Journal of Interdisciplinary Research and
Sustainability, 1(1), 29-41, December 2025

e-ISSN: xxxx-yyyy

jirs.lppm.unila.ac.id

DOI: 10.23960/jirs.v?.i?.year??

Research Article

Rasch Model Analysis on The Two-Tier Multiple Choice Questions of Physics Understanding to Measure Students' Higher Order Thinking Ability in Gender Perspective

Abdurrahman Abdurrahman[1*], Hervin Maulina[2], Novinta Nurulsari[2]
Affiliation

Article Info

Received: 20 April 2025

Revised: 29 June 2025

Accepted: 28 October 2025

Available online: 30 December 2025

Keywords:

higher-order thinking skills,

Rasch model,

two-tier questions, gender

Published by LP2M University of Lampung, Indonesia.

Abstract

The purpose of this study is to design and validate the two-tier multiple-choice (TTMC) questions on fluid concepts through Rasch Model analysis. The design of the study was Research and Development. The step of development included determining the format of the item, determining the construction of the test item, determining the assessment guidelines, expert validation, and revision of the test item. The research involved 57 senior high school students of XI-grade MIA who more interested in learning physics at Lampung Province, Indonesia. The item qualities were analyzed using the Rasch Model with the Winstep 3.73 version of application software. The research results showed that the instrument developed is valid (reached an acceptable range of PT-Measure, Outfit Mean Square (MNSQ), and Outfit z-standardized (ZSTD)} values. Furthermore, the instrument has a very good reliability with alpha Cronbach's value of 0.94, and there is no bias for gender. Besides, based on the results of the pilot study through the ongoing assessment process, the TTMC can improve students' higher-order thinking skills and contribute to the mastery of physics concepts.

To cite this article:

Abdurrahman, A., Maulina, H., & Nuruldari, N. (2020). Rasch Model Analysis on The Two-Tier Multiple Choice Questions of Physics Understanding to Measure Students' Higher Order Thinking Ability in Gender Perspective. *Journal of Interdisciplinary Research and Sustainability*, *3*(2), 47-59.

Introduction

In the 21st century learning, students are required to be more active in learning in order to be able to develop higher-order thinking skills to achieve success, especially in the field of education (Romli et al., 2019; Anwar et al., 2019; Sapriadil et al., 2018). According to Conklin, the characteristics of higher-order thinking skills encompass both critical-thinking and creative-thinking (Conklin, 2012). The ability to think critically and creatively is a higher-order thinking ability. Note that students' success depends on their 21st-century skills (Marteel-parrish & Lipchock, 2017; Ongardwanich, Kanjanawasee, & Tuipae, 2015). Thus, the ideal conditions of 21st century learning in physics will be realized if higher-order thinking skills are applied (Putra, Nur Kholifah, Subali, & Rusilowati, 2018; Wartono, Hudha, & Batlolona, 2018).

Physics achievements in 2011 were measured in reasoning aspects, Indonesia was ranked 40th out of 42 countries (Simbolon, Simanjuntak, & Simarmata, 2018; Zaniewski & Reinholz, 2016). That is, students' higher-order thinking skills are still low (Madhuri, Kantamreddi, & Prakash Goteti, 2012). Therefore, teachers should direct their students to think higher in learning. Learning by applying higher-order thinking skills makes students not only able to remember and understand a concept, but students are also able to analyze, evaluate, and create a concept well (Mutakinati, Anwari, & Yoshisuke, 2018; Putra et al., 2018). Krathworl and Anderson (2001) reveal that higher-order thinking skills in revised Bloom Taxonomy involve analytical skills (C4), evaluating (C5), and creating (C6) (Anderson & Krathwohl, 2001). Higher-order thinking skills are able to find new ways to solve their daily problems and resolve them according to decisions (Yee, 2015).

In addition, gender issues in 21st-century science have been discussed regarding its pedagogical practices, curriculum, techniques, patterns of participation in science learning, and the work in science in all stages of formal and informal education. A nonequivalent educational background onset the existence of gender inequality (Chi, Liu, Wang, & Won Han, 2018). The results of the Program for International Student Assessment (PISA), which measures capabilities, especially in the field of science in 2015, shows that Indonesia ranked 62nd out of 70 countries. The index of male and female learners in their involvement in science was 8.6 and 22.1 with the overall gender gap of -0.06. The 2015 Trends International Mathematics and Science Study (TIMSS) study which measured mathematics and science achievements showed that Indonesia was ranked 45 of 50 countries, with a score of 397 from the average TIMSS score of 500 (Mullis, Martin, Foy, & Hooper, 2015). These results indicate that there is still a gap between male and female students in science. Even though the success of students in learning physics can be seen from mastery of concepts but gender tends to bring an understanding of concepts and misconceptions derived from personal experience and the results of the social interaction.

The low mastery of concepts causes students to experience learning difficulties, especially physics. One of the high school physics materials that are considered difficult for students is fluid material (static fluid and dynamic fluid). This is consistent with the field data analysis that has been carried out which shows that students have difficulty in learning the material and the questions given by the teacher are very difficult. Difficult questions can improve students' higher-order thinking skills. Higher-order thinking skills can build students' understanding of concepts so that students are able to solve a problem. The learning process needs to integrate content learning knowledge with activities that shape higher-order thinking skills and problem-solving which of course requires continuous scaffolding by the teacher (Trilling & Fadel, 2009; Abdurrahman et al., 2019).

The success of education can be known by an evaluation. Through evaluation, the teacher can find out whether students already have higher-order thinking skills. Assessment in the form of tests can be used to hone students' thinking skills. If students are expected to have higher-order thinking skills, then the tests used must represent that aspect (Abdurrahman, Saregar, & Umam, 2018). Bernett & Francis strengthen this statement by proposing that questions used to test higher-order thinking skills can encourage students to think deeply about a material (Barnett & Francis, 2011).

The average teacher has applied higher-order thinking skills to the fluid material but it is limited to the level of cognitive analysis (C4). This is because the questions given by the teacher are mostly questions of critical-thinking ability categories, and also the lack of availability of questions to measure higher-order thinking skills at the evaluation level (C5), and creations (C6), so that students are less trained in completing higher-order thinking problems. In addition, the test instruments used by the teacher are in the form of descriptions and multiple choices because the assessment of the questions is more objective and the scoring is easier whereas instruments that measure higher-order thinking skills students can use various types of questions such as modified multiple-choice, short answer construction, and long answer construction (Saregar et al., 2018). A modified multiple-choice question that has been developed to measure higher-order thinking ability is two-tier multiple-choice questions (Treagust, 2006).

Two-tier multiple-choice questions can also be used to test students' understanding and identify student misconceptions. The second level in two-tier multiple-choice questions can be used to see students' higher-order thinking skills and see students' ability to give reasons. The second level of inclusion to reduce the occurrence of luck is often a weakness of the form of multiple-choice questions in general (Cullinane & Liston, 2011). The accuracy and consistency of the developed two-tier multiple-choice questions were analyzed using the Rasch Model with the Winstep 3.73 application. Tabatabaee-Yazdi used the Rasch model with the Winstep 3.73 application to see the validity and constancy of 40 questionnaire questions to measure teacher success (Tabatabaee-Yazdi, Motallebzadeh, Ashraf, & Baghaei, 2018). Arsad also used the Rasch Model to generate questions that were effective in assessing students' level of ability (Arsad et al., 2013).

Rasch analysis is an alternative modern measurement method that creates a measurement basis that matches the content standard unit criteria and acts as an instrument with a clear measurement unit and can function as a good model (Yuniasti & Wulandari, 2018). The analysis was carried out using empirical data obtained directly from the assessment of the questions tested to students and then converted to the logit scale. logit is used as a unit of measurement. Then the results are estimated linear correlation. Rasch allows change from defining the concept of the most compatible data line reliability to produce repeatable measurement tools that can be trusted. It is more focused on making measurement tools rather than looking for data to adjust the measurement model (Osman et al., 2011).

Based on the above problems, to be able to complete the demands in the 2013 curriculum learning, two-tier multiple-choice test questions have been developed to measure higher-order thinking skills in fluid material for high school students. The purposes of this study are (1) to produce valid questions, (2) to describe the reliability of the questions, and (3) to analyze differences in students' abilities based on gender.

Method

This study employed the Research and Development (R & D) method in developing two-tier multiple-choice test questions. The questions that have been designed are 30 items used to measure higher-order thinking skills and see the differences in students' abilities in terms of gender. The material used in developing the problem was static and dynamic fluid material. The subjects of the trial in the study were the 57 first semester students of class XI MIA in the 2017/2018 academic year. The development research model was adapted from Adams and Wieman which stages include (1) determining the format of item questions, (2) determining the construction of items, (3) determining the assessment guidelines, (4) Expert validation, and (5) revision of items (Adams & Wieman, 2011).

A. Determining the Format of the Item

The format of the item that can be applied can be a multiple-choice, essay, wrong choice, and so on. This study uses the two-tier multiple-choice item format. This is reinforced by other researchers who found that two-tier multiple-choice instruments can be used to measure students' higher-order thinking skills (Shidiq, Masykuri, & Susanti VH, 2014; Treagust, 2006). Then, determining the construction of the items in preparing the items must be in accordance with the indicators of the higher-order thinking skills that have been established, and the language used must be clear and easy to understand.

B. Determining Assessment Guidelines

Assessment guidelines are used to determine and know the achievement of students' higher-order thinking skills. The guidelines must be adjusted to the items that have been made. The assessment guideline used in this study is that; if the answers on the first-tier and second-tier are correct, then 3 points will be given. If the first-tier is correct and the second-tier is incorrect, then 2 points will be given. If the first-tier is incorrect and the second-tier is correct, then 1 point will be given. Lastly, if the first and second tier's answers are incorrect, then 0 points will be given.

C. Expert Validation

At this stage, the validity and reliability of the question design results are tested through expert validation regarding the material, language, and construction aspects. Expert validation in this development research was carried out by three lecturers who are experts in the field of physics and learning evaluation. The questions can be declared as valid and reliable if the questions have a coefficient of validity and reliability in the category of fairly high to high categories. The validity of the instrument was examined by experts, and instrument reliability tests were obtained using Cronbach alpha.

D. Items Revision

Based on the results of the validity and reliability test, the items that are not good will be revised again and the inappropriate questions will be replaced with new questions. The results of the revision will produce feasible items and can be used as valid and reliable questions in measuring higher-order thinking skills.

The data analysis technique used in this study was the Rasch Model with Winstep 3.73 application to analyze the validity and reliability of the questions and to analyze the items that are biased towards gender.

Results and Discussions

In developing the questions, the researchers used Anderson's 3 cognitive level revisions and then developed them into several cognitive levels. The indicators used to develop the problem of higher-order thinking skills (HOTS) include stimulus in the form of discourse, graphics, or images, operational verbs that describe high levels of cognitive, and material to be achieved. Indicators on this HOTS problem used Anderson and Krathwohl's learning outcome indicators which have been revised and adjusted to the cognitive level measured. The description in detail can be seen in Table 1.

Table 1.

Developed High-Order Thinking Indicators.

Cognitive Level	**Cognitive Sub Level**	**Total**
C_4	Test Ideas	8
	Analyzing the argument	2
	Identifying Arguments	2
C_5	Assessing statements	5
	Assessing the argument	4
C_6	Designing a Design	3
	Providing Solutions	6

The developed higher-order thinking questions were given to three experts in the field of education and activities to assess aspects of the construct, material, and language. Based on the results, the average score of the instrument validity from the constructed aspect was 84% which was categorized as valid with minor revisions, the strongest aspect was material that scored 85% which was categorized as valid with minor revisions, and the language content was 90% which was categorized as highly valid. Thus, several important things for fluid materials were developed to measure the students' higher-order thinking ability. Furthermore, the readability of the questions was tested to 3 students. The students stated that the questions were very easy to understand in terms of the meaning contained in each question related to the clarity of language, letters, numbers, pictures, symbols, and graphics.

After the trial and readability tests were carried out, the questions were tried-out to 57 students of XI-grade MIA. The trial was conducted on May 7, 2018. The questions consisted of 30 items, covering 14 static fluid questions and 16 dynamic fluid questions. Based on the results of the trial, the data obtained were analyzed using the Rasch Model through Winstep 3.73 application to measure the validity and reliability, as well as the biased analysis of gender

A. Analysis of Test Structure

Figure 1 shows the distribution of the difficulty level of the item on the right and the ability of the respondent on the left. Based on the map it is known that the item consists of 30 items labeled S1-S30. The most difficult item to solve by the respondent is S26, this is indicated by the *logit* value of 1.59 which is the highest *logit* value compared to the other questions. In addition, the item that are most easily solved by the respondents are S4, this is indicated by the *logit* value of -1.50 which is the lowest *logit* value of the other items. In other words, S26 item have the highest difficulty level (1.59 logit) and S4 item have a low difficulty level (-1.50 logit). Based on the sample response of students in this study, the instrument had item that were fairly easy and did not have enough difficult items.

The level of student ability is used to identify the level of students' ability to answer questions. Students' ability levels have been sorted from highest to lowest based on the *logit* value of each person. 29P respondents had the highest logit value of 1.89, meaning that respondents 29P had the highest level of ability compared to others. Meanwhile, 52P respondents have the lowest *logit* value of -1.88, this means that respondents 52P have a low level of ability. There were two respondents, namely 39L and 52P who were unable to solve the questions correctly on the problem with the lowest difficulty, namely items about S4. This means that these students need extra attention from the teacher because what the teacher has taught so far is not well understood.

```
                               Person - MAP - Item
                                     <more>|<rare>
 2                                         +
                                      29P  |
                                           |
                                      57P  |
                                           |T S26
                                      22P T|
                                      51P  |  S8
                                           |  S21
                                 18L  53P  |
                                      02P  |
 1                                         +  S25    S27
                                           |
                                      38P  |S S28
                                      36L  |  S10
                                 10P  15P S|
                                           |  S20
                       05L  09L  30L  37P  |  S17
                                           |
                            03P  06P  25L  |  S29
                                 21L  33P  |  S15    S3     S30
 0                                    23P  +M S19    S7
                                 40P  47P  |  S18    S23    S6
                                 01P  45L  |  S5
                                 12P  26P M|  S9
                            14P  24P  27L  |  S1     S24
                                      11P  |  S16    S22
                                 20L  32P  |  S12
                       04P  17P  44P  49P  |
                                      35L  |S
                                 13P  55P  |
-1                          07L  28P  34L  +  S11    S14
   08P  16P  31P  41P  48P  54L  56L  |  S2
                                      43P S|
                                      46P  |
                            19L  42P  50P  |  S13
                                           |  S4
                                           |T
                                           |
                                      39L  |
                                      52P  |
-2                                         +
                                     <less>|<frequ>
```

Figure 1. Person-Item *Wright Map*

B. Validity

The level of suitability of the item (item fitness) was used to explain whether the item was functioning normally or not. If a problem is found to be not fit, it means that there is an indication that there is a students' misconception on the item. According to Boone et al. (2014) and Bond and Fox (2015), mean-square outfit, z-standard outfit, and point measure correlation are the criteria used to see item fitness where the mean-square (MNSQ) outfit obtained was 0.5 <MNSQ <1.5, the obtained Z-standard (ZSTD) outfit value was -2.0 <ZSTD <+2.0, and the point measure correlation value (Pt Mean Corr) obtained was 0.4 <Pt Mean Corr <0.85 (Sumintono & Widhiarso, 2014). Based on these three criteria, Table 2 shows that all developed items have met the criteria so that it can be concluded that all questions are valid and do not need to be changed or replaced. All questions were fit, meaning there was no misconception on the item.

Table 2.

Item Fitness Order.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
20	55	57	.49	.17	1.05	.4	1.28	1.4	A .52	.62	50.9	47.3	S20
30	70	57	.08	.16	1.27	1.5	1.20	1.1	B .72	.62	33.3	41.4	S30
1	88	57	-.37	.16	1.25	1.5	1.23	1.2	C .56	.61	38.6	42.6	S1
23	77	57	-.10	.16	1.17	1.0	1.25	1.3	D .43	.62	38.6	40.4	S23
10	49	57	.67	.18	1.23	1.2	1.24	1.1	E .50	.61	43.9	49.0	S10
9	84	57	-.27	.16	1.23	1.4	1.24	1.3	F .54	.61	29.8	42.4	S9
16	93	57	-.50	.16	1.22	1.3	1.19	1.0	G .46	.60	40.4	41.9	S16
12	96	57	-.57	.16	1.17	1.0	1.07	.4	H .67	.60	31.6	42.0	S12
19	72	57	.03	.16	1.13	.8	1.12	.7	I .53	.62	43.9	41.0	S19
14	114	57	-1.04	.16	1.12	.7	1.00	.1	J .55	.55	45.6	46.4	S14
22	93	57	-.50	.16	1.12	.7	1.09	.5	K .45	.60	47.4	41.9	S22
13	127	57	-1.40	.17	.93	-.3	1.09	.5	L .41	.50	54.4	51.1	S13
7	72	57	.03	.16	1.07	.5	1.05	.4	M .66	.62	38.6	41.0	S7
17	58	57	.41	.17	.91	-.4	1.05	.3	N .53	.62	52.6	45.8	S17
3	69	57	.11	.16	.96	-.2	1.05	.3	O .55	.62	38.6	41.4	S3
24	91	57	-.45	.16	.96	-.2	1.04	.3	o .51	.60	42.1	42.3	S24
11	112	57	-.98	.16	1.02	.2	.94	-.3	n .62	.56	36.8	45.3	S11
2	117	57	-1.12	.17	.99	.0	.95	-.2	m .48	.54	50.9	46.6	S2
5	83	57	-.25	.16	.98	.0	.95	-.2	l .71	.62	40.4	42.3	S5
27	38	57	1.04	.19	.97	-.1	.81	-.7	k .67	.59	57.9	56.9	S27
4	130	57	-1.50	.18	.87	-.6	.90	-.3	j .49	.49	57.9	52.8	S4
28	45	57	.80	.18	.89	-.5	.75	-1.1	i .77	.60	63.2	52.5	S28
29	67	57	.16	.16	.89	-.6	.82	-.9	h .77	.63	43.9	41.3	S29
15	70	57	.08	.16	.84	-.9	.88	-.6	g .63	.62	47.4	41.4	S15
26	25	57	1.59	.22	.88	-.4	.82	-.5	f .56	.53	63.2	67.4	S26
8	28	57	1.45	.21	.86	-.6	.87	-.4	e .62	.54	70.2	64.1	S8
18	78	57	-.12	.16	.85	-.9	.80	-1.1	d .74	.62	40.4	40.5	S18
6	77	57	-.10	.16	.80	-1.2	.79	-1.2	c .64	.62	56.1	40.4	S6
25	39	57	1.01	.19	.74	-1.3	.65	-1.5	b .75	.59	73.7	56.2	S25
21	31	57	1.32	.20	.71	-1.4	.72	-1.1	a .65	.56	63.2	61.9	S21
MEAN	74.9	57.0	.00	.17	1.00	.1	.99	.1			47.8	46.9	
S.D.	28.2	.0	.79	.02	.16	.9	.18	.9			11.2	7.5	

C. Reliability

The results of data analysis showed that 30 questions were declared reliable with Cronbach's alpha value of 0.94 which was in the excellent reliability category. Cronbach's alpha value is used to measure reliability, namely the interaction between person and item as a whole as shown in tables 3 and 4.

Table 3.

Person Reliability

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	39.4	30.0	-.27	.24	1.01	.1	.99	.0
SD	17.6	.0	.90	.02	.26	1.0	.22	.8
MAX.	78.0	30.0	1.89	.31	1.59	2.3	1.47	1.7
MIN.	12.0	30.0	-1.88	.21	.62	-1.9	.58	-1.5

REAL RMSE .25 TRUE SD .87 SEPARATION 3.49 Person RELIABILITY .92
MODEL RMSE .24 TRUE SD .87 SEPARATION 3.69 Person RELIABILITY .93
SE. OF Person MEAN = .12

Person RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY= .94

Table 3 shows the average values of MNSQ INFIT MNSQ and OUTFIT sequentially are 1.01 and 0.99, which means that the value is close to the perfect value of 1.00. The mean values of ZSTD and OUTFIT ZFI INFIT respectively are 0.1 and 0.0, meaning that the quality of the person is getting better because the value is near perfect, which is 0.0. The person's reliability value is 0.92, which shows that the consistency of answers from respondents is very good. The separation value is 3.49, so that the value of the separation is 4.99 (rounded to

5) means that there are five groups of respondents. This shows that the greater the score, the better the quality of the respondent as a whole. This reflects a variety of abilities (heterogeneous), which shows the representativeness of the students' ability to take the test.

Table 4.
Items Reliability

	TOTAL SCORE	COUNT	MEASURE	MODEL EROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	74.9	57.0	.00	.17	1.00	.1	.99	.1
SD	28.2	.0	.79	.02	.16	.9	.18	.9
MAX.	130.0	57.0	1.59	.22	1.27	1.5	1.28	1.4
MIN.	25.0	57.0	-1.50	.16	.71	-1.4	.65	-1.5
REAL RMSE .18 TRUE SD			.77 SEPARATION 4.38		Person RELIABILITY .95			
MODEL RMSE .17 TRUE SD			.77 SEPARATION 4.52		Person RELIABILITY .95			
SE. OF Person MEAN = .15								

Table 4 shows the MNSQ and OUTFIT INFIT MNSQ mean values of 1.00 and 0.99 respectively, which means that the value is close to the perfect value of 1.00. The average value of ZSTD and OUTFIT ZFID INFIT sequentially is 0.99 and 0.1, meaning that the quality of the items is better because the value is close to perfect, which is 0.0. The value of item reliability is 0.95, which indicates that the quality of the items is special. The separation value is 4.38, so that the value of the separation is 6.17 (rounded to 6), meaning that there are six groups of items. This shows that the greater the separation value, the better the overall item quality.

D. Differential Item Functioning (DIF)

A valid instrument is an instrument used does not contain bias. An instrument is said to be biased if one of the individuals with certain characteristics is more advantageous than individuals with other characteristics. For example, instruments are more easily answered by male students than females, and this indicates the existence of gender-biased items. Detection of biased items can be seen in Table 5 and Figure 1.

Table 5.
Differential Item Monitoring (DIF)

Person CLASSES	SUMMARY DIF CHI-SQUARE	D.F.	PROB.	BETWEEN-CLASS MEAN-SQUARE	t=ZSTD	Item Number	Name
2	.2181	1	.6405	.0953	-.6810	1	S1
2	.7177	1	.3969	.3119	-.2113	2	S2
2	.5778	1	.4472	.2537	-.3070	3	S3
2	.2616	1	.6090	.1114	-.6292	4	S4
2	.4958	1	.4814	.2184	-.3724	5	S5
2	.0048	1	.9447	.0034	-1.3316	6	S6
2	.1941	1	.6596	.0842	-.7201	7	S7
2	.0936	1	.7597	.0371	-.9423	8	S8
2	.2933	1	.5881	.1282	-.5803	9	S9
2	.3751	1	.5403	.1682	-.4789	10	S10
2	.5378	1	.4633	.2380	-.3353	11	S11
2	.0000	1	1.0000	.0001	-1.5539	12	S12
2	.8137	1	.3670	.3545	-.1486	13	S13
2	3.4806	1	.0621	1.6296	.8464	14	S14
2	.4269	1	.5135	.1867	-.4375	15	S15
2	.0853	1	.7702	.0373	-.9409	16	S16
2	.2528	1	.6151	.1133	-.6234	17	S17
2	.2803	1	.5965	.1237	-.5931	18	S18
2	.8595	1	.3539	.3841	-.1080	19	S19
2	.0368	1	.8478	.0153	-1.1232	20	S20
2	.2888	1	.5910	.1275	-.5822	21	S21
2	.0853	1	.7702	.0373	-.9409	22	S22
2	.5833	1	.4450	.2586	-.2985	23	S23
2	.6884	1	.4067	.3035	-.2243	24	S24
2	.4142	1	.5199	.1847	-.4419	25	S25
2	.0559	1	.8131	.0254	-1.0268	26	S26
2	.0865	1	.7687	.0355	-.9531	27	S27
2	.2951	1	.5869	.1324	-.5689	28	S28
2	.0000	1	1.0000	.0019	-1.3878	29	S29
2	.0962	1	.7564	.0414	-.9159	30	S30

Table 5 shows that all items do not contain gender bias because their probability values are> 0.05. This is consistent with the statement that the item is said to contain biases if it is found that the bitter probability value is below 5% (Sumintono & Widhiarso, 2014).

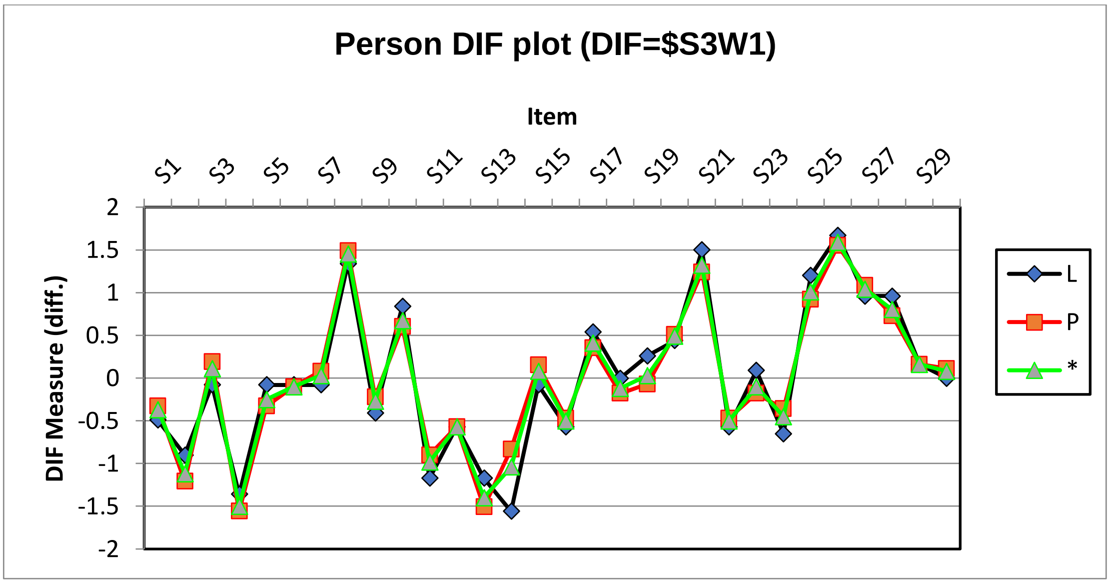


Figure 1. Person DIF Plot

Figure 1 shows there are three curves based on the sex of the student, namely L (male), P (female), and * (star) indicates the average value. It can be seen that the curve that approaches the upper limit (question S26) shows the high difficulty of the problem, while the curve below shows an easy problem (S4). The difference in the ability of students to work on all items correctly is not much different. Relatively invisible items differ in the ability to solve problems such as S8, S9, S22, and S29.

Discussion

The developed higher-order thinking questions had been validated by validators and then tested on 57 students to analyze the validity, reliability, and higher-order thinking skills. After analyzing the overall test questions that have been tested and revised, the result obtained was in the form of two-tier multiple-choice items that have a reliability value of 0.95 which is included in the excellent category (Sumintono & Widhiarso, 2014). The items that have been developed in this study are 30 fluid matter questions consisting of 14 static fluid questions and 16 dynamic fluid questions which can be seen in the appendix by referring to 3 cognitive levels of higher-order thinking from Anderson which is equipped with cognitive sub-level. The student's higher-order thinking activities are giving arguments, identifying arguments, evaluating arguments, evaluating statements, making designs, and providing solutions. The higher-order thinking test questions developed to refer to the cognitive level of Anderson and Krathworl by using Anderson's revised Bloom's Taxonomy indicator which measures students' higher-order thinking skills consist of analyzing (C4), evaluating (C5), and creating (C6). The test questions developed consist of 12 C4 questions, 9 C5 questions, and 9 C6 questions.

The questions developed after being declared feasible were tested by three validators, and the questions were stated to be easy to understand and did not contain a double interpretation by three students. Then the field-trial was conducted to 57 first semester high school students of class XI MIA. This limited trial was conducted on May 7, 2018.

Analysis of data from the results of the trial was done using Winstep 3.73 which obtained 30 questions declared as valid because all items have met the three criteria set by Boone et al. (2014). Where the mean square (MNSQ) outfit received is 0.5 <MNSQ <1.5, the received Z-standard (ZSTD) outfit value is -2.0 <ZSTD <+2.0, and the point measure correlation value (Pt Mean Corr) received is 0.4 <Pt Mean Corr <0.85. It was concluded that all questions were valid and all were fit. This means that all items function normally in making measurements and there is no misconception on students about the items.

In addition, the items were declared reliable. The item developed has a person reliability value of 0.92 which is included in the excellent category. It means that the respondent answers the whole item earnestly and not carelessly. The value of item reliability is 0.95 which is included in the excellent category. It means that all items identified have accuracy with the model and are quality items. The alpha Chonbrach value is 0.94. It means that the problems have high reliability (Sumintono & Widhiarso, 2014). Thus, Higher-order thinking test can be used as an alternative to measuring students' higher-order thinking skills.

Instrument quality in terms of overall respondents and items in this study is categorized as excellent because it has a large separation index. This shows that the heterogeneous response of items are carefully placed. All items have valid options, so the respondent is able to understand the choices given to each difficulty level of the item.

Valid instruments do not contain bias. Bias in items can be known through the item probability value. If an item has a probability value of<0.05, the item is said to contain bias. In this study, it was found that all items did not have a bias because the probability value was> 0.05 which can be seen in Table 5.

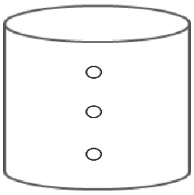
Figure 1 shows the most difficult problem for students to do is S26. This is indicated by the logic value of 1.59 which is the highest logit value compared to other questions. In addition, the easiest problem for students to do in S4, this is indicated by the logic value of -1.50 which is the lowest logit value of other questions. In other words, item S26 has the highest difficulty level (1.59 logit) and the S4 item has a low level of difficulty (-1.50 logit).In addition, the graph in Figure 1 shows the difference in the ability of students to work on all items correctly. Relatively invisible items differ in the ability to solve problems such as S8, S9, S22, and S29. This can occur because the percentage of the total number of student answers is true or false.

In S8, the percentage of students who answered questions on first-tier and second-tier correctly was 16%, the percentage of students who answered the questions correctly on first-tier and incorrectly on second-tier was 47.4%. The percentage of students who answered the questions incorrectly on the first-tier and correctly on second-tier was 17.5%. The students who answered incorrectly on the first-tier and second-tier questions were 19.3%. In S9, the percentage of students who answered questions on first-tier and tertiary tier correctly was 15.7%, the percentage of students who answered the questions on the first-tier correctly and second-tier incorrectly was 36.8%, and the percentage of students who answered the questions incorrectly at first-tier and correctly on second-tier was 24.6%. The students who answered incorrectly on the first-tier and second-tier was 45.9%.

Furthermore, In S22, the percentage of students who answered questions on first-tier and second-tier correctly was 7%. The percentage of students who answered the questions on first-tier correctly and second-tier incorrectly was 17.5%. The percentage of students who answered the questions incorrectly on the first-tier and correctly on the second-tier was 45.6%. The students who answered incorrectly on the first-tier and second-tier was 29.9%. In S29 (see Figure 2), the percentage of students who answered questions on first-tier and second-tier correctly was 0%. The percentage of students who answered the questions on first-tier correctly and second-tier incorrectly was 14%. The percentage of students who answered questions incorrectly on first-tier and correctly on the second-tier was 23%. The students who answered incorrectly on the first-tier and the second-tier was 56.3%.

S29. Afra merupakan salah satu peserta ujian praktikum. Pada saat ujian, ia diminta untuk memperagakan Asas Bernaulli pada fluida. Apabila di laboratorium disediakan beberapa alat percobaan yaitu pompa hidrolik, venturimeter, dan toricelli seperti gambar berikut.


Venturimeter


Toricelli


Pompa Hidrolik

Untuk memperagakan Asas Bernaulli pada fluida maka hal yang seharusnya dilakukan oleh Afra adalah….

A. Mengambil alat venturimeter, lalu melakukan pengamatan pengaruh luas penampang terhadap laju aliran fluida.
B. Mengambil alat pompa hidrolik, lalu melakukan pengamatan pengaruh tekanan terhadap laju aliran fluida.
C. Mengambil alat toricelli, lalu melakukan pengamatan pengaruh kedalaman terhadap laju aliran fluida.

(***First-tier***)

Alasan yang tepat memilih jawaban di atas adalah….

1) Venturimeter merupakan alat yang digunakan untuk mengukur kelajuan fluida. Semakin besar luas penampang fluida, maka semakin cepat laju aliran fluida.
2) Toricelli merupakan alat yang digunakan untuk mengukur kelajuan fluida. Semakin dalam posisi lubang kebocoran, maka semakin cepat laju aliran fluida.
3) Pompa hidrolik merupakan alat yang digunakan untuk mengukur kelajuan fluida. Semakin besar tekanan yang diberikan pada fluida, maka semakin cepat laju aliran fluida.

(***Second-tier***)

Figure 2. An Example of the S29 Item Number of the Fluid Concept of Two-Tier Multiple Choice

The data above can be concluded that male and female students have the ability to answer questions that tend to be the same in the four questions. Of the four questions, male and female students cannot answer the questions correctly. This can be caused by the weak understanding of the students, not yet trained students in working on two-tier multiple choice questions, limited learning resources, students forget about past material, and lack of time to do the questions.

Based on the explanation above, it was concluded that 30 two-tier multiple choice questions were valid and could be used as an alternative to measure students' higher-order thinking ability towards fluid material (static and dynamic fluid). This is in accordance with the results of Shidiq et al. (2014) and Treagust (2006), which shows that two-tier multiple-choice instruments can be used to measure students' higher-order thinking skills (Shidiq et al., 2014; Treagust, 2006).

Conclussion and Recommendations

Based on the results of the research and discussion, it can be concluded that (1) the higher-order thinking test questions of the fluid material developed have met the instrument's eligibility standards which are valid and reliable. The results of data analysis using Winstep 3.73 show that all questions have met the three criteria of the suitability of items according to Boone et al. (2014), (2) HOTS test questions have very high reliability with alpha Cronbach of 0.94. Thus the test questions can be used as an alternative to measuring students' higher-order thinking skills, (3) there is no bias in the item on gender because the probability value of the problem is> 0.05.

Acknowledgement

The authors are grateful to the Indonesian of Research, Technology, and Higher Education Ministry for funding the research with contract number of No: 065/SP2H/LT/DPRM/2019.

References

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessment: A revision of Bloom's taxonomy of educational objectives*. New York: Addison Wesley Longman.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.

Conklin, W. (2012). *Higher Order Thinking Skills to Develop 21st Century Learners*. Shell Educational Publishing.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Understanding person measures. In *Rasch analysis in the human sciences* (pp. 69–92). Springer, Dordrecht.

Abdurrahman, A., Nurulsari, N., Maulina, H., Rahman, B., Umam, R., & Jermsittiparsert, K. (2019). Multi-level scaffolding: A novel approach of physics teacher development program for promoting content knowledge mastery. *International Journal of Innovation, Creativity and Change*, 7(8).

Abdurrahman, A., Saregar, A., & Umam, R. (2018). The effect of feedback as soft scaffolding on ongoing assessment toward the quantum physics concept mastery of the prospective physics teachers. *Jurnal Pendidikan IPA Indonesia*, 7(1), 34–40.

Adams, W. K., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education*, 33(9), 1289–1312.

Anwar, C., Saregar, A., Yuberti, Zellia, N., Widayanti, Diani, R., & Wekke, I. S. (2019). Effect size test of learning model ARIAS and PBL: Concept mastery of temperature and heat on senior high school students. *EURASIA Journal of Mathematics, Science and Technology Education*, 15(3), 1–9.

Arsad, N., Kamal, N., Ayob, A., Sarbani, N., Tsuey, C., Misran, N., & Husein, H. (2013). Rasch model analysis on effectiveness of early evaluation questions as a benchmark for new student ability. *International Education Studies*, 6(6), 185–190.

Barnett, J. E., & Francis, A. L. (2011). Using higher order thinking questions to foster critical thinking: A classroom study. *Educational Psychology*, 32(2), 201–211.

Chi, S., Liu, X., Wang, Z., & Won Han, S. (2018). Moderation of the effects of scientific inquiry activities on low SES students' PISA 2015 science achievement by school teacher support and disciplinary climate in science classroom across gender. *International Journal of Science Education*.

Madhuri, V. G., Kantamreddi, S. V., & Prakash Goteti, N. L. (2012). Promoting higher order thinking skills using inquiry-based learning. *European Journal of Engineering Education*, 37, 117–123.

Marteel-Parrish, A. E., & Lipchock, J. M. (2017). Preparing chemistry majors for the 21st century through a comprehensive one-semester course focused on professional preparation, contemporary issues, scientific communication, and research skills. *Journal of Chemical Education*.

Mutakinati, L., Anwari, I., & Yoshisuke, K. (2018). Analysis of students' critical thinking skill of middle school through STEM education project-based learning. *Jurnal Pendidikan IPA Indonesia*, 7(1), 54–65.

Ongardwanich, N., Kanjanawasee, S., & Tuipae, C. (2015). Development of 21st century skill scales as perceived by students. *Procedia - Social and Behavioral Sciences,* 191, 737–741.

Putra, F., Nur Kholifah, I. Y., Subali, B., & Rusilowati, A. (2018). 5E-learning cycle strategy: Increasing conceptual understanding and learning motivation. *Jurnal Ilmiah Pendidikan Fisika Al-Biruni*, 7(2), 171.

Saregar, A., Irwandani, I., Abdurrahman, A., Parmin, P., Septiana, S., Diani, R., & Sagala, R. (2018). Temperature and heat learning through SSCS model with scaffolding: Impact on students' critical thinking ability. *Journal for the Education of Gifted Young Scientists*, 6(3), 39–54.

Shidiq, A., Masykuri, M., & Susanti VH, E. (2014). Development of two-tier multiple choice instrument assessment for measuring higher order thinking skills in solubility materials and solubility times for high school students. *Journal of Chemical Education*.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2015). *TIMSS 2015 International Results in Mathematics*. Boston College: International Study Center.

Osman, S. A., Badaruzzaman, W. H. W., Hamid, R., Taib, K., Khalim, A. R., Jaafar, O., et al. (2011). Assessment on students' performance using Rasch models in reinforced concrete design course examination. In *WSEAS Recent Researches in Education Proceedings* (pp. 193–198).

Romli, S., Abdurrahman, A., & Riyadi, B. (2018). Designing students' worksheet based on open-ended approach to foster students' creative thinking skills. In *Journal of Physics: Conference Series* (Vol. 948, No. 1, pp. 1–6). IOP Publishing.

Sapriadil, S., Setiawan, A., Suhandi, A., Malik, A., Safitri, D., Lisdiani, S. A. S., & Hermita, N. (2018). Optimizing students' scientific communication skills through higher order thinking virtual laboratory (HOTVL). In *Journal of Physics: Conference Series*. IOP Publishing.