (LMS) for the Academic Institutions using Moodle. *ICAL 2009 - Technology, Policy and Innovation*, 249–254.

Raykov, T., & Marcoulides, G. A. (2015). On the Relationship Between Classical Test Theory and Item Response Theory: From One to the Other and Back. *Educational and Psychological Measurement*, 76(2), 325–338. <https://doi.org/10.1177/0013164415576958>

Skultety, L., Gonzalez, G., & Vargas, G. (2017). Using Technology to Support Teachers' Lesson Adaptations during Lesson Study. *Journal of Technology and Teacher Education*, 25(2), 185–213. Retrieved from <https://www.learntechlib.org/p/172139>

Tanujaya, B., Mumu, J., & Margono, G. (2017). The Relationship between Higher Order Thinking Skills and Academic Performance of Student in Mathematics Instruction. *International Education Studies*, 10(11), 78–85.

Tee, S. S., Siti, T., Tengku, M., & Zainudin, S. (2013). User Testing for Moodle Application. *International Journal of Software Engineering and Its Applications*, 7(5), 243–252.

Winarti, Cari, Widha, S., & Istiyono, E. (2015). Analysis of Higher Order Thinking Skills Content of Physics Examinations In Madrasah Aliyah. In *International Conference on Mathematics, Science, and Education 2015 (ICMSE 2015)* (Vol. 2015, pp. 32–38).

Yildiz, E. P., Tezer, M., & Uzunboylu, H. (2018). Student Opinion Scale Related to Moodle LMS in an Online Learning Environment: Validity and Reliability Study. *International Journal of Interactive Mobile Technologies (IJIM)*, 12(4), 97–108.

Yusuf, I., & Widyaningsih, S. W. (2018). Profil Kemampuan Mahasiswa dalam Menyelesaikan Soal HOTS di Jurusan Pendidikan Fisika Universitas Papua. *Jurnal Komunikasi Pendidikan*, 2(14), 42–49.

Yusuf, I., & Widyaningsih, S. W. (2019). HOTS profile of physics education students in

Author surnames go here

19

STEM-based classes using PhET media. *Journal of Physics: Conference Series*, 1157(032021), 1–5.

Yusuf, I., Widyaningsih, S. W., & Sebayang, S. R. B. (2018). Implementation of E-learning based-STEM on Quantum Physics Subject to Student HOTS Ability. *Turkish Science Education*, 15(December), 67–75.

Perbaikan oleh Penulis (Round 2)

The following changes have been made on the Manuscript “...” in accordance with reviewers’ comments

Reviewer’s comments	Changes made	Page (see highlights)
The study aimed, do not discussed	Adding the purpose of research, namely: This research aims to develop HOTS physics questions based on <i>Modern Test Theory</i> designed and presented with LMS Moodle on e-learning, which can be accessed online. This study also serves as one of the efforts to expand students’ HOTS by applying a variety of HOTS-based learning sources.	1
Poor figure. It must be revised	Image has been replaced with the best resolution	8 and 9

The Development of the HOTS Test of Physics Based on Modern Test Theory: Question Modeling through E-learning of Moodle LMS

This research aims to develop HOTS physics questions based on Modern Test Theory designed and presented with LMS Moodle on e-learning, which can be accessed online. This study also serves as one of the efforts to expand students' HOTS by applying a variety of HOTS-based learning sources. Further, this research employed the ADDIE model with analysis, design, development, implementation, and evaluation stages. The instrument consisted of 24 multiple-choice physics questions; the questions were designed by following the aspects and sub-aspects of HOTS and validated by the assessment of physics experts, physicists, and lecturers. Moreover, the validity analysis was based on Aiken's V formula, in which every aspect was confirmed to be valid. The instrument had been tested on 34 students of the Physics Education Department, Universitas Papua. Dichotomy data analysis used the Rasch Model (RM) 1-PL through the Quest program, and the test characteristics comprised item fitness, reliability, and difficulty. The trial result obtained the criteria of INFIT MNSQ mean and standard deviation of 1.0 and 0.0, respectively, indicating that the items fitted the RM1-PL. In addition, the value of item reliability based on the item estimate summary arrived at 0.66; meanwhile, the case reliability under the summary of the case estimate accounted for 0.85. The reliability value in the range of 0.67- 0.80 was categorized as quite reliable. Drawing upon the criteria of minimum and maximum INFIT MNSQ of 0.77 and 1.30, 24 question items fitted the RM 1-PL model. The Quest output result also suggested that the average values of Thresholds and its standard deviation were 0.00 ± 0.71 , or in the acceptance range of -2 to 2. Overall, all 24 question items that had been tested have fitted the model with a good category. They can be used in the HOTS measurement and can increase students' HOTS.

Keywords: E-learning, HOTS Test, and Modern Test Theory.

INTRODUCTION

Assessment, especially in the cognitive domain, is central to the learning process and should be carried out accurately and in compliance with the subject to be assessed or measured. Students' cognitive skills in the learning process can be categorized into Lower-Order Thinking Skills (LOTS) and Higher-Order Thinking Skills (HOTS). The LOTS includes remembering, understanding, and applying; the HOTS, on the other hand, consists of analyzing, evaluating, and creating. HOTS is thinking skills that require not only the remembering skill but also other higher skills. Indicators to measure HOTS encompass analyzing (C4), evaluating (C5), and creating (C6) skills (Krauthwohl & Anderson, 2010).

HOTS also refers to thinking skills when one takes new information, connects it with initial information s/he has, and finally delivers the information to achieve goals or answer questions (Istiyono, Dwandaru, & Muthmainah, 2019). This is in line with skill characteristics in the 21st century published by Partnership of 21st Century Skill stating that 21st-century learners should be able to develop competitive skills, such as critical

thinking, problem-solving, communication, information and communication technology (ICT) literacy, ICT, information literacy, and media literacy (Brun & Hinostroza, 2014); these focus on HOTS development.

Physics serves as part of science, comprising abstract concepts that are difficult to be directly described. Learning physics is expected to help students develop their thinking skills, in which they are not only demanded to master LOTS, but also HOTS. Teachers are also urged to deliver learning materials to students, including the HOTS, that can be improved by the HOTS instrument. A previous study has reported that the majority of teachers find it challenging to formulate an assessment instrument of learning outcomes, HOTS questions, in particular (Istiyono, 2018). For this reason, teachers' creativity is highly required to measure student learning outcomes. Today's development of ICT can be utilized to design and habituate students to learn anywhere at any time (Yusuf, Widyaningsih, & Sebayang, 2018). Relying on ICT during the learning process is one of the significant innovations, including the evaluation of student learning outcomes.

Evaluation questions can be posed in an integrated manner through e-learning systems, such as Moodle Learning Management System (LMS) (Azevedo, 2015; Bogdanović, Barać, Jovanić, Popović, & Radenković, 2014). The Moodle provides different types of questions, namely multiple choices, true or false, and short answers; these are stored in the taught course database and can be reapplied (Limongelli, Sciarrone, & Vaste, 2011). Teachers are also able to offer feedback directly to the students and give them correct answers to questions they have worked on (Pandey & Pandey, 2009). One of the advantages of an online evaluation through Moodle LMS is that students can figure out their assessment results right away.

Teachers need to prepare a good test to measure student learning outcomes. There are two paradigms developed to assess student learning outcomes through the used test, i.e., classical and modern approaches. The classical paradigm being utilized is classical test theory or widely known as classical true-score theory; meanwhile, the modern paradigm is item response theory (IRT). The classical test theory is selected due to its ease in the application despite its limitations in measuring the item difficulty level and discrimination since both indicators' calculation is based on the test taker's total score. In contrast, the IRT frees up the dependence between the test item and the test taker (a concept of parameter invariance); the test taker's response to a test item does not affect another item (a concept of local independence), and; the test item does only measure one measurement dimension (Raykov & Marcoulides, 2015). Therefore, the application answers the needs of modern measurement to date, i.e., comparing test taker's skills, question development, and even adaptive test development. It is considered able to overcome the limitations of the classical test theory.

On account of the simplicity of the analysis, most teachers have analyzed assessment tools using classical analysis techniques. The use of classical analytical techniques features some limitations, including the difficulty of defining individual learners' skills. The calculated error of measurement does not include persons but groups together. This is because each test taker's response to the questions cannot be clarified by classical test

theory. Efforts are thereby required to free the measuring tool from attachment to the sample (sample-free) employing the IRT.

This is a preliminary study with a long-term purpose of developing general physics questions with good quality at the Department of Physics Education, Universitas Papua. As the first stage, this study focuses on students at the department mentioned previously who enroll in General Physics subject taught by the researcher. This study also serves as one of the efforts to expand students' HOTS by applying a variety of HOTS-based learning sources. This research aims to develop HOTS physics questions based on IRT designed and presented with LMS Moodle on e-learning, which can be accessed online.

METHOD

As employed by this study, the ADDIE model refers to a general and systematic model of development study with a phased framework, allowing each element to connect (Aldoobie, 2015). The stages of this model used in the development of the HOTS instrument are presented in Figure 1.

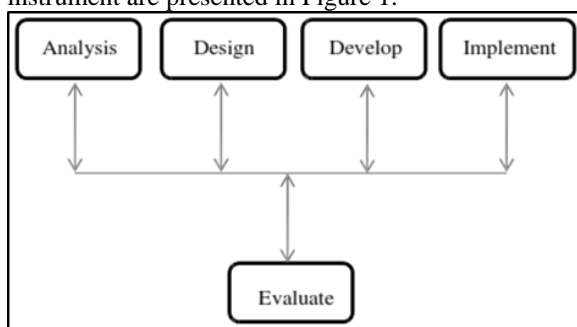


Figure 1
Stages of ADDIE Development Model in Designing Moodle LMS-based HOTS Test

Analysis

The analysis stage was a process of needs analysis to determine test objectives, identify problems, analyze tasks, and determine question formats to be applied. It was shown that the problems were related to the needs of HOTS instrument design for students at the Department of Physics Education, Universitas Papua.

Design

This stage comprised the process of designing HOTS questions to be used; the design process encompassed creating a question matrix and outline that covered question distribution in every aspect and sub-aspect of HOTS.

Develop

Every single thing required in the arrangement of HOTS questions has been prepared in the next stage. This stage also covered the process of making the questions regarding HOTS, as well as validating the questions that involved the experts of measurement,

physics education, and practitioners. The validity analysis technique to assess the content validity of the developed questions relied on the Aiken's V formula (Aiken, 1980, 1985).

$$V = \frac{\sum s}{n(c-1)} \quad (1)$$

"V" refers to the agreement index of validators in regards to item validity; "s" is the assessment score of validators subtracted by the assessment lowest score; "n" refers to the number of validators; "c" is the number of categories that can be chosen by validators. All test items are considered valid if the value of the Aiken's V index falls under the range of 0.37 to 1.00 (Kowsalya, Venkat Lakshmi, & Suresh, 2012). The value of Aiken's V of every test item was calculated based on the assessment items of every validator. In this stage, there was also an evaluation process, i.e., revising questions by following validators' corrections and suggestions.

Implementation

Another stage was applying HOTS questions that had been developed to 34 students in the site area who enrolled in general physics subject. This number followed the sample size for data stability in Rasch Model (RM) 1- PL, which is from 30 to 300, with the limit of INFIT t is from -2 to +2 (Bond, Yan, & Heene, 2020). Question item analysis was performed based on the raw score of the students by employing the Quest program.

Evaluation

The evaluation was a process of finding out whether HOTS's developed questions had met the expectation. The evaluation stage is carried out in every stage and is called a formative evaluation intended for revisions (Lee & Zainal, 2017). For instance, in the design stage, the expert's review is necessary to provide input towards the design. Besides, the evaluation stage was undertaken after analyzing empirical questions mathematically by using the Quest software program by referring to the Rasch model. The Quest program can do the Rasch measurement, i.e., a comprehensive empirical test of question items. There were three parameters being measured mathematically based on the empirical test of question items, as follows.

1. The first parameter is item fitness with the Rasch model by following the value of INFIT MNSQ or INFIT t of the item. The expected values of the unweighted mean square (Outfit MNSQ) in the Quest program and weighted mean square are 1; the variance is 0. On the contrary, the expected value of Mean INFIT t is equal to 0, with the variance equal to 1 (Adams & Khoo, 1996). The provision of INFIT MNSQ for the Rasch Model is presented in Table 1 and Table 2 below.

Table 1

Criteria of Question Item Fitness with the Rasch Model

MNSQ INFIT Value	Criteria
>1.33	Does Not Fit the Rasch Model
0.77 to 1.33	Fits the Rasch Model
<0.77	Does Not Fit the Rasch Model

Table 2

The Provision of Outfit t for the Rasch Model.

t OUTFIT Value	Criteria
OUTFIT $t \leq 2.00$	Fits the Rasch Model
OUTFIT $t \geq 2.00$	Does Not Fit the Rasch Model

2. The second parameter is reliability. The analysis result of the Quest program also showed the item and case reliability. The reliability value based on the item estimate is also called sample reliability; the higher the value, the more the items that fit the tested model. Whereas, the lower the value, the less the items that fit the tested model, so that it does not give the expected information. The reliability category is provided in the following table (Istiyono, 2017).

Table 3

Interpretation of Reliability Value

Reliability Value	Criteria
> 0.94	Excellent
$0.91 - 0.94$	Very Good
$0.81 - 0.90$	Good
$0.67 - 0.80$	Fair
< 0.67	Poor

3. The third parameter is the item difficulty index and respondents' skills presented as difficulty index in the Quest output. Thresholds (THRSHL) show the item difficulty index in the logit scale along with its standard deviation (Hambleton & Rogers, 1989). The provision of the THRSHL value is in Table 4.

Table 4

Criteria of THRSHL Value to Categorize Item Difficulty Level

THRSHL Value	Criteria
$b > 2.00$	Very Difficult
$1.00 < b \leq 2.00$	Difficult
$-1.00 < b \leq 1.00$	Medium
$-1.00 > b \geq 2.00$	Easy
$b < -2.00$	Very Easy

Respondents' skills were shown by the value of the estimate error, in which the criteria of the estimate value of respondents' skills are given in Table 5.

Table 5

Criteria of Estimate Value to Categorize Respondents' Skills

THRSHL Value	Criteria
$b > 2.00$	Very Difficult
$1.00 < b \leq 2.00$	Difficult
$-1,00 < b \leq 1.00$	Medium
$-1.00 > b \geq 2.00$	Easy
$b < -2.00$	Very Easy

The evaluation stage also included the process of analyzing the HOTS of students on the whole. The level of HOTS is categorized based on the ideal mean and standard deviation. This was applied with the assumption that students' HOTS of physics were normally distributed. The ideal mean (Im) and ideal standard deviation (Isd) are based

on the highest and lowest score of research variables. Table 6 shows the criteria of students' HOTS of physics.

Table 6
Criteria of Students' HOTS of Physics

Interval	Criteria
$Im + 1.5 Isb < \theta$	Very high
$Im + 0.5 Isb < \theta \leq Im + 1.5 Isb$	High
$Im - 0.5 Isb < \theta \leq Im + 0.5 Isb$	Moderate
$Im - 1.5 Isb < \theta \leq Im - 0.5 Isb$	Low
$0 < Im - 1.5 Isb$	Very Low

Meaning:

Im : ideal mean

Isb : ideal standard deviation

Xmak : highest score

Xmin : lowest score

RESULTS

The ADDIE development model can be used for different product developments in education, and one of which is the development of HOTS questions. This model is simple and systematically structured in its implementation stages. The following is a description of each stage result.

Analysis

A needs analysis was the first stage being done by observation and interview to gather any information required in physics learning at the Department of Physics Education, Universitas Papua. The researchers' experience indicated that the lecturers had applied HOTS learning in the classroom. However, a test to measure students' HOTS has not been conducted. The arrangement of HOTS instrument is required to train and develop students' HOTS. Accordingly, to facilitate the students in accessing other learning sources, this study designed HOTS questions in an online system through an e-learning program using the Moodle LMS.

Design

In the design stage, the test instrument was designed based on the analysis result in the first stage. The test instrument design was in the form of a question matrix and outline adjusted to students' needs and characteristics and learning sources. The test was in a multiple-choice format, in which 24 questions were adjusted to the formulation of a HOTS test that had been created in the test matrix and outline. The question matrix is provided in Table 7.

Table 7
The Question Matrix

Aspects	Sub Aspects	Theories		
		Electric current, Ohm's law, and electrical power	Series and parallel circuits of resistor and capacitor	Electric Force, Kirchoff's law, and RC circuit.
Analyze	Differentiating	8	12	21
	Organizing	3	15	20
	Attributing	2	9	23
Evaluate	Checking	4	11	22
	Critiquing	1	16	18
Create	Generating	5	13	19
	Planning	7	14	17
	Producing	6	10	24

Develop

The development of HOTS questions was based on the question matrix and outline that had been designed. In addition, the questions were formulated online through e-learning by utilizing the Moodle LMS. Figure 2 below shows all question items in the e-learning program.

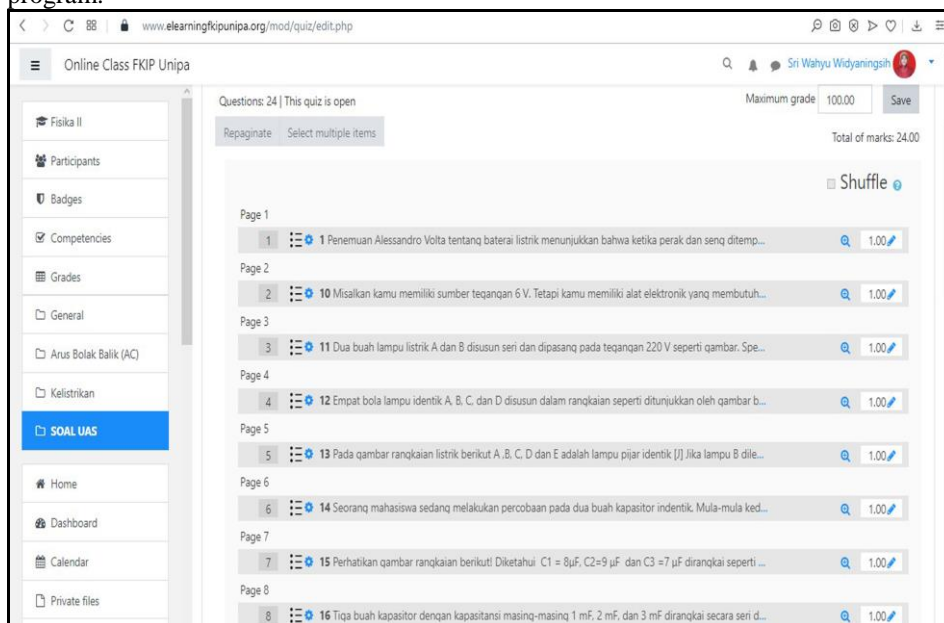


Figure 2

All Question Items in the E-Learning Program

The questions are displayed interactively, and students can randomly work on the questions. Moodle LMS can present questions with a picture or other contents to make it easier for teachers to design the questions as expected. Figure 3 illustrates one of the HOTS questions displayed on the e-learning through the Moodle LMS.

Question 18
Not yet answered
Marked out of 1.00
Flag question
Edit question

Seorang anak merancang suatu percobaan baterai sederhana yaitu mencelupkan batang karbon dan seng pada suatu larutan asam seperti gambar berikut:

Diagram showing a beaker labeled "Asam" containing two electrodes: "Elektroda Karbon" (Carbon Electrode) and "Elektroda Seng" (Zinc Electrode). Labels above the electrodes indicate "Terminal" and "Terminal".

Apabila kedua terminal batang tersebut dihubungkan dengan lampu LED maka lampu akan menyala dan lama kelamaan akan padam, hal ini disebabkan karena....

Select one:

- a. Semakin banyak ion seng yang terlarut menyebabkan elektroda atau yang lainnya terpakai habis menyebabkan sel mati
- b. Larutan asam melarutkan ion karbon lebih banyak
- c. Elektron-elektron terlepas dari elektroda seng sehingga elektroda seng bermuatan negatif dan akan habis
- d. Semakin banyak ion karbon yang memasuki larutan sehingga beda potensial sulit dipertahankan
- e. Elektron-elektron terlepas dari elektroda karbon menyebabkan elektroda seng bermuatan negatif dan lama kelamaan akan netral

Previous page Next page

Quiz navigation
1 2 3 4 5
6 7 8 9 10
11 12 13 14 15
16 17 18 19 20
21 22 23 24
Finish attempt ...
Start a new preview

Figure 3
HOTS Questions Displayed on the E-learning Through the Moodle LMS

The development stage aims to produce a HOTS test instrument that has been validated by experts and practitioners. Product validation is a process of assessing the designed product, or in this case, the test instrument of HOTS in general physics subject in the site area. Product validation was carried out by involving seven validators, i.e., experts of measurement, physics education, physics, and practitioners. The validity test of the instrument included material, construction, and language. The analysis result of the question validity assessed by validators obtained the value of Aiken's V in the range of 0.76 to 1.00, showing a valid result. The questions validated by experts and practitioners were then revised following the provided corrections and suggestions.

Implementation

The implementation stage in this study was the product trial, in which HOTS questions were tried out to 34 students in the research site. The students worked on these questions online through e-learning by using their own Moodle account upon completing all learning stages. Results of the students' learning can be accessed after this process.

Evaluation

Before conducting the estimate analysis of respondents' skills and item difficulty level, the analysis of item fitness was performed using INFIT and OUTFIT for mean square and t. The determination of the item fitness with the model is based on the value of INFIT MNSQ and the standard deviation or Infit t (Adams & Khoo, 1996). The fitness of each case is also based on the value of INFIT MNSQ or INFIT t of the item. Table 8 provides the testing result through the Quest program to obtain the values of item estimate and case estimate in the HOTS questions trial.

Table 8

Values of Item Estimate and Case Estimate in the HOTS Questions Trial

No	Measurement	Estimates for Items	Estimates for Testing
1.	Average values and standard deviations	0.00 ± 0.57	0.01 ± 1.24
2.	Reliability Estimates	0.66	0.85
3.	The mean and standard deviation of INFIT MNSQ	1.00 ± 0.14	0.99 ± 0.15
4.	The mean and standard deviation of OUTFIT MNSQ	1.09 ± 0.52	1.09 ± 0.52
5.	The mean and standard deviation of INFIT t	-0.03 ± 0.81	0.00 ± 0.72
6.	The mean and standard deviation of OUTFIT t	0.21 ± 0.91	0.17 ± 0.81

The analysis result suggested that the INFIT MNSQ got the range of 0.86 to 1.14, and INFIT t is -0.28 to 0.72. This signified that all 24 questions fit the model as they reached the range of INFIT MNSQ value from 0.77 to 1.30 and used INFIT t with the limit of -2.0 to 2.0. In addition to testing the fitness, the Quest program's output also presented the reliability estimate of the test instrument. The above table shows the value of item reliability based on the value of the item estimate summary, which is 0.66. On the other hand, the value of person reliability, as based on the case estimate summary, gets 0.85. These results were in line with the Rasch model, in which the reliability value fell under the range of 0.67 to 0.80 (quite reliable). On that ground, the instrument can be employed to measure students' HOTS in the General Physics subject.

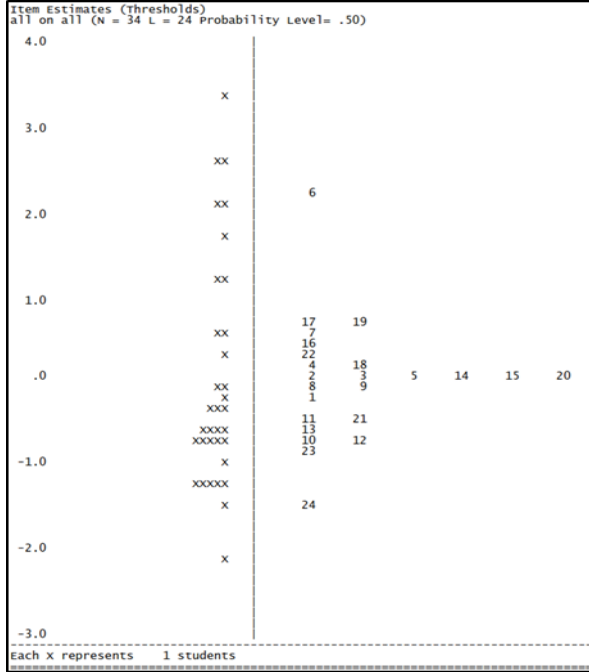


Figure 4
Distribution of Item Difficulty Level and Respondents' Skills

Figure 4 presents the distribution of the respondents according to the difficulty level in the logit scale from -4.0 to +4.0. This map displays the item difficulty level compared to the respondents' skills. Case and item difficulty levels in the Rasch model are expressed in one line in the form of abscissa in the graph with a log-odd unit. The graph of respondents' skills shows a normal curve, meaning that there are only a few respondents with low and high skills; and many respondents with moderate skills. The level of item difficulty of threshold revealed that item 6 was the most difficult question, and item 24 was the easiest one.

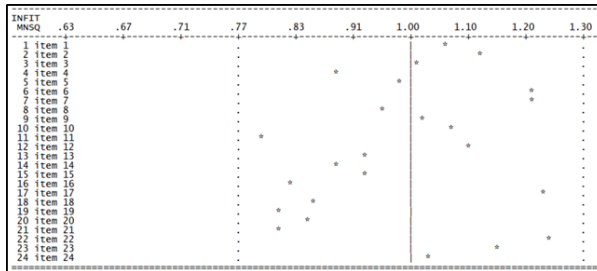


Figure 5
Distribution of INFIT MNSQ Values of Each Question Item of HOTS

Author surnames go here

11

Question items that fit the Rasch model are in the range of 0.77 to 1.33. By referring to Figure 5, we can see that all 24 question items are in the line, implying that they fit the Rasch model.

Item Estimates (Thresholds) In input Order							
all on all (N = 34 L = 24 Probability Level= .50)							
ITEM NAME	SCORE	MAXSCR	THRSH	INFT	OUTFT	INFT	OUTFT
			1	MNSQ	MNSQ	t	t
1 item 1	18	34	-.26 .39	1.06	1.15	.4	.5
2 item 2	16	34	.04 .40	1.12	1.17	.7	.6
3 item 3	16	34	.04 .40	1.01	.91	.1	-.2
4 item 4	15	34	.19 .40	.88	.93	-.6	-.1
5 item 5	16	34	.04 .40	.98	.89	.0	-.2
6 item 6	5	34	2.27 .57	1.21	2.16	.7	1.4
7 item 7	13	34	.52 .42	1.21	1.27	1.0	.9
8 item 8	17	34	-.11 .40	.96	1.00	-.2	.1
9 item 9	17	34	-.11 .40	1.02	.91	.2	-.2
10 item 10	21	34	-.70 .39	1.07	1.16	.6	.5
11 item 11	19	34	-.41 .39	.79	.66	-1.6	-.9
12 item 12	21	34	-.70 .39	1.10	1.14	.8	.5
13 item 13	20	34	-.55 .39	.93	1.09	-.5	.4
14 item 14	16	34	.04 .40	.88	.78	-.7	-.6
15 item 15	16	34	.04 .40	.93	.82	-.4	-.5
16 item 16	13	33	.47 .42	.82	.69	-.8	-.9
17 item 17	12	34	.69 .43	1.23	1.16	1.0	.6
18 item 18	15	34	.19 .40	.86	.73	-.8	-.8
19 item 19	12	34	.69 .43	.81	.71	-.8	-.8
20 item 20	16	34	.04 .40	.85	.75	-.9	-.7
21 item 21	19	34	-.41 .39	.81	.68	-1.4	-.8
22 item 22	14	34	.35 .41	1.24	1.23	1.2	.8
23 item 23	22	34	-.85 .40	1.15	3.04	1.1	3.1
24 item 24	26	34	-1.50 .43	1.03	1.02	.2	.2
Mean			.00	1.00	1.09	.0	.1
SD			.71	.14	.52	.8	.9

Figure 6
Item Estimates of HOTS Questions

The previous figure presents the Item Estimate of HOTS questions based on the trial result. In this figure, there is SCORE-MAXSCR successively showing the respondents who answer correctly and the number of total respondents. Item 24 was the most correctly-answered, in which 26 out of 34 respondents could work on this item. Figure 6

also provides the value of THRSHL that shows the item difficulty index in the logit scale along with its standard deviation. Item 6 got a THRSHL or difficulty index of 2.27 that was greater than 2.0, or in other words, this item was very difficult since only five students could give a correct answer. Also, the average value of THRSHL and its standard deviation accounted for 0.00 ± 0.71 and fell under the range of -2 to 2 (Hambleton & Rogers, 1989). The average value of INFIT MNSQ was 1.00 ± 0.14 and achieved the acceptance range of 0.77 to 1.33; the average value of OUTFIT t arrived at 0.10 ± 0.90 and was included in the acceptance range of ≤ 2.00 . Accordingly, these results indicate that all question items being developed can be utilized to measure students' HOTS.

Case Estimates In Input Order all on all (N = 34 L = 24 Probability Level= .50)							
NAME	SCORE MAXSCR	ESTIMATE	ERROR	INFIT MNSQ	OUTFIT MNSQ	INFIT t	OUTFIT t
1 01	12 24	-.02	.43	1.06	1.02	.58	.18
2 02	6 24	-1.21	.49	1.17	1.09	.75	.36
3 03	8 24	-.77	.45	.98	1.01	-.06	.13
4 04	8 24	-.77	.45	.83	.81	-1.04	-.48
5 05	8 24	-.77	.45	.89	.83	-.59	-.41
6 06	6 24	-1.21	.49	.79	.70	-.84	-.67
7 07	10 24	-.38	.43	.99	.95	-.01	-.09
8 08	6 24	-1.21	.49	1.07	2.30	.36	2.44
9 09	3 24	-2.10	.63	.98	.85	.11	.00
10 10	9 24	-.57	.44	.88	.83	-.80	-.48
11 11	22 24	2.61	.77	.73	.46	-.30	-.56
12 12	5 24	-1.46	.52	.89	.85	-.29	-.18
13 13	20 24	1.75	.57	1.21	1.45	.64	.93
14 14	11 24	-.20	.43	.86	.83	-1.33	-.59
15 15	21 24	2.12	.64	1.18	1.05	.52	.29
16 16	9 24	-.57	.44	1.08	1.06	.59	.28
17 17	7 24	-.98	.47	1.29	2.20	1.38	2.55
18 18	6 24	-1.21	.49	1.23	1.28	.96	.75
19 19	14 24	.35	.43	.92	.87	-.56	-.40
20 20	15 24	-.54	.44	.97	1.09	-.13	-.40
21 21	18 24	1.19	.49	.94	.86	-.16	-.25
22 22	21 24	2.12	.64	.93	1.23	-.01	-.55
23 23	9 24	-.57	.44	1.07	1.01	.54	.15
24 24	8 24	-.77	.45	1.01	.95	.13	-.03
25 25	10 24	-.38	.43	.87	.82	-1.06	-.57
26 26	15 24	-.54	.44	1.05	1.22	.36	.80
27 27	6 24	-1.21	.49	.82	.74	-.69	-.56
28 28	22 24	2.61	.77	.73	.46	-.30	-.56
29 29	12 24	-.02	.43	.92	.88	-.73	-.39
30 30	9 24	-.57	.44	.90	.90	-.64	-.23
31 31	23 24	3.40	1.05	1.18	3.14	.49	1.53
32 32	18 24	1.19	.49	.85	.75	-.53	-.58
33 33	10 23	-.32	.44	1.11	1.11	.97	.45
34 34	8 24	-.77	.45	1.29	1.34	1.61	1.03
Mean		.01		.99	1.09	.00	.17
SD		1.35		.15	.52	.72	.81

Figure 7
Case Estimates of Every Student

Figure 7 serves as the case estimate or the skill level of each student. Information obtained from the case estimate is that the SCORE-MAXSCR shows each respondent's score from the maximum score sequentially. Respondent 31 answered the majority of the questions (23 out of 24 questions) correctly compared to other respondents. The average estimate value and its standard deviation got 0.01 ± 1.35 and were in a moderate category. The analysis result of the case estimate revealed that students' skills were in the moderate category.

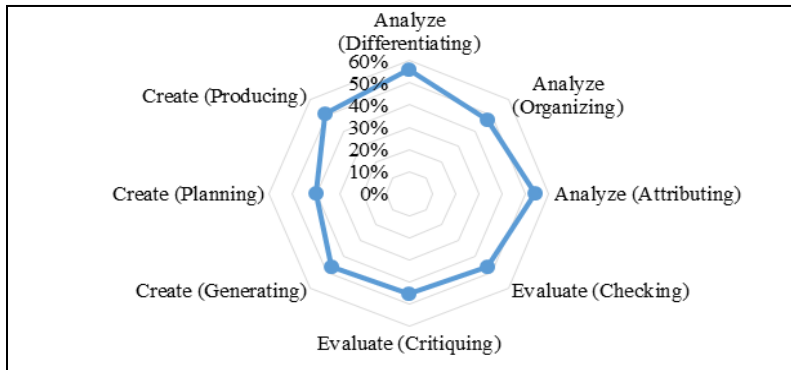


Figure 8
Distribution of Students' Answer Percentage HOTS

Figure 8 provides the percentage of students' answers based on the aspects and sub-aspects of HOTS. The analysis result pointed out that students tended to find it difficult to answer questions regarding the creating aspect, specifically the planning sub-aspect. Creating is the highest level of HOTS in Bloom's taxonomy; therefore, students need to practice developing their creating skills. This figure also signifies that most students find it easy to answer HOTS questions related to the analysis aspect, differentiating sub-aspect in particular.

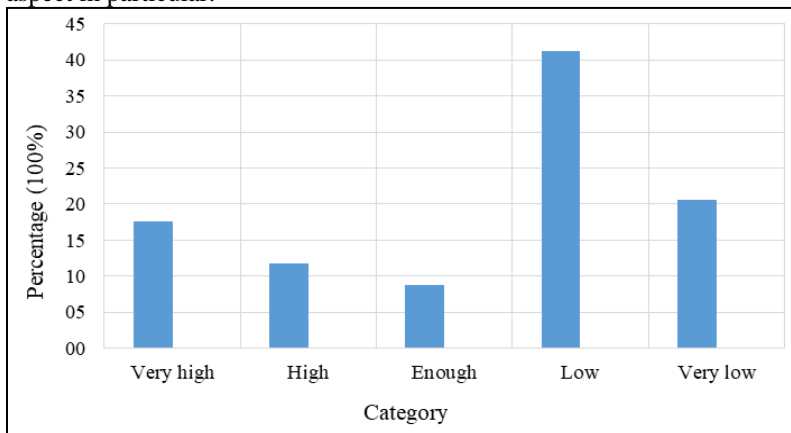


Figure 9
Percentage of Students' HOTS

The above figure shows the percentage of students' HOTS. It is seen that most students (41.2%) still have low HOTS; the categories consist of very low (20.6%), moderate (8.8%), high (11.8%), and very high (17.6%).

DISCUSSION

This study aims to produce the HOTS instrument presented in e-learning using Moodle LMS and determine the number of HOTS after using the instrument. The findings were valid and useable. The HOTS instrument validity was seen from the construct validity and face validity. Construct validity intends to investigate the HOTS instrument's accuracy and collect responses from experts and practitioners. Based on validator evaluation, the Aiken's V value was obtained from 0.76 to 1.00, suggesting a valid result. This result indicated that the HOTS instrument featured good material, design, and language aspects. The material aspect relates to the question items according to the indicators; has only one correct answer key; contents follow the calculation goal and the education level; the item distractors work properly. The construction feature of the HOTS instrument associates with the subject matter; has clearly-formulated answer choices; the subject matter does not lead to a correct answer; no multiple negative shapes; has homogeneous answer choices; has a similar length of answer choices; the items do not depend on each other; and the options are type. Next, it relates to the formulation of communicative language, grammatical sentences, non-multi-significant sentences, and standard/general/neutral vocabulary in the language aspect. Using Moodle LMS as a medium to serve HOTS instruments will promote the access of the students to online questions. E-learning using LMS Moodle is equipped with various facilities supporting online learning implementation that allows students to learn independently (Martín-Blas & Serrano-Fernández, 2009; Yildiz, Tezer, & Uzunboylu, 2018). Moodle LMS program presents an interesting display and is user-friendly (Martín-Blas & Serrano-Fernández, 2009). Students can work on the questions interactively and see the results directly.

Face validity in this analysis was obtained and evaluated based on students' HOTS instrument tests. Analyzing the HOTS instrument used IRT analysis methodology. It was suggested that all 24 items were fit as they reached the range of 0.77 to 1.30 in the MNSQ INFIT value, and -2.0 to 2.0 in the INFIT t. The item reliability value following the item estimate value summary measured at 0.66; meanwhile, the person's reliability based on the case estimate summary was 0.85 or very accurate (0.67 to 0.80). Thus, the instrument produced is appropriate for measuring students' HOTS as it has met the requirements according to the IRT analysis result.

The analysis result of students' HOTS obtained the average approximate value or skill level of each student, along with the standard deviation of 0.01 ± 1.35 (moderate category). The case estimate result indicated that the HOTS skills of the students were in the moderate category. The low category of students' HOTS was influenced by several factors, one of which was that the students were not used to working on HOTS questions (Tanujaya, Mumu, & Margono, 2017; Yusuf & Widyaningsih, 2019). They needed to practice developing their HOTS by being exposed to HOTS-based learning sources. To realize HOTS, students are required to be more active in learning (Winarti, Cari, Widha, & Istiyono, 2015; Yusuf & Widyaningsih, 2019). Lecturers are also expected to act as facilitators who provide various learning resources and provide feedback on the students' tasks (Masruroh & Prasetyo, 2018). The use of e-learning allows students to

access different learning resources in the form of texts, animations, simulations, multimedia, or virtual laboratories that can be accessed directly (Skultety, Gonzalez, & Vargas, 2017; Tee, Siti, Tengku, & Zainudin, 2013). It is expected that these e-learning facilities can facilitate students in learning so that their HOTS can be developed. Students' HOTS can also be improved through assignments and exercises in the learning process (Istiyono, Dwandaru, Megawati, & Ermansah, 2018; Yusuf & Widyaningsih, 2018). On this ground, it is of major importance to train the students' HOTS by applying learning technologies and quality instrument presentations through the IRT analysis.

CONCLUSION

The HOTS instrument presented by Moodle LMS in e-learning obtains a good performance. The IRT analysis, including item fit, reliability, and difficulty, acquires the mean and standard deviation parameters for INFIT MNSQ of 1.0 and 0.0; the items have proven to fit RM 1-PL. Additionally, test characteristics comprised item fitness, reliability, and difficulty. The trial result obtains the criteria of INFIT MNSQ mean and standard deviation of 1.0 and 0.0, respectively, implying that the items fit the RM1-PL. In addition, the value of item reliability based on the value of item estimate summary arrives at 0.66; meanwhile, the person reliability under the case estimate summary reaches 0.85, i.e., the reliability value is in the range of 0.67 - 0.80 (quite reliable). As based on the criteria of minimum and maximum INFIT MNSQ of 0.77 and 1.30, 24 question items fit the RM 1-PL model. The Quest output result also reveals that the average values of THRSHL and its standard deviation are 0.00 ± 0.71 , or in the acceptance range of -2 to 2. To sum up, all 24 question items that had been tried out have fit the model with a good category, so that they can be used in the HOTS measurement. Every student's average estimate or skill level along with the standard deviation is 0.01 ± 1.35 or in the moderate category. Students' HOTS must be practiced by providing HOTS-based learning resources.

ACKNOWLEDGMENT

We would like to acknowledge the contribution of the Ministry of Research and Higher Education in funding this study through the Inter-University Cooperation scheme with the contract number: 198/SP2H/AMD/LT/DRPM/2020.

REFERENCES

- Adams, R. J., & Khoo, S.-T. (1996). *Quest: the interactive test analysis system*. Camberwell, Vic.: Australian Council for Educational Research.
- Aiken, L. R. (1980). Content Validity and Reliability of Single Items or Questionnaires. *Educational and Psychological Measurement*, 40(4), 955–959.
- Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, 45(1), 131–142.
- Aldoobie, N. (2015). ADDIE Model. *American International Journal of Contemporary Research*, 5(6), 72.

- Azevedo, J. M. (2015). e-Assessment in mathematics courses with multiple-choice questions tests. *CSEDU 2015 - 7th International Conference on Computer Supported Education, Proceedings*, 2, 260–266. <https://doi.org/10.5220/0005452702600266>
- Bogdanović, Z., Barać, D., Jovanić, B., Popović, S., & Radenković, B. (2014). Evaluation of Mobile Assessment in A Learning Management System. *British Journal of Educational Technology*, 45(2), 231–244. <https://doi.org/10.1111/bjet.12015>
- Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York: Routledge.
- Brun, M., & Hinostroza, J. E. (2014). Learning to become a teacher in the 21st century: ICT integration in Initial Teacher Education in Chile. *Journal of Educational Technology & Society*, 17(3), 222–238.
- Hambleton, R. K., & Rogers, H. J. (1989). Detecting Potentially Biased Test Items: Comparison of IRT Area and Mantel-Haenszel Methods. *Applied Measurement in Education*, 2(4), 313–334. https://doi.org/10.1207/s15324818ame0204_4
- Istiyono, E. (2017). The Analysis of Senior High School Students' Physics HOTS in Bantul District Measured using PhysRemChoTHOTS. *AIP Conference Proceedings*, 1868(August), 1–7. <https://doi.org/10.1063/1.4995184>
- Istiyono, E. (2018). IT-based HOTS assessment on physics st learning as the 21 century demand at senior high schools: Expectation and reality IT-Based HOTS Assessment on Physics Learning as the 21 st Century Demand at Senior High Schools : Expectation and Reality. *AIP Conference Proceedings*, 2014(020014), 1–6.
- Istiyono, E., Dwandaru, W. S. B., Megawati, I., & Ermansah. (2018). Application of Bloomian and Marzanoian Higher Order Thinking Skills in the Physics Learning Assessment: an Inevitability. *Advances in Social Science, Education and Humanities Research*, 164(ICLI 2017), 136–142. <https://doi.org/10.2991/icli-17.2018.26>
- Istiyono, E., Dwandaru, W. S. B., & Muthmainah. (2019). Developing of Bloomian HOTS Physics Test: Content and Construct Validation of The PhysTeBloHOTS. *Journal of Physics: Conference Series*, 1397(012017), 1–9.
- Kowsalya, D. N., Venkat Lakshmi, H., & Suresh, K. P. (2012). Development and Validation of a Scale to assess Self-Concept in Mild Intellectually Disabled Children. *International Journal of Social Sciences & Education*, 2(4).
- Krathwohl, D. R., & Anderson, L. W. (2010). Merlin C. Wittrock and the Revision of Bloom's Taxonomy. *Educational Psychologist*, 45(1), 64–65. <https://doi.org/10.1080/00461520903433562>
- Lee, M. F., & Zainal, N. A. (2017). Development of needham model based E-module for electromagnetic field & wave. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 120–124). <https://doi.org/10.1109/IEEM.2017.8289863>

Author surnames go here

17

Limongelli, C., Sciarrone, F., & Vaste, G. (2011). Personalized e-learning in Moodle: the Moodle_LS System. *Journal of E-Learning and Knowledge Society*, 7(1), 49–58. Retrieved from <https://www.learntechlib.org/p/43340>

Martín-Blas, T., & Serrano-Fernández, A. (2009). The role of new technologies in the learning process: Moodle as a teaching tool in Physics. *Computers & Education*, 52(1), 35–44. <https://doi.org/10.1016/J.COMPEDU.2008.06.005>

Masrurroh, A. N., & Prasetyo, Z. K. (2018). Effect of E-Module with Guided Inquiry Approach Containing Nature of Science to Student's Science Literacy. *E-Journal Pend. IPA*, 7(3), 165–171.

Pandey, S. R., & Pandey, S. (2009). Developing a More Effective and Flexible Learning Management System (LMS) for the Academic Institutions using Moodle. *ICAL 2009 - Technology, Policy and Innovation*, 249–254.

Raykov, T., & Marcoulides, G. A. (2015). On the Relationship Between Classical Test Theory and Item Response Theory: From One to the Other and Back. *Educational and Psychological Measurement*, 76(2), 325–338. <https://doi.org/10.1177/0013164415576958>

Skultety, L., Gonzalez, G., & Vargas, G. (2017). Using Technology to Support Teachers' Lesson Adaptations during Lesson Study. *Journal of Technology and Teacher Education*, 25(2), 185–213. Retrieved from <https://www.learntechlib.org/p/172139>

Tanujaya, B., Mumu, J., & Margono, G. (2017). The Relationship between Higher Order Thinking Skills and Academic Performance of Student in Mathematics Instruction. *International Education Studies*, 10(11), 78–85.

Tee, S. S., Siti, T., Tengku, M., & Zainudin, S. (2013). User Testing for Moodle Application. *International Journal of Software Engineering and Its Applications*, 7(5), 243–252.

Winarti, Cari, Widha, S., & Istiyono, E. (2015). Analysis of Higher Order Thinking Skills Content of Physics Examinations In Madrasah Aliyah. In *International Conference on Mathematics, Science, and Education 2015 (ICMSE 2015)* (Vol. 2015, pp. 32–38).

Yildiz, E. P., Tezer, M., & Uzunboylu, H. (2018). Student Opinion Scale Related to Moodle LMS in an Online Learning Environment: Validity and Reliability Study. *International Journal of Interactive Mobile Technologies (IJIM)*, 12(4), 97–108.

Yusuf, I., & Widyaningsih, S. W. (2018). Profil Kemampuan Mahasiswa dalam Menyelesaikan Soal HOTS di Jurusan Pendidikan Fisika Universitas Papua. *Jurnal Komunikasi Pendidikan*, 2(14), 42–49.

Yusuf, I., & Widyaningsih, S. W. (2019). HOTS profile of physics education students in STEM-based classes using PhET media. *Journal of Physics: Conference Series*, 1157(032021), 1–5.

Yusuf, I., Widyaningsih, S. W., & Sebayang, S. R. B. (2018). Implementation of E-learning based-STEM on Quantum Physics Subject to Student HOTS Ability. *Turkish Science Education*, 15(December), 67–75.



The Development of the HOTS Test of Physics Based on Modern Test Theory: Question Modeling through E-learning of Moodle LMS

Sri Wahyu Widyaningsih

Asst. Prof., corresponding author, Faculty of Teacher Training and Education, Universitas Papua, Indonesia, s.widyaningsih@unipa.ac.id

Irfan Yusuf

Asst. Prof., Faculty of Teacher Training and Education, Universitas Papua, Indonesia, i.yusuf@unipa.ac.id

Zuhdan Kun Prasetyo

Prof., Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Indonesia, zuhdan@uny.ac.id

Edi Istiyono

Prof., Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Indonesia, edi_istiyono@uny.ac.id

This research aims to develop HOTS physics questions based on Modern Test Theory designed and presented with LMS Moodle on e-learning, which can be accessed online. This study also serves as one of the efforts to expand students' HOTS by applying a variety of HOTS-based learning sources. Further, this research employed the ADDIE model with analysis, design, development, implementation, and evaluation stages. The instrument consisted of 24 multiple-choice physics questions; the questions were designed by following the aspects and sub-aspects of HOTS and validated by the assessment of physics experts, physicists, and lecturers. Moreover, the validity analysis was based on Aiken's V formula, in which every aspect was confirmed to be valid. The instrument had been tested on 34 students of the Physics Education Department, Universitas Papua. Dichotomy data analysis used the Rasch Model (RM) 1-PL through the Quest program, and the test characteristics comprised item fitness, reliability, and difficulty. The trial result obtained the criteria of INFIT MNSQ mean and standard deviation of 1.0 and 0.0, respectively, indicating that the items fitted the RM1-PL. In addition, the value of item reliability based on the item estimate summary arrived at 0.66; meanwhile, the case reliability under the summary of the case estimate accounted for 0.85.

Keywords: e-learning, HOTS test, modern test theory, physics, test

Citation: Widyaningsih, S. W., Yusuf, I., Prasetyo, Z. K., & Istiyono, E. (2021). The development of the HOTS test of physics based on modern test theory: Question Modeling through e-learning of moodle LMS. *International Journal of Instruction*, 14(4), 51-68.

INTRODUCTION

Assessment, especially in the cognitive domain, is central to the learning process and should be carried out accurately and in compliance with the subject to be assessed or measured. Students' cognitive skills in the learning process can be categorized into Lower-Order Thinking Skills (LOTS) and Higher-Order Thinking Skills (HOTS). The LOTS includes remembering, understanding, and applying; the HOTS, on the other hand, consists of analyzing, evaluating, and creating. HOTS is thinking skills that require not only the remembering skill but also other higher skills. Indicators to measure HOTS encompass analyzing (C4), evaluating (C5), and creating (C6) skills (Krathwohl & Anderson, 2010).

HOTS also refers to thinking skills when one takes new information, connects it with initial information s/he has, and finally delivers the information to achieve goals or answer questions (Istiyono, Dwandaru, & Muthmainah, 2019). This is in line with skill characteristics in the 21st century published by Partnership of 21st Century Skill stating that 21st-century learners should be able to develop competitive skills, such as critical thinking, problem-solving, communication, information and communication technology (ICT) literacy, ICT, information literacy, and media literacy (Brun & Hinostroza, 2014); these focus on HOTS development.

Physics serves as part of science, comprising abstract concepts that are difficult to be directly described. Learning physics is expected to help students develop their thinking skills, in which they are not only demanded to master LOTS, but also HOTS. Teachers are also urged to deliver learning materials to students, including the HOTS, that can be improved by the HOTS instrument. A previous study has reported that the majority of teachers find it challenging to formulate an assessment instrument of learning outcomes, HOTS questions, in particular (Istiyono, 2018). For this reason, teachers' creativity is highly required to measure student learning outcomes. Today's development of ICT can be utilized to design and habituate students to learn anywhere at any time (Yusuf, Widyaningsih, & Sebayang, 2018). Relying on ICT during the learning process is one of the significant innovations, including the evaluation of student learning outcomes.

Evaluation questions can be posed in an integrated manner through e-learning systems, such as Moodle Learning Management System (LMS) (Azevedo, 2015; Bogdanović, Barać, Jovanić, Popović, & Radenković, 2014). The Moodle provides different types of questions, namely multiple choices, true or false, and short answers; these are stored in the taught course database and can be reapplied (Limongelli, Sciarrone, & Vaste, 2011). Teachers are also able to offer feedback directly to the students and give them correct answers to questions they have worked on (Pandey & Pandey, 2009). One of the advantages of an online evaluation through Moodle LMS is that students can figure out their assessment results right away.

Teachers need to prepare a good test to measure student learning outcomes. There are two paradigms developed to assess student learning outcomes through the used test, i.e., classical and modern approaches. The classical paradigm being utilized is classical test theory or widely known as classical true-score theory; meanwhile, the modern paradigm

is item response theory (IRT). The classical test theory is selected due to its ease in the application despite its limitations in measuring the item difficulty level and discrimination since both indicators' calculation is based on the test taker's total score. In contrast, the IRT frees up the dependence between the test item and the test taker (a concept of parameter invariance); the test taker's response to a test item does not affect another item (a concept of local independence), and; the test item does only measure one measurement dimension (Raykov & Marcoulides, 2015). Therefore, the application answers the needs of modern measurement to date, i.e., comparing test taker's skills, question development, and even adaptive test development. It is considered able to overcome the limitations of the classical test theory.

On account of the simplicity of the analysis, most teachers have analyzed assessment tools using classical analysis techniques. The use of classical analytical techniques features some limitations, including the difficulty of defining individual learners' skills. The calculated error of measurement does not include persons but groups together. This is because each test taker's response to the questions cannot be clarified by classical test theory. Efforts are thereby required to free the measuring tool from attachment to the sample (sample-free) employing the IRT.

This is a preliminary study with a long-term purpose of developing general physics questions with good quality at the Department of Physics Education, Universitas Papua. As the first stage, this study focuses on students at the department mentioned previously who enroll in General Physics subject taught by the researcher. This study also serves as one of the efforts to expand students' HOTS by applying a variety of HOTS-based learning sources. This research aims to develop HOTS physics questions based on IRT designed and presented with LMS Moodle on e-learning, which can be accessed online.

METHOD

As employed by this study, the ADDIE model refers to a general and systematic model of development study with a phased framework, allowing each element to connect (Aldoobie, 2015). The stages of this model used in the development of the HOTS instrument are presented in Figure 1.

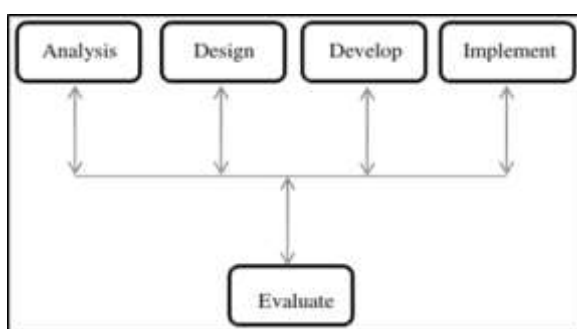


Figure 1
Stages of ADDIE development model in designing moodle LMS-based HOTS test

Analysis

The analysis stage was a process of needs analysis to determine test objectives, identify problems, analyze tasks, and determine question formats to be applied. It was shown that the problems were related to the needs of HOTS instrument design for students at the Department of Physics Education, Universitas Papua.

Design

This stage comprised the process of designing HOTS questions to be used; the design process encompassed creating a question matrix and outline that covered question distribution in every aspect and sub-aspect of HOTS.

Develop

Every single thing required in the arrangement of HOTS questions has been prepared in the next stage. This stage also covered the process of making the questions regarding HOTS, as well as validating the questions that involved the experts of measurement, physics education, and practitioners. The validity analysis technique to assess the content validity of the developed questions relied on the Aiken's V formula (Aiken, 1980, 1985).

$$V = \frac{\sum s}{n(c-1)} \quad (1)$$

"V" refers to the agreement index of validators in regards to item validity; "s" is the assessment score of validators subtracted by the assessment lowest score; "n" refers to the number of validators; "c" is the number of categories that can be chosen by validators. All test items are considered valid if the value of the Aiken's V index falls under the range of 0.37 to 1.00 (Kowsalya, Venkat Lakshmi, & Suresh, 2012). The value of Aiken's V of every test item was calculated based on the assessment items of every validator. In this stage, there was also an evaluation process, i.e., revising questions by following validators' corrections and suggestions.

Implementation

Another stage was applying HOTS questions that had been developed to 34 students in the site area who enrolled in general physics subject. This number followed the sample size for data stability in Rasch Model (RM) 1- PL, which is from 30 to 300, with the limit of INFIT t is from -2 to +2 (Bond, Yan, & Heene, 2020). Question item analysis was performed based on the raw score of the students by employing the Quest program.

Evaluation

The evaluation was a process of finding out whether HOTS's developed questions had met the expectation. The evaluation stage is carried out in every stage and is called a formative evaluation intended for revisions (Lee & Zainal, 2017). For instance, in the design stage, the expert's review is necessary to provide input towards the design. Besides, the evaluation stage was undertaken after analyzing empirical questions mathematically by using the Quest software program by referring to the Rasch model. The Quest program can do the Rasch measurement, i.e., a comprehensive empirical test

of question items. There were three parameters being measured mathematically based on the empirical test of question items, as follows.

1. The first parameter is item fitness with the Rasch model by following the value of INFIT MNSQ or INFIT t of the item. The expected values of the unweighted mean square (Outfit MNSQ) in the Quest program and weighted mean square are 1; the variance is 0. On the contrary, the expected value of Mean INFIT t is equal to 0, with the variance equal to 1 (Adams & Khoo, 1996). The provision of INFIT MNSQ for the Rasch Model is presented in Table 1 and Table 2 below.

Table 1

Criteria of question item fitness with the rasch model

MNSQ INFIT Value	Criteria
>1.33	Does Not Fit the Rasch Model
0.77 to 1.33	Fits the Rasch Model
<0.77	Does Not Fit the Rasch Model

Table 2

The provision of outfit t for the rasch model.

t OUTFIT Value	Criteria
OUTFIT $t \leq 2.00$	Fits the Rasch Model
OUTFIT $t \geq 2.00$	Does Not Fit the Rasch Model

2. The second parameter is reliability. The analysis result of the Quest program also showed the item and case reliability. The reliability value based on the item estimate is also called sample reliability; the higher the value, the more the items that fit the tested model. Whereas, the lower the value, the less the items that fit the tested model, so that it does not give the expected information. The reliability category is provided in the following table (Istiyono, 2017).

Table 3

Interpretation of reliability value

Reliability Value	Criteria
> 0.94	Excellent
0.91 – 0.94	Very Good
0.81 – 0.90	Good
0.67 – 0.80	Fair
< 0.67	Poor

3. The third parameter is the item difficulty index and respondents' skills presented as difficulty index in the Quest output. Thresholds (THRSHL) show the item difficulty index in the logit scale along with its standard deviation (Hambleton & Rogers, 1989). The provision of the THRSHL value is in Table 4.

Table 4
Criteria of THRSHL value to categorize item difficulty level

THRSHL Value	Criteria
$b > 2.00$	Very Difficult
$1.00 < b \leq 2.00$	Difficult
$-1.00 < b \leq 1.00$	Medium
$-1.00 > b \geq 2.00$	Easy
$b < -2.00$	Very Easy

Respondents' skills were shown by the value of the estimate error, in which the criteria of the estimate value of respondents' skills are given in Table 5.

Table 5
Criteria of estimate value to categorize respondents' skills

THRSHL Value	Criteria
$b > 2.00$	Very Difficult
$1.00 < b \leq 2.00$	Difficult
$-1.00 < b \leq 1.00$	Medium
$-1.00 > b \geq 2.00$	Easy
$b < -2.00$	Very Easy

The evaluation stage also included the process of analyzing the HOTS of students on the whole. The level of HOTS is categorized based on the ideal mean and standard deviation. This was applied with the assumption that students' HOTS of physics were normally distributed. The ideal mean (I_m) and ideal standard deviation (I_{sd}) are based on the highest and lowest score of research variables. Table 6 shows the criteria of students' HOTS of physics.

Table 6
Criteria of students' HOTS of physics

Interval	Criteria
$I_m + 1.5 I_{sb} < \theta$	Very high
$I_m + 0.5 I_{sb} < \theta \leq I_m + 1.5 I_{sb}$	High
$I_m - 0.5 I_{sb} < \theta \leq I_m + 0.5 I_{sb}$	Moderate
$I_m - 1.5 I_{sb} < \theta \leq I_m - 0.5 I_{sb}$	Low
$0 < I_m - 1.5 I_{sb}$	Very Low

Meaning:

I_m : ideal mean

I_{sb} : ideal standard deviation

X_{mak} : highest score

X_{min} : lowest score

FINDINGS

The ADDIE development model can be used for different product developments in education, and one of which is the development of HOTS questions. This model is simple and systematically structured in its implementation stages. The following is a description of each stage result.

Analysis

A needs analysis was the first stage being done by observation and interview to gather any information required in physics learning at the Department of Physics Education, Universitas Papua. The researchers' experience indicated that the lecturers had applied HOTS learning in the classroom. However, a test to measure students' HOTS has not been conducted. The arrangement of HOTS instrument is required to train and develop students' HOTS. Accordingly, to facilitate the students in accessing other learning sources, this study designed HOTS questions in an online system through an e-learning program using the Moodle LMS.

Design

In the design stage, the test instrument was designed based on the analysis result in the first stage. The test instrument design was in the form of a question matrix and outline adjusted to students' needs and characteristics and learning sources. The test was in a multiple-choice format, in which 24 questions were adjusted to the formulation of a HOTS test that had been created in the test matrix and outline. The question matrix is provided in Table 7.

Table 7
The question matrix

Aspects	Sub Aspects	Theories		
		Electric current, Ohm's law, and electrical power	Series and parallel circuits of resistor and capacitor	Electric Force, Kirchoff's law, and RC circuit.
Analyze	Differentiating	8	12	21
	Organizing	3	15	20
	Attributing	2	9	23
Evaluate	Checking	4	11	22
	Critiquing	1	16	18
Create	Generating	5	13	19
	Planning	7	14	17
	Producing	6	10	24

Develop

The development of HOTS questions was based on the question matrix and outline that had been designed. In addition, the questions were formulated online through e-learning by utilizing the Moodle LMS. Figure 2 below shows all question items in the e-learning program.

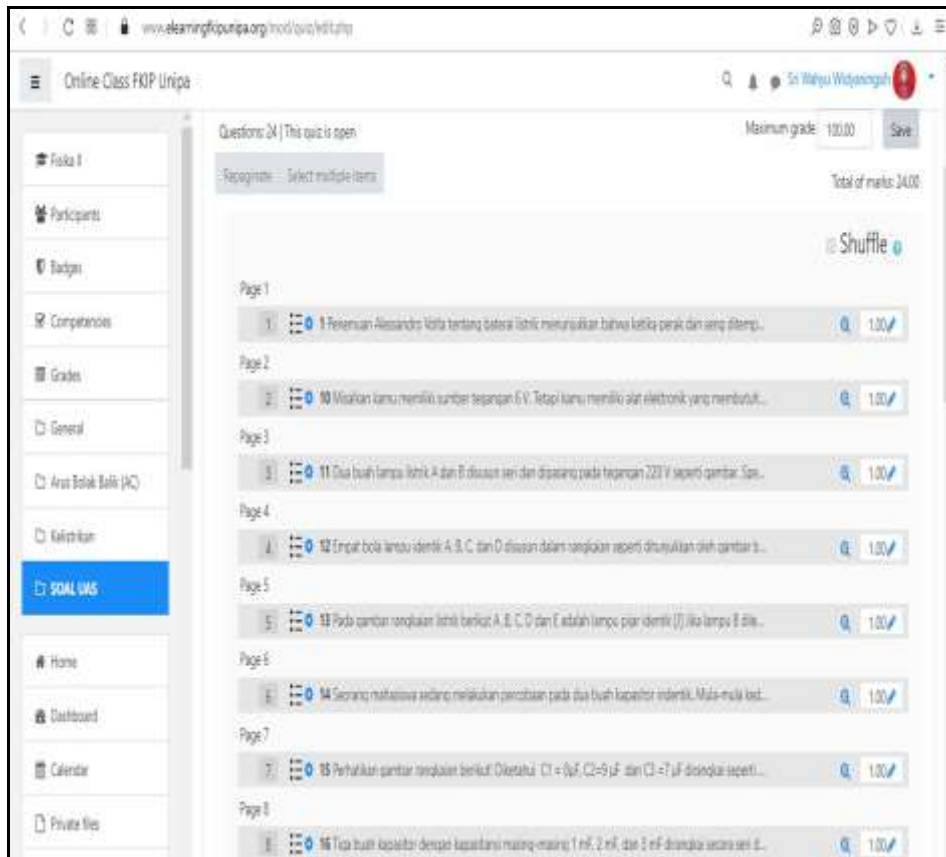


Figure 2
All question items in the e-learning program

The questions are displayed interactively, and students can randomly work on the questions. Moodle LMS can present questions with a picture or other contents to make it easier for teachers to design the questions as expected. Figure 3 illustrates one of the HOTS questions displayed on the e-learning through the Moodle LMS.

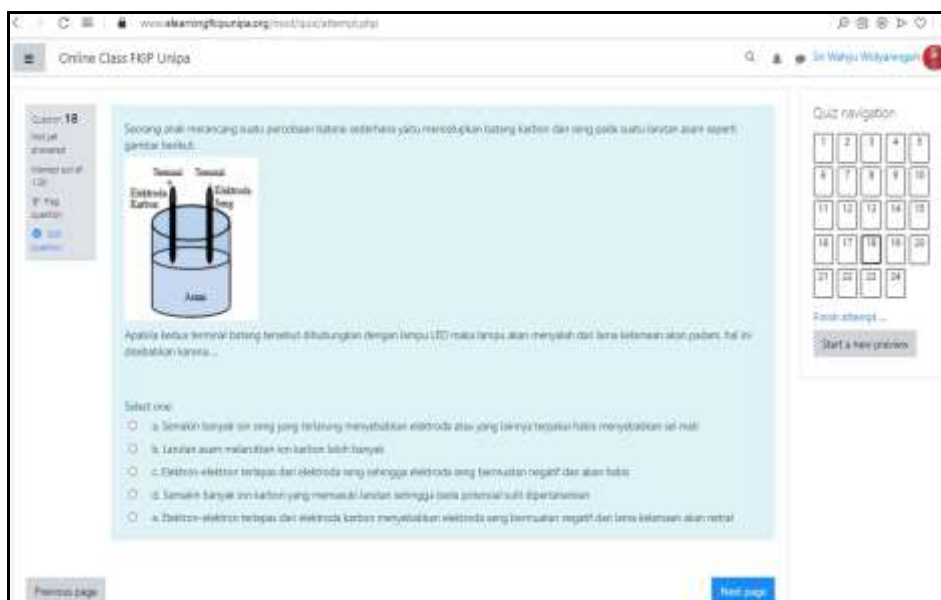


Figure 3
HOTS questions displayed on the e-learning through the moodle LMS

The development stage aims to produce a HOTS test instrument that has been validated by experts and practitioners. Product validation is a process of assessing the designed product, or in this case, the test instrument of HOTS in general physics subject in the site area. Product validation was carried out by involving seven validators, i.e., experts of measurement, physics education, physics, and practitioners. The validity test of the instrument included material, construction, and language. The analysis result of the question validity assessed by validators obtained the value of Aiken's V in the range of 0.76 to 1.00, showing a valid result. The questions validated by experts and practitioners were then revised following the provided corrections and suggestions.

Implementation

The implementation stage in this study was the product trial, in which HOTS questions were tried out to 34 students in the research site. The students worked on these questions online through e-learning by using their own Moodle account upon completing all learning stages. Results of the students' learning can be accessed after this process.

Evaluation

Before conducting the estimate analysis of respondents' skills and item difficulty level, the analysis of item fitness was performed using INFIT and OUTFIT for mean square and t . The determination of the item fitness with the model is based on the value of INFIT MNSQ and the standard deviation or Infit t (Adams & Khoo, 1996). The fitness of each case is also based on the value of INFIT MNSQ or INFIT t of the item. Table 8

provides the testing result through the Quest program to obtain the values of item estimate and case estimate in the HOTS questions trial.

Table 8
Values of item estimate and case estimate in the HOTS questions trial

No	Measurement	Estimates for Items	Estimates for Testing
1.	Average values and standard deviations	0.00 ± 0.57	0.01 ± 1.24
2.	Reliability Estimates	0.66	0.85
3.	The mean and standard deviation of INFIT MNSQ	1.00 ± 0.14	0.99 ± 0.15
4.	The mean and standard deviation of OUTFIT MNSQ	1.09 ± 0.52	1.09 ± 0.52
5.	The mean and standard deviation of INFIT t	-0.03 ± 0.81	0.00 ± 0.72
6.	The mean and standard deviation of OUTFIT t	0.21 ± 0.91	0.17 ± 0.81

The analysis result suggested that the INFIT MNSQ got the range of 0.86 to 1.14, and INFIT t is -0.28 to 0.72. This signified that all 24 questions fit the model as they reached the range of INFIT MNSQ value from 0.77 to 1.30 and used INFIT t with the limit of -2.0 to 2.0. In addition to testing the fitness, the Quest program's output also presented the reliability estimate of the test instrument. The above table shows the value of item reliability based on the value of the item estimate summary, which is 0.66. On the other hand, the value of person reliability, as based on the case estimate summary, gets 0.85. These results were in line with the Rasch model, in which the reliability value fell under the range of 0.67 to 0.80 (quite reliable). On that ground, the instrument can be employed to measure students' HOTS in the General Physics subject.

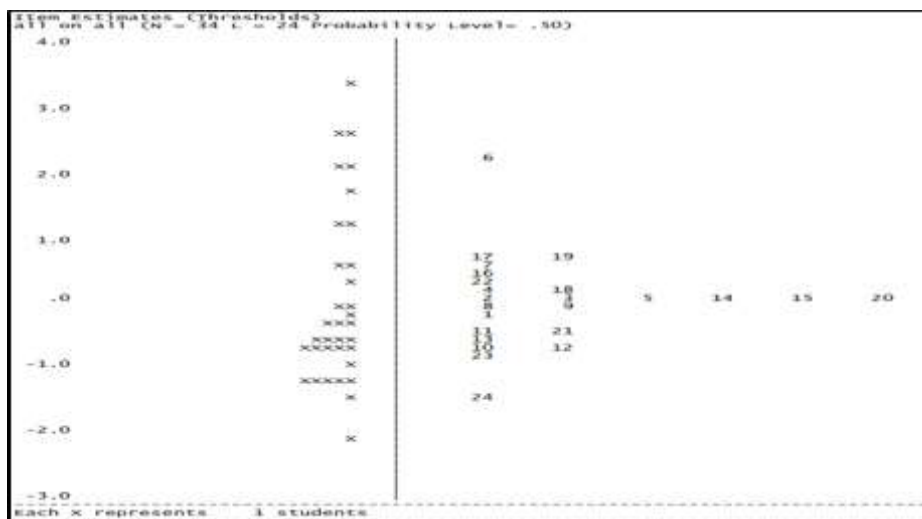


Figure 4
Distribution of item difficulty level and respondents' skills

Figure 4 presents the distribution of the respondents according to the difficulty level in the logit scale from -4.0 to +4.0. This map displays the item difficulty level compared to the respondents' skills. Case and item difficulty levels in the Rasch model are expressed in one line in the form of abscissa in the graph with a log-odd unit. The graph of respondents' skills shows a normal curve, meaning that there are only a few respondents with low and high skills; and many respondents with moderate skills. The level of item difficulty of threshold revealed that item 6 was the most difficult question, and item 24 was the easiest one.

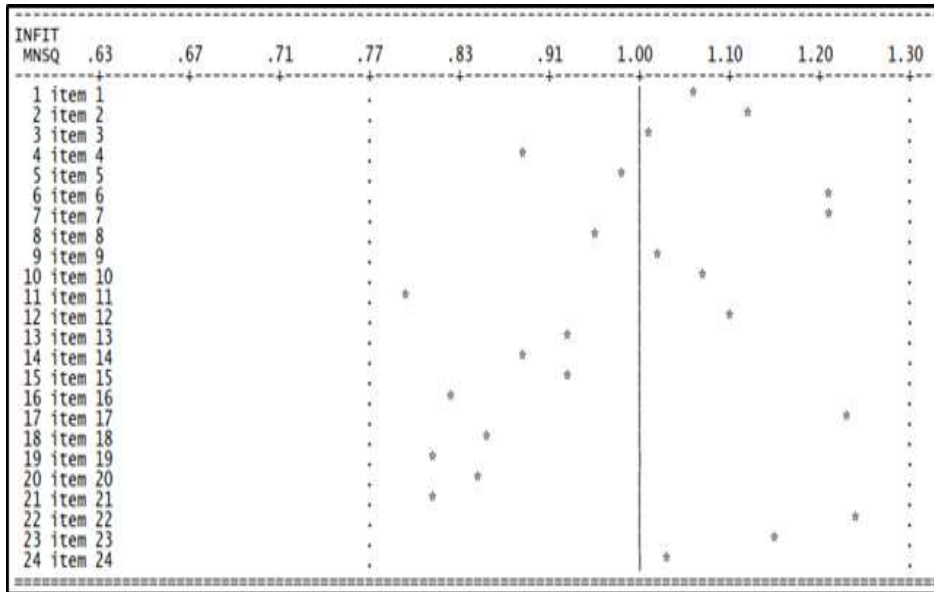


Figure 5
Distribution of INFIT MNSQ values of each question item of HOTS

Question items that fit the Rasch model are in the range of 0.77 to 1.33. By referring to Figure 5, we can see that all 24 question items are in the line, implying that they fit the Rasch model.

Item Estimates (Thresholds) in Input Order all on all (N = 34 L = 24 Probability Level = .50)									
ITEM	ITEM NAME	SCORE	MAXSCR	THRSHL	INFT MNSQ	OUTFIT MNSQ	INFT T	OUTFIT T	
1	Item 1	18	34	-.26 .39	1.06	1.15	.4	.5	
2	Item 2	16	34	.04 .40	1.12	1.17	.7	.6	
3	Item 3	16	34	-.04 .40	1.01	.91	-.1	-.2	
4	Item 4	15	34	-.19 .40	.88	.93	-.6	-.1	
5	Item 5	16	34	.04 .40	.98	.89	.0	-.2	
6	Item 6	5	34	2.27 .57	1.21	2.16	.7	1.4	
7	Item 7	13	34	-.52 .42	1.21	1.27	1.0	.9	
8	Item 8	17	34	-.11 .40	.96	1.00	-.2	-.1	
9	Item 9	17	34	-.11 .40	1.02	.91	.2	-.2	
10	Item 10	21	34	-.70 .39	1.07	1.16	.6	.5	
11	Item 11	19	34	-.41 .39	.79	.66	-1.6	-.9	
12	Item 12	21	34	-.70 .39	1.10	1.14	.8	.5	
13	Item 13	20	34	-.55 .39	.93	1.09	-.5	.4	
14	Item 14	16	34	.04 .40	.88	.78	-.7	-.6	
15	Item 15	16	34	.04 .40	.93	.82	-.4	-.5	
16	Item 16	13	33	.47 .42	.82	.69	-.8	-.9	
17	Item 17	12	34	.69 .43	1.23	1.16	1.0	.6	
18	Item 18	15	34	-.19 .40	.86	.73	-.8	-.8	
19	Item 19	12	34	.69 .43	.81	.71	-.8	-.8	
20	Item 20	16	34	-.04 .40	.85	.75	-.9	-.7	
21	Item 21	19	34	-.41 .39	.81	.68	-1.4	-.8	
22	Item 22	14	34	-.35 .41	1.24	1.23	1.2	.8	
23	Item 23	22	34	-.85 .40	1.15	3.04	1.1	3.1	
24	Item 24	26	34	-1.50 .43	1.03	1.02	.2	.2	
Mean				.00	1.00	1.09	.0	.1	
SD				.71	.14	.52	.8	.9	

Figure 6
Item estimates of HOTS questions

The previous figure presents the Item Estimate of HOTS questions based on the trial result. In this figure, there is SCORE-MAXSCR successively showing the respondents who answer correctly and the number of total respondents. Item 24 was the most correctly-answered, in which 26 out of 34 respondents could work on this item. Figure 6 also provides the value of THRSHL that shows the item difficulty index in the logit scale along with its standard deviation. Item 6 got a THRSHL or difficulty index of 2.27 that was greater than 2.0, or in other words, this item was very difficult since only five students could give a correct answer. Also, the average value of THRSHL and its standard deviation accounted for 0.00 ± 0.71 and fell under the range of -2 to 2 (Hambleton & Rogers, 1989). The average value of INFIT MNSQ was 1.00 ± 0.14 and achieved the acceptance range of 0.77 to 1.33; the average value of OUTFIT t arrived at 0.10 ± 0.90 and was included in the acceptance range of ≤ 2.00 . Accordingly, these results indicate that all question items being developed can be utilized to measure students' HOTS.

Case Estimates In input Order
all on all (N = 34 L = 24 Probability Level= .50)

NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFIT MNSQ	OUTFIT MNSQ	INFT t	OUTFT t
1 01	12	24	-.02	.43	1.06	1.02	.58	.18
2 02	5	24	-1.21	.49	1.17	1.09	.75	.36
3 03	8	24	-.77	.45	.98	1.01	-.06	.13
4 04	8	24	-.77	.45	.83	.81	-1.04	-.48
5 05	8	24	-.77	.45	.89	.83	-.59	-.41
6 06	6	24	-1.21	.49	.79	.70	-.84	-.67
7 07	10	24	-.38	.43	.99	.95	-.01	-.09
8 08	6	24	-1.21	.49	1.07	2.30	.36	2.44
9 09	3	24	-2.10	.63	.98	.85	.11	.00
10 10	9	24	-.57	.44	.88	.83	-.80	-.48
11 11	22	24	2.61	.77	.73	.46	-.30	-.56
12 12	5	24	-1.46	.52	.89	.85	-.79	-.18
13 13	20	24	1.75	.57	1.21	1.45	.64	.93
14 14	11	24	-.20	.43	.86	.83	-1.33	-.59
15 15	21	24	2.12	.64	1.18	1.05	.52	.29
16 16	7	24	-.57	.44	1.08	1.06	.59	.28
17 17	9	24	-.57	.44	1.29	2.20	1.38	2.28
18 18	6	24	-1.21	.49	1.23	1.28	.36	.72
19 19	14	24	-.35	.43	.92	.87	-.56	-.40
20 20	15	24	-.54	.44	.97	1.09	-.13	-.40
21 21	18	24	1.19	.49	.94	.86	-.16	-.23
22 22	21	24	2.12	.64	.93	1.23	-.01	.52
23 23	9	24	-.57	.44	1.07	1.01	.54	.13
24 24	8	24	-.77	.45	1.01	.95	.13	-.03
25 25	10	24	-.38	.43	.87	.82	-1.06	-.57
26 26	15	24	-.54	.44	1.05	1.22	.36	.80
27 27	6	24	-1.21	.49	.82	.74	-.69	-.56
28 28	22	24	2.61	.77	.73	.46	-.30	-.56
29 29	12	24	-.02	.43	.92	.88	-.73	-.39
30 30	9	24	-.57	.44	.90	.90	-.64	-.23
31 31	23	24	3.40	1.05	1.18	3.14	.49	1.53
32 32	18	24	1.19	.49	.85	.75	-.53	-.58
33 33	10	23	-.32	.44	1.11	1.11	.97	.43
34 34	8	24	-.77	.45	1.29	1.34	1.61	1.03
Mean			.01		.99	1.09	.00	.17
SD			1.35		.15	.52	.72	.81

Figure 7
Case estimates of every student

Figure 7 serves as the case estimate or the skill level of each student. Information obtained from the case estimate is that the SCORE-MAXSCR shows each respondent's score from the maximum score sequentially. Respondent 31 answered the majority of the questions (23 out of 24 questions) correctly compared to other respondents. The average estimate value and its standard deviation got 0.01 ± 1.35 and were in a moderate category. The analysis result of the case estimate revealed that students' skills were in the moderate category.

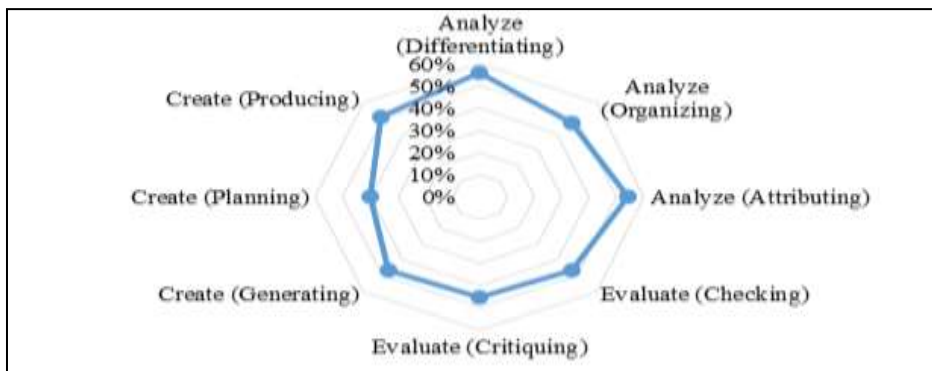


Figure 8
Distribution of students' answer percentage HOTS

Figure 8 provides the percentage of students' answers based on the aspects and sub-aspects of HOTS. The analysis result pointed out that students tended to find it difficult to answer questions regarding the creating aspect, specifically the planning sub-aspect. Creating is the highest level of HOTS in Bloom's taxonomy; therefore, students need to practice developing their creating skills. This figure also signifies that most students find it easy to answer HOTS questions related to the analysis aspect, differentiating sub-aspect in particular.

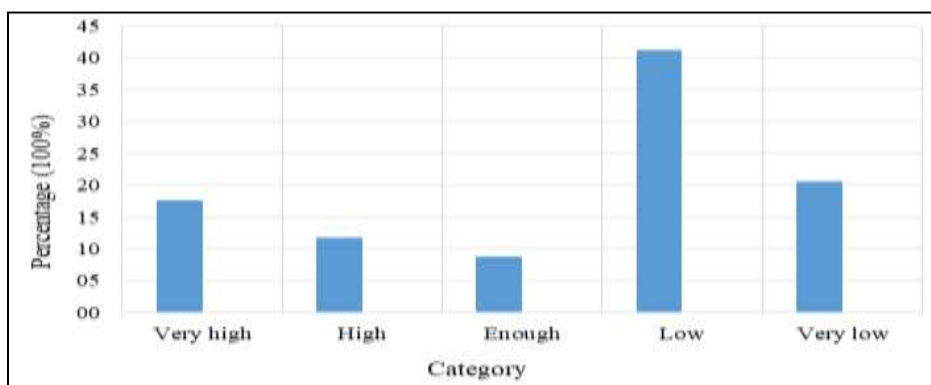


Figure 9
Percentage of students' HOTS

The above figure shows the percentage of students' HOTS. It is seen that most students (41.2%) still have low HOTS; the categories consist of very low (20.6%), moderate (8.8%), high (11.8%), and very high (17.6%).

DISCUSSION

This study aims to produce the HOTS instrument presented in e-learning using Moodle LMS and determine the number of HOTS after using the instrument. The findings were valid and useable. The HOTS instrument validity was seen from the construct validity and face validity. Construct validity intends to investigate the HOTS instrument's accuracy and collect responses from experts and practitioners. Based on validator evaluation, the Aiken's V value was obtained from 0.76 to 1.00, suggesting a valid result. This result indicated that the HOTS instrument featured good material, design, and language aspects. The material aspect relates to the question items according to the indicators; has only one correct answer key; contents follow the calculation goal and the education level; the item distractors work properly. The construction feature of the HOTS instrument associates with the subject matter; has clearly-formulated answer choices; the subject matter does not lead to a correct answer; no multiple negative shapes; has homogeneous answer choices; has a similar length of answer choices; the items do not depend on each other; and the options are type. Next, it relates to the formulation of communicative language, grammatical sentences, non-multi-significant sentences, and standard/general/neutral vocabulary in the language aspect. Using Moodle LMS as a medium to serve HOTS instruments will promote the access of the

students to online questions. E-learning using LMS Moodle is equipped with various facilities supporting online learning implementation that allows students to learn independently (Martín-Blas & Serrano-Fernández, 2009; Yildiz, Tezer, & Uzunboylu, 2018). Moodle LMS program presents an interesting display and is user-friendly (Martín-Blas & Serrano-Fernández, 2009). Students can work on the questions interactively and see the results directly.

Face validity in this analysis was obtained and evaluated based on students' HOTS instrument tests. Analyzing the HOTS instrument used IRT analysis methodology. It was suggested that all 24 items were fit as they reached the range of 0.77 to 1.30 in the MNSQ INFIT value, and -2.0 to 2.0 in the INFIT t. The item reliability value following the item estimate value summary measured at 0.66; meanwhile, the person's reliability based on the case estimate summary was 0.85 or very accurate (0.67 to 0.80). Thus, the instrument produced is appropriate for measuring students' HOTS as it has met the requirements according to the IRT analysis result.

The analysis result of students' HOTS obtained the average approximate value or skill level of each student, along with the standard deviation of 0.01 ± 1.35 (moderate category). The case estimate result indicated that the HOTS skills of the students were in the moderate category. The low category of students' HOTS was influenced by several factors, one of which was that the students were not used to working on HOTS questions (Tanujaya, Mumu, & Margono, 2017; Yusuf & Widyaningsih, 2019). They needed to practice developing their HOTS by being exposed to HOTS-based learning sources. To realize HOTS, students are required to be more active in learning (Winarti, Cari, Widha, & Istiyono, 2015; Yusuf & Widyaningsih, 2019). Lecturers are also expected to act as facilitators who provide various learning resources and provide feedback on the students' tasks (Masruroh & Prasetyo, 2018). The use of e-learning allows students to access different learning resources in the form of texts, animations, simulations, multimedia, or virtual laboratories that can be accessed directly (Skultety, Gonzalez, & Vargas, 2017; Tee, Siti, Tengku, & Zainudin, 2013). It is expected that these e-learning facilities can facilitate students in learning so that their HOTS can be developed. Students' HOTS can also be improved through assignments and exercises in the learning process (Istiyono, Dwandaru, Megawati, & Ermansah, 2018; Yusuf & Widyaningsih, 2018). On this ground, it is of major importance to train the students' HOTS by applying learning technologies and quality instrument presentations through the IRT analysis.

CONCLUSION

The HOTS instrument presented by Moodle LMS in e-learning obtains a good performance. The IRT analysis, including item fit, reliability, and difficulty, acquires the mean and standard deviation parameters for INFIT MNSQ of 1.0 and 0.0; the items have proven to fit RM 1-PL. Additionally, test characteristics comprised item fitness, reliability, and difficulty. The trial result obtains the criteria of INFIT MNSQ mean and standard deviation of 1.0 and 0.0, respectively, implying that the items fit the RM1-PL. In addition, the value of item reliability based on the value of item estimate summary arrives at 0.66; meanwhile, the person reliability under the case estimate summary reaches 0.85, i.e., the reliability value is in the range of 0.67 - 0.80 (quite reliable). As

based on the criteria of minimum and maximum INFIT MNSQ of 0.77 and 1.30, 24 question items fit the RM 1-PL model. The Quest output result also reveals that the average values of THRSHL and its standard deviation are 0.00 ± 0.71 , or in the acceptance range of -2 to 2. To sum up, all 24 question items that had been tried out have fit the model with a good category, so that they can be used in the HOTS measurement. Every student's average estimate or skill level along with the standard deviation is 0.01 ± 1.35 or in the moderate category. Students' HOTS must be practiced by providing HOTS-based learning resources.

ACKNOWLEDGMENT

We would like to acknowledge the contribution of the Ministry of Research and Higher Education in funding this study through the Inter-University Cooperation scheme with the contract number: 198/SP2H/AMD/LT/DRPM/2020.

REFERENCES

- Adams, R. J., & Khoo, S.-T. (1996). *Quest: the interactive test analysis system*. Camberwell, Vic.: Australian Council for Educational Research.
- Aiken, L. R. (1980). Content Validity and Reliability of Single Items or Questionnaires. *Educational and Psychological Measurement*, 40(4), 955–959.
- Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, 45(1), 131–142.
- Aldoobie, N. (2015). ADDIE Model. *American International Journal of Contemporary Research*, 5(6), 72.
- Azevedo, J. M. (2015). e-Assessment in mathematics courses with multiple-choice questions tests. *CSEDU 2015 - 7th International Conference on Computer Supported Education, Proceedings*, 2, 260–266. <https://doi.org/10.5220/0005452702600266>
- Bogdanović, Z., Barać, D., Jovanić, B., Popović, S., & Radenković, B. (2014). Evaluation of Mobile Assessment in A Learning Management System. *British Journal of Educational Technology*, 45(2), 231–244. <https://doi.org/10.1111/bjet.12015>
- Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York: Routledge.
- Brun, M., & Hinostroza, J. E. (2014). Learning to become a teacher in the 21st century: ICT integration in Initial Teacher Education in Chile. *Journal of Educational Technology & Society*, 17(3), 222–238.
- Hambleton, R. K., & Rogers, H. J. (1989). Detecting Potentially Biased Test Items: Comparison of IRT Area and Mantel-Haenszel Methods. *Applied Measurement in Education*, 2(4), 313–334. https://doi.org/10.1207/s15324818ame0204_4
- Istiyono, E. (2017). The Analysis of Senior High School Students' Physics HOTS in Bantul District Measured using PhysRemChoTHOTS. *AIP Conference Proceedings*, 1868(August), 1–7. <https://doi.org/10.1063/1.4995184>

- Istiyono, E. (2018). IT-based HOTS assessment on physics st learning as the 21 century demand at senior high schools: Expectation and reality IT-Based HOTS Assessment on Physics Learning as the 21 st Century Demand at Senior High Schools : Expectation and Reality. *AIP Conference Proceedings*, 2014(020014), 1–6.
- Istiyono, E., Dwandaru, W. S. B., Megawati, I., & Ermansah. (2018). Application of Bloomian and Marzanoian Higher Order Thinking Skills in the Physics Learning Assessment: an Inevitability. *Advances in Social Science, Education and Humanities Research*, 164(ICLI 2017), 136–142. <https://doi.org/10.2991/icli-17.2018.26>
- Istiyono, E., Dwandaru, W. S. B., & Muthmainah. (2019). Developing of Bloomian HOTS Physics Test: Content and Construct Validation of The PhysTeBloHOTS. *Journal of Physics: Conference Series*, 1397(012017), 1–9.
- Kowsalya, D. N., Venkat Lakshmi, H., & Suresh, K. P. (2012). Development and Validation of a Scale to assess Self-Concept in Mild Intellectually Disabled Children. *International Journal of Social Sciences & Education*, 2(4).
- Krathwohl, D. R., & Anderson, L. W. (2010). Merlin C. Wittrock and the Revision of Bloom's Taxonomy. *Educational Psychologist*, 45(1), 64–65. <https://doi.org/10.1080/00461520903433562>
- Lee, M. F., & Zainal, N. A. (2017). Development of needham model based E-module for electromagnetic field & wave. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 120–124). <https://doi.org/10.1109/IEEM.2017.8289863>
- Limongelli, C., Sciarrone, F., & Vaste, G. (2011). Personalized e-learning in Moodle: the Moodle_LS System. *Journal of E-Learning and Knowledge Society*, 7(1), 49–58. Retrieved from <https://www.learntechlib.org/p/43340>
- Martín-Blas, T., & Serrano-Fernández, A. (2009). The role of new technologies in the learning process: Moodle as a teaching tool in Physics. *Computers & Education*, 52(1), 35–44. <https://doi.org/10.1016/J.COMPEDU.2008.06.005>
- Masruroh, A. N., & Prasetyo, Z. K. (2018). Effect of E-Module with Guided Inquiry Approach Containing Nature of Science to Student's Science Literacy. *E-Journal Pend. IPA*, 7(3), 165–171.
- Pandey, S. R., & Pandey, S. (2009). Developing a More Effective and Flexible Learning Management System (LMS) for the Academic Institutions using Moodle. *ICAL 2009 - Technology, Policy and Innovation*, 249–254.
- Raykov, T., & Marcoulides, G. A. (2015). On the Relationship Between Classical Test Theory and Item Response Theory: From One to the Other and Back. *Educational and Psychological Measurement*, 76(2), 325–338. <https://doi.org/10.1177/0013164415576958>

- Skultety, L., Gonzalez, G., & Vargas, G. (2017). Using Technology to Support Teachers' Lesson Adaptations during Lesson Study. *Journal of Technology and Teacher Education*, 25(2), 185–213. Retrieved from <https://www.learntechlib.org/p/172139>
- Tanujaya, B., Mumu, J., & Margono, G. (2017). The Relationship between Higher Order Thinking Skills and Academic Performance of Student in Mathematics Instruction. *International Education Studies*, 10(11), 78–85.
- Tee, S. S., Siti, T., Tengku, M., & Zainudin, S. (2013). User Testing for Moodle Application. *International Journal of Software Engineering and Its Applications*, 7(5), 243–252.
- Winarti, Cari, Widha, S., & Istiyono, E. (2015). Analysis of Higher Order Thinking Skills Content of Physics Examinations In Madrasah Aliyah. In *International Conference on Mathematics, Science, and Education 2015 (ICMSE 2015)* (Vol. 2015, pp. 32–38).
- Yildiz, E. P., Tezer, M., & Uzunboylu, H. (2018). Student Opinion Scale Related to Moodle LMS in an Online Learning Environment: Validity and Reliability Study. *International Journal of Interactive Mobile Technologies (IJIM)*, 12(4), 97–108.
- Yusuf, I., & Widyaningsih, S. W. (2018). Profil Kemampuan Mahasiswa dalam Menyelesaikan Soal HOTS di Jurusan Pendidikan Fisika Universitas Papua. *Jurnal Komunikasi Pendidikan*, 2(14), 42–49.
- Yusuf, I., & Widyaningsih, S. W. (2019). HOTS profile of physics education students in STEM-based classes using PhET media. *Journal of Physics: Conference Series*, 1157(032021), 1–5.
- Yusuf, I., Widyaningsih, S. W., & Sebayang, S. R. B. (2018). Implementation of E-learning based-STEM on Quantum Physics Subject to Student HOTS Ability. *Turkish Science Education*, 15(December), 67–75.