Jurnal Penelitian Fisika dan Aplikasinya (JPFA) *Volume 11, Issue 2, December 2021*
p-ISSN: 2087-9946 DOI: 10.26740/jpfa.v11n2.p127-137
e-ISSN: 2477-1775 https://journal.unesa.ac.id/index.php/jpfa

Research Article

Analysis of the Characteristics of Higher Order Thinking Skills (HOTS) Test on Momentum and Impulse for Senior High School Student Using Item Response Theory

Aulia Rahman [1,a,*], Heni Rusnayati [1,b], and Muslim [1,c]

[1]Departement of Physics Education, Universitas Pendidikan Indonesia
Jl. Dr. Setiabudi. No. 229, Bandung 40154, Jawa Barat, Indonesia

e-mail: [a] rahmanaul4@student.upi.edu, [b] heni@upi.edu, and [c] muslim@upi.edu
* Corresponding Author

Abstract

Currently, 21st-century skills are needed by students, especially Higher Order Thinking Skills (HOTS), to meet the needs of the rapidly growing world of work. It is necessary to provide debriefing during learning activities to meet these needs. Besides that, a quality instrument is needed to measure the skills of each student, the more information that the instrument can provide shows that the instrument is getting better. This study aimed to determine the characteristics of the Higher Order Thinking Skills test on momentum and impulse consisting of validity, reliability, difficulty level, and discrimination index based on item response theory analysis. The method used in this research is a descriptive method with a quantitative approach and a One-Shot Design research design. The population was second-year senior high school students in Bandung. Meanwhile, the sample of this study consisted of 122 second-year senior high school students who were selected using a purposive sampling technique. The instrument used was the Higher Order Thinking Skills test on momentum and impulse. Based on the result, 16 items were categorized as valid. Besides, the reliability of the test instrument was good. For the level of difficulty, an item was categorized as very difficult, an item was difficult, twelve items were medium, three items were easy, and an item was very easy. Lastly, for the discrimination index, thirteen items were considered good, and five items did not classify as good.

Keywords: *Higher Order Thinking Skills; Item Response Theory; Momentum Impulse; Techniques of testing; Theory of testing and techniques.*

Analisis Karakteristik Tes *Higher Order Thinking Skills* (HOTS) pada Materi Momentum dan Impuls Bagi Siswa Sekolah Menengah Atas Menggunakan Teori Respon Butir

Abstrak

Keterampilan abad 21 sangat dibutuhkan oleh para pelajar khususnya keterampilan berpikir tingkat tinggi (HOTS) untuk memenuhi kebutuhan dunia kerja yang semakin berkembang pesat. Dalam memenuhi kebutuhan tersebut, perlu adanya pembekalan yang diberikan selama kegiatan pembelajaran. Selain itu, diperlukan juga sebuah instrument yang berkualitas dan mampu mengukur keterampilan dari setiap peserta didik, di mana semakin banyak informasi yang dapat diberikan oleh instrument menunjukan bahwa instumen tersebut semakin baik. Penelitian ini bertujuan untuk mengetahui karakteristik tes Higher Order Thinking Skills pada materi momentum dan impuls yang terdiri dari validitas, reliabilitas, taraf kesukaran, dan daya pembeda berdasarkan analisis teori respon butir. Metode penelitian yang digunakan adalah deskriptif dengan pendekatan kuantitaif dan desain penelitian One-Shot Design. Populasi penelitian adalah seluruh peserta didik kelas XI di salah satu sekolah menegah atas di Kota Bandung. Sampel penelitian ini terdiri dari 122 peserta didik kelas XI yang dipilih menggunakan teknik purposive sampling. Instrumen yang digunakan adalah tes Higher Order Thinking Skills pada materi momentum dan impuls. Berdasarkan hasil analisis, 16 butir soal dinyatakan valid. Reliabilitas instumen tes termasuk kategori bagus. Untuk taraf kesukaran, satu soal dikatakan sangat sukar, satu soal dikatakan sukar, dua belas soal termasuk kategori sedang, tiga soal termasuk mudah, dan satu soal termasuk sangat mudah. Untuk daya pembeda, tiga belas soal termasuk kategori baik dan lima soal tidak memenuhi kategori baik.

Kata Kunci: *Higher Order Thinking Skills; Teori Respon Butir; Momentum Impuls; Teknik Pengujian; Teori dan Teknik Pengujian.*

PACS: 01. 40.-d; 01. 40. Fk; 01. 40. gf; 01. 50. Kw.

Article History: Received: 18 January 2021 Approved with minor revision: 28 June 2021
Accepted: 23 July 2021 Published: 30 December 2021
How to cite: Rahman A, et al. Analysis of the Characteristics of Higher Order Thinking Skills (HOTS) Test on Momentum and Impulse for Senior High School Student Using Item Response Theory. *Jurnal Penelitian Fisika dan Aplikasinya (JPFA)*. 2021; **11**(2): 127-137. DOI: https://doi.org/10.26740/jpfa.v11n2.p127-137.

I. INTRODUCTION

Currently, learning is expected to practice and facilitate students to have 21st-century skills such as communication, collaboration, critical thinking, problem-solving, and creativity [1]. This is closely related to thinking skills differentiated by Anderson & Krathwohl in the revised Bloom's Taxonomy. On the taxonomy, thinking skill was differentiated into Higher Order Thinking Skill (HOTS) and Lower Order Thinking Skill (LOTS). LOTS involved remembering, understanding, and applying. Meanwhile, HOTS involves analyzing, synthesizing, evaluating, and creating [2].

The skill that the student must achieve is not only LOTS but also includes HOTS because it is a sequential learning process. The student is said to be skilled at HOTS only if the student can analyze, evaluate, and create. HOTS has become very important to be managed by students to improve their competence in facing the globalization era, advances in ICT, the convergence of science and technology as an impact of technoscience, and the rise of the creative industry in the future [3].

HOTS in a person can be measured by an assessment process [4]. The assessment process certainly requires a medium that can measure the skills and competencies of the object of research, one of which is through a test. The test intends to measure the achievement of abilities or skills in a certain competency and produce quantitative data [5]. In Indonesia, the average physics national exam score designed to demand HOTS is still low, comprising 44.22 in 2018 and 46.35 in 2019 [6]. This number is relatively low when compared to other countries. Indonesian students' average score on the Program for International Student Assessment (PISA) survey in the science field ranked 74th out of 79 countries that participated in 2018 [7].

The test instrument designed to measure HOTS is certainly needed by the teacher to diagnose students' weaknesses, to differentiate between a superior and less superior group of students, and to help students understand better [8]. However, teachers still find it challenging to make a test instrument to measure HOTS. Based on research, 50% of Physics teachers who compiled the test instrument tended to only measure LOTS, and the items were not contextual. In addition, 75% of the items compiled tended to only measure recall skill [9]. Furthermore, the context used in the items are mostly in-class contexts and are very theoretical.

A survey was conducted among high school teachers in Bandung. Based on the result, 14.3% of respondents have never made a HOTS test instrument, 28.6% of respondents rarely made a HOTS test instrument, and 57.1% of respondents used the HOTS test instrument, even though only half of the entire items. Some of the reasons include the lack of reference to HOTS test items, the lack of experience in making HOTS test instruments, and many students still not used to working on HOTS tests.

The objectivity of the learning outcome assessment depends on the test instrument's quality. A multiple-choice test is a test that consists of several alternative answers. There is only one correct answer, whereas the others are false. This test is objective because there is only one correct answer [10]. To see the quality of a test instrument, an item analysis can be done. An item analysis aims to examine each item to obtain the quality before use, improve the quality of the item through revision or remove the ineffective item, and get diagnostic information from students [11].

In analyzing the test, classical test theory (CTT) and modern test theory named item response theory (IRT) can be used. Classical theory has several weaknesses, such as 1) the test item statistics are very dependent on the characteristics of the test subject; 2) the participant's estimated ability is very dependent

on the test presented; 3) the standard error of the score estimator applies to all participants so that the standard error of measurement for each participant and the items does not exist; 4) the information presented is limited to true or false answers without paying attention to the participant's answer pattern; 5) the parallel test assumption is difficult to fulfill [12]. Alas, item response theory is a theory of item analysis that contains improvements on the weaknesses of classical theory, especially on the dependence of item size on the test participant group and the dependence of the participant's characteristic size on the item group [13].

With the item response theory, the item difficulty level and other item characteristics remain (invariant) to the test participant group; it does not matter which group of participants is working on the test [13]. As mentioned, the framework of classical test theory (CTT) for test assessment has some important limitations. To overcome these shortcomings, the item response theory (IRT) was introduced [14]. So, it can be said that this item response theory complements classical test theory (CTT). The item response theory does not have a dependency on the analysis of the respondent; the item response theory also determines the standard error for each test item so that each test item has a different error value [15].

The results of research conducted by Rakkapao [14], which examined the function of the test to measure the ability of students in vector concepts (Test of Understanding of Vectors / TUV), found that analysis with item response theory could tell that the test was able to measure students' understanding from low-ability students to high-ability students. In other words, using item response theory analysis, the test has the same characteristics for all participants. Item response theory provides a relationship between the ability of the respondent to answer correctly in a test item, in which respondents with high ability will have a greater probability of answering correctly when compared to respondents with the lower ability.

Item response theory has various types of models, one of which is the logistic parameter model (1-PL, 2-PL, 3-PL), where each of these parameters describes the character information of the item being tested by connecting the test taker's estimated ability (Θ) with its probability to respond correctly to the given item [16]. In determining the most suitable parameter logistic model to analyze the character of the item being tested, it is necessary to do statistical calculations using Chi-Square Statistics [17] or by testing the items on each logistic parameter model and looking at the total information curve of each logistics parameter model. At the end, the most suitable model is indicated by the highest total information value.

II. METHOD

The descriptive method with a quantitative approach was used in this research. The quantitative approach aims to describe or explain the occurrence in meaningful numbers [18]. The design research used was One-Shot Design. According to Arikunto [19], One-Shot Design is a research design that uses a one-time data collection. This research was carried out from the beginning of the year for the preparation and testing of instruments. Meanwhile, the data collection was carried out in the 9th month of 2020.

The population of this study was second-year students in a high school in Bandung. The sample in this study consisted of 122 second-year students in a high school in Bandung selected using a purposive sampling technique. Purposive sampling is a type of nonprobability sampling technique that does not provide a similar opportunity for each population member to be selected as a sample [20].

The instrument used in this study was a HOTS test about momentum and impulses, which consists of 18 multiple-choice items. Data analysis in this study was carried out using the one-parameter logistic (1-PL) item response theory known as the Rasch Model to determine the validity and reliability of the items. While the two-item logistic parameter response theory (2-PL) is used to determine the level of difficulty and discrimination index of the items.

III. RESULTS AND DISCUSSION

The research was conducted online because it was not possible to carry out any face-to-face activities at school. Google Classroom was used as an online class facility, and Google Form as an online test facility. The characteristics of the HOTS test instrument can be known through the test characteristic curve (TCC),

where the curve can show the level of difficulty and the discrimination index of a test instrument by displaying the score obtained by students. In this study, the TCC can be seen in Figure 1.

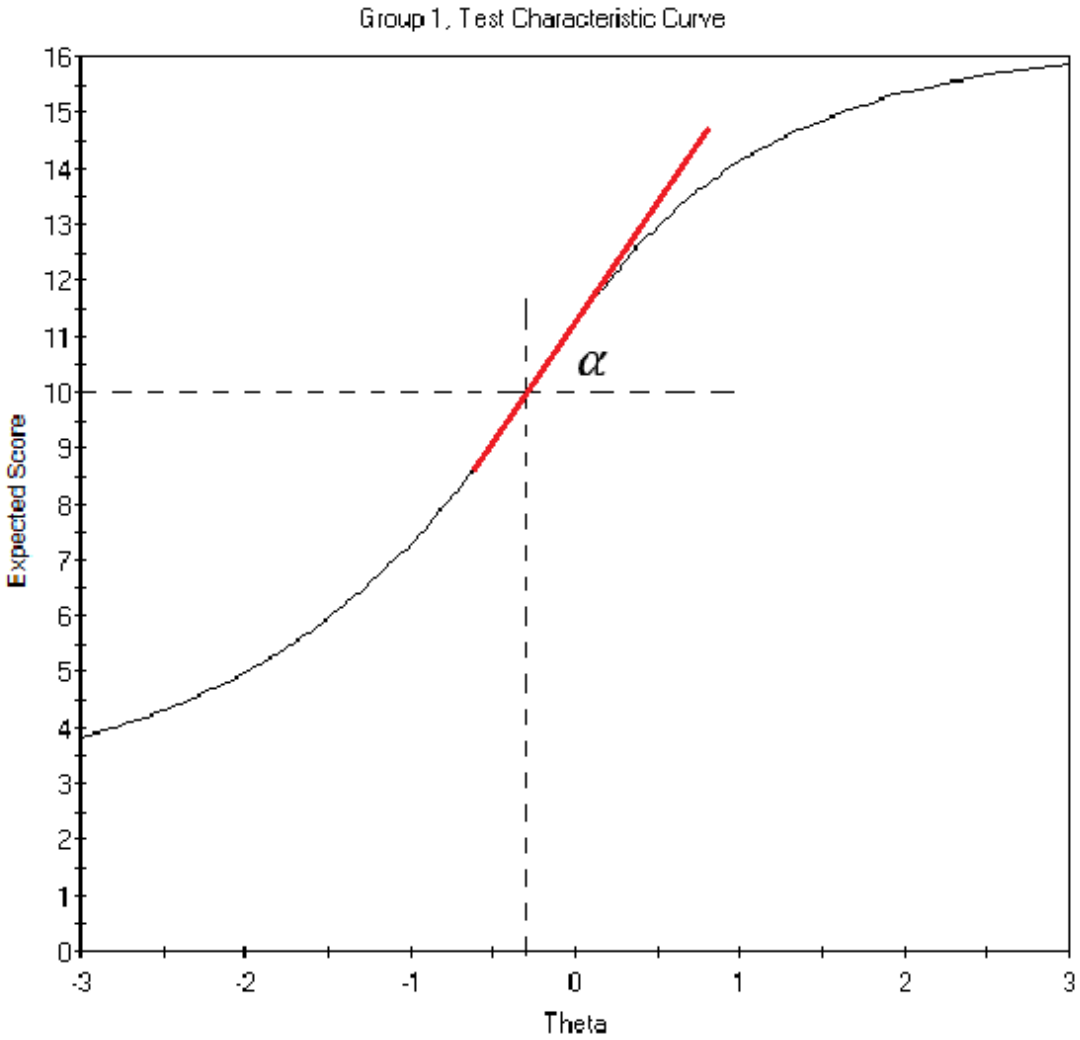

Figure 1. Test Characteristic Curve

The test instrument had 18 items in multiple choices. Each item has a score of 1, so the total score of this test is 18. Based on Figure 1, it can be seen that students with the ability of -3 get a score of 4, which means that students with that ability can work on four items of the 18 items given correctly. Whereas respondents with the ability +3 get a score of 16, meaning that students with this ability can work on 16 of the 18 items given correctly. So, the highest score on this test is 16 and the lowest score obtained is 4.

The difficulty level is known through the vertical line that crosses the x-axis (student's level of ability) on the TCC graph. Because the highest score is 16 and the lowest score is 4, the midpoint is at point 10. From that point, a line is drawn horizontally until it is right on the tangent. Then when the horizontal line is at the tangent, the line is drawn back vertically. The level of difficulty can be seen from the number shown by the point on the vertical line test on the x-axis, where the *b* value shown is -0.30, so the level of difficulty for the HOTS test instrument is classified in the easy category. According to Hambleton, the difficulty level criteria can be seen in Table 1.

Table 1. Item Difficulty Level Criteria [21]

Difficulty Level Value (b)	Criteria
$b \leq -2$	Very easy
$-2 < b \leq -1$	Easy
$-1 < b \leq 1$	Average
$1 < b \leq 2$	Difficult
$b > 2$	Very difficult

The discrimination index (a) is obtained from the slope of the curve, which is the value of tan α. Based on the TCC, the tan α obtained is 55°, so the discrimination index obtained in this

research instrument is 1.43. According to the discrimination index, the whole test instrument can be categorized as good because the value is between 0 to 2.

Item Validity

The research data collected from 122 students were analyzed using the item response theory with one logistical parameter (1-PL), known as the Rasch Model, to determine the validity of each item. According to Sumintono and Widhiarso [22], each item should occupy three criteria:

1) Outfit mean square value (MnSq): $0.5 < MnSq < 1.5$
2) Outfit Z-Standart value (ZStd): $-2.0 < ZStd < +2.0$
3) Point measure correlation (PtMeaCorr): $0.4 < PtMeaCorr < 0.85$

Any item is considered valid if it occupies all of those criteria. If an item only occupied two of the criteria, the item is categorized as Fit 2 but still can be considered valid. If only one criterion is occupied, the item is said to be Fit 1 and the item is considered invalid. Likewise, if no criteria are occupied, the item is included in the misfit category or can also be regarded as invalid. With the help of the Ministep software, the results obtained from the test instrument are shown in Table 2.

Based on Table 2, it can be seen that nine items occupied all of the criteria and were considered valid, seven items were categorized as Fit 2 and still be regarded as valid, and two items were categorized as Fit 1 and were considered invalid. So, it can be concluded that 16 of 18 items were considered valid and can be used as a measuring instrument. When an item is declared invalid, there are two things that the researcher can do, the first is if the category of validity includes a misfit or the three existing criteria are not met, then the question must be discarded or replaced, while for other options when the item is invalid, and only 1 criterion is met (Fit 1), the item may be used on the condition that repairs must be made based on recommendations from expert validators. While instrument users can use the instrument that is considered the best, namely the one that has been declared valid, the test instrument in this study is intended to measure HOTS, which consists of analyzing, evaluating, and creating. Analyzing is a process of breaking matter down into smaller parts and determining the relationships between parts and the whole structure. Evaluating is defined as the process of making decisions based on criteria and standards. Meanwhile, creating is the process of arranging elements into a coherent or functional system [23].

Table 2. Item Validity

Item	Outfit MnSq	Outfit ZStd	PtMeaCorr	Category	Interpretation
I17	1.43	2.15	.13	Fit 1	Invalid
I13	1.42	2.55	.11	Fit 1	Invalid
I9	1.26	1.94	.26	Fit 2	Valid
I16	1.23	1.79	.39	Fit 2	Valid
I12	1.12	.62	.38	Fit 2	Valid
I15	1.07	.50	.41	Outfit	Valid
I7	1.06	.45	.34	Fit 2	Valid
I3	1	0	.38	Fit 2	Valid
I6	.94	-.49	.39	Fit 2	Valid
I2	.90	-.80	.45	Outfit	Valid
I4	.88	-.73	.41	Outfit	Valid
I14	.93	-.62	.44	Outfit	Valid
I1	.92	-.51	.42	Outfit	Valid
I5	.86	-.51	.38	Outfit	Valid
I8	.84	-1.20	.46	Outfit	Valid
I10	.82	-.89	.45	Outfit	Valid
I18	.69	-1.41	.49	Outfit	Valid
I11	.62	-2.09	.64	Fit 2	Valid

Item Reliability

With the help of Ministep, the researchers can obtain Cronbach Alpha value, person reliability, and item reliability. The Cronbach Alpha value measures the reliability of the interaction between the person (students) and the item (test item) as a whole [24]. Before categorizing the Cronbach Alpha value, person reliability, and item reliability, we need to consider the criteria of each reliability value. The reliability value of the instruments can be seen in Figure 2. Meanwhile, the Cronbach Alpha value, person reliability, and item reliability are grouped into categories that can be seen in Tables 3 and 4.

```
-------------------------------------------------------------------------------
|          TOTAL                          MODEL        INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       9.1      18.0       50.68    5.59                                |
|  SEM        .4       .0        1.18     .18                                |
| P.SD       3.3       .0       10.18    1.58                                |
| S.SD       3.3       .0       10.25    1.59                                |
| MAX.      18.0      18.0       93.51   18.43                                |
| MIN.       3.0      18.0       31.89    5.06                                |
|-----------------------------------------------------------------------------|
| REAL RMSE   6.02 TRUE SD    8.21  SEPARATION  1.36  Person RELIABILITY  .65 |
|MODEL RMSE   5.81 TRUE SD    8.36  SEPARATION  1.44  Person RELIABILITY  .67 |
| S.E. OF Person MEAN = 1.18                                                  |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .66  SEM = 1.91
STANDARDIZED (50 ITEM) RELIABILITY = .85

     SUMMARY OF 18 MEASURED (NON-EXTREME) Item
-------------------------------------------------------------------------------
|          TOTAL                          MODEL        INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      37.9      75.0       50.00    2.65       .99   -.02   1.00    .04 |
|  SEM       2.9       .0        1.93     .03       .03    .29    .05    .31 |
| P.SD      11.8       .0        7.97     .13       .14   1.19    .22   1.29 |
| S.SD      12.2       .0        8.20     .14       .15   1.22    .23   1.32 |
| MAX.      58.0      75.0       62.38    2.93      1.32   2.69   1.43   2.55 |
| MIN.      20.0      75.0       36.08    2.52       .75  -1.96    .62  -2.09 |
|-----------------------------------------------------------------------------|
| REAL RMSE   2.72 TRUE SD    7.49  SEPARATION  2.75  Item   RELIABILITY  .88 |
|MODEL RMSE   2.66 TRUE SD    7.51  SEPARATION  2.83  Item   RELIABILITY  .89 |
| S.E. OF Item MEAN = 1.93                                                    |
-------------------------------------------------------------------------------
```

Figure 2. Test Instrument Reliability

Table 3. Person dan Item Reliability Category

Correlation (r)	Category
0.94 < r	Excellent
0.91 < r ≤ 0.94	Very Good
0.81 < r ≤ 0.91	Good
0.67 < r ≤ 0.81	Moderate
r ≤ 0.67	Poor

Table 4. Cronbach Alpha Category

Correlation (r)	Category
0.80 < r	Very Good
0.70 < r ≤ 0.80	Good
0.60 < r ≤ 0.70	Acceptable
0.50 < r ≤ 0.60	Poor
r ≤ 0.50	Unacceptable

From Figure 2, the Cronbach alpha value is 0.66. According to Sumintono and Widhiarso's criteria, this value is interpreted as acceptable. In addition, the person's reliability value is 0.65. This shows that the reliability of the research sample is included in the poor category. In contrast, the item reliability is 0.88, which means that the reliability of this item test is good.

Reliability relates to the determination of test results. Reliability is a coefficient that shows the level of consistency of the measurement results of a test [25]. A test instrument can be considered good if it can consistently provide data that is under reality. A test may be reliable but may not be valid. So, it can be concluded that the test instrument in this study can provide steady results [26].

Item Difficulty Level

The item difficulty level indicates the extent to which the items are easy or difficult for students [27]. The difficulty level of each item was obtained using the two-parameter logistic response theory (2-PL) with the help of IRTPro for Student software. The item difficulty level in IRTPro software is called the threshold parameter (*b*) and is one of the useful parameters in analyzing a test because we can find out how good each item is in a test. The criteria for categorizing the item difficulty level can be seen in Table 5.

Based on the processed data, there is one item that was categorized as very difficult, which is number 18 with a *b* value of 10.32. Item number 5 with *b* value = 1.86 is included in the difficult category. Besides, some items have a very small *b* value, so they are included in the very easy category, such as item number 10 with a *b* value = -10.56 and item numbers 1, 4, and 7 with a *b* value of -1.84, -1.31, and -1.08 respectively. Meanwhile, the remaining 12 items are categorized as moderate. It can be concluded that most items are in the moderate category. A good item is one that is neither too easy nor too difficult [26]. Item level difficulty can be seen in Table 5.

Table 5. Item Level Difficulty

Item	Level Difficulty	Interpretation
1	-1.84	Easy
2	-0.43	Average
3	-0.46	Average
4	-1.31	Easy
5	1.86	Difficult
6	-0.64	Average
7	-1.08	Easy
8	-0.66	Average
9	-0.50	Average
10	-10.56	Very Easy
11	-0.16	Average
12	-0.25	Average
13	-0.40	Average
14	-0.47	Average
15	-0.21	Average
16	-0.56	Average
17	-0.44	Average
18	10.32	Very Difficult

Item Discrimination Index

Item discrimination index was obtained by two-parameter (2-PL) logistic Item Response Theory. With the help of IRTPro for Student software, the item discrimination index can be obtained from the table and item characteristic curves (ICC), where each graph for each item shows different results. The item discrimination index can be seen in Table 6.

Based on Table 6, it can be seen that the item discrimination index for each item is varied. Out of 18 items tested, the item discrimination index ranged from -0.27 to 3.12, and some of the items had a poor discrimination index.

Table 6. Item Discrimination Index

Item	Discrimination Index	Interpretation
1	0.27	Good
2	1.15	Good
3	1.08	Good
4	0.38	Good
5	-0.27	Poor
6	0.78	Good
7	0.46	Good
8	0.76	Good
9	0.99	Good
10	0.05	Good
11	3.12	Poor
12	2.03	Poor
13	1.25	Good
14	1.06	Good
15	2.37	Poor
16	0.90	Good
17	1.14	Good
18	-0.05	Poor

The discrimination index value (a) can be seen from the curve slope [29]. Based on Figures 3 and 4, item number 4 has a slower curve than item number 12. The greater the curve's degree of slope, the greater the discrimination index value of an item [30]. It means that item number 4 has a greater discrimination index value because the probability of students answering correctly has a wider range of ability than item number 12, which has the same probability but a smaller range of ability.

Items that can be answered correctly by students in the superior or less superior groups have a poor discrimination index. Likewise, if students in the superior or less superior groups cannot answer the question items correctly, then these items have a poor discrimination index. Nonetheless, good items are items that can be answered correctly by the superior group of students only [26].

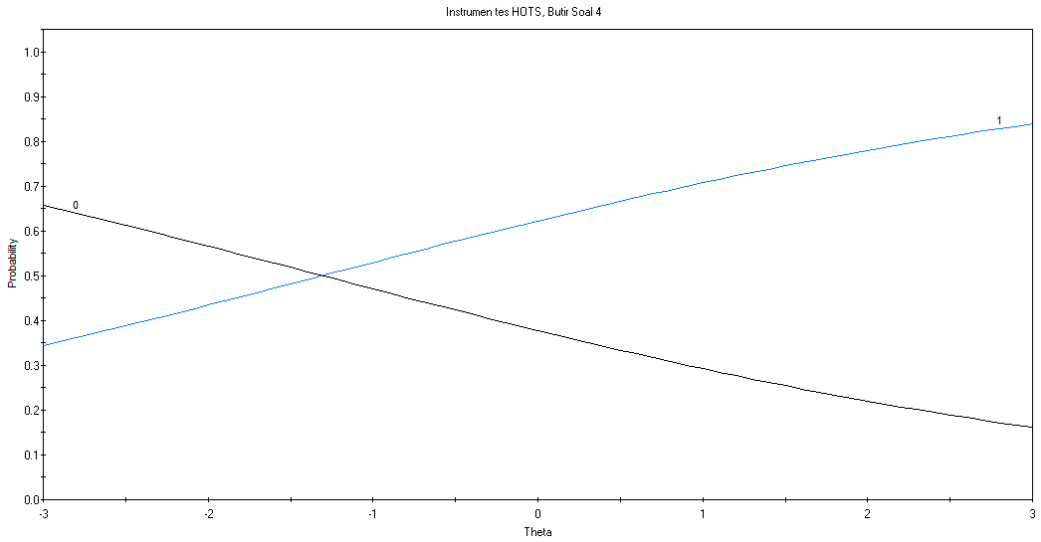

Figure 3. Item characteristic curves number 4 (good discrimination index)

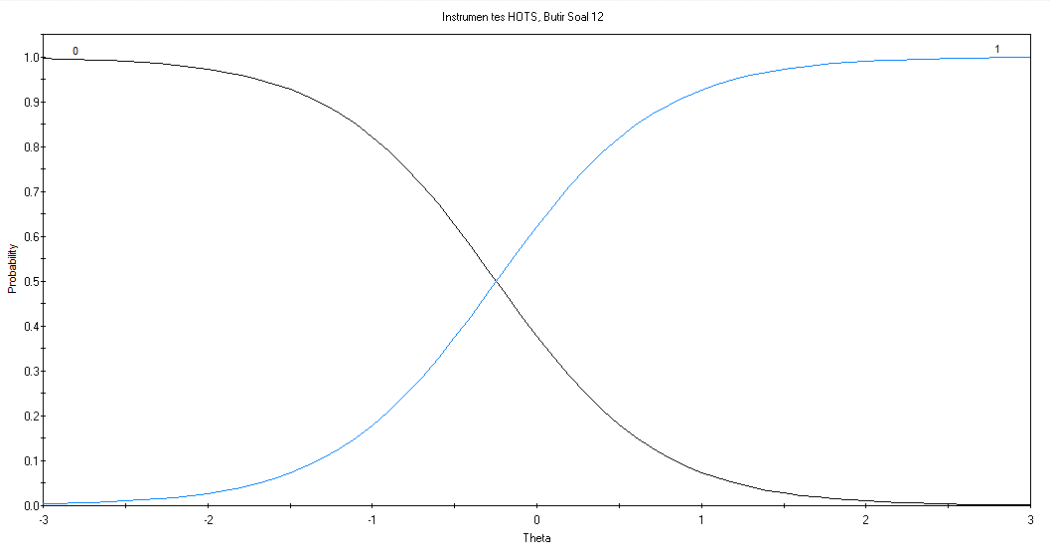

Figure 4. Item characteristic curves number 12 (poor discrimination index)

The scope of this research is limited to learning physics in the case study of Momentum and Impulse at the high school level. Therefore, researchers in the field of education must conduct further research with different materials and levels of education (classes) to detect students' higher-order thinking abilities.

Several recommendations from the author can be made to improve this research in the future so that this research can be more useful and can also be a reference for further research. Higher Order Thinking Skills (HOTS) instruments on momentum and impulse materials that were made in such a way in this study could be tested and distributed more widely to students who have studied momentum and impulse materials. In addition, as a recommendation for the future, it is hoped that more research samples will be sought, and samples shall come from various different schools.

This research can be used as a reference by educators or teachers in an effort to train students to get used to and have higher-order thinking skills. In addition, it can also be used as an evaluation material for students, which aims to obtain information about the higher-order thinking skills possessed by each student. At last, this research can also be used as a reference in developing assessment instruments.

IV. CONCLUSION

Based on the validity analysis, 16 of 18 items were considered valid and could be used to measure HOTS. Based on the reliability test, this test instrument is included in the good criteria. In other words, this test can be used to measure consistently. The level of difficulty of this test instrument varies: an item was considered as very difficult, an item was considered as difficult, three items were considered as easy, an item was considered as very easy, and 12 items were considered as medium or moderate. Based on its item discrimination index, 13 items have a good discrimination index so that these items can differentiate students from the superior and less superior groups of students well, then the remaining five items have a poor discrimination index. The High Order Thinking Skills (HOTS) test instrument on momentum and impulse material made in such a way in this study can be tested and distributed more widely to students who have studied momentum and impulse material.

REFERENCES

[1] Kementerian Pendidikan dan Kebudayaan. *Pendidikan Karakter Dorong Tumbuhnya Kompetensi Peserta didik Abad 21*. Jakarta: Kementerian Pendidikan dan Kebudayaan; 2017. Available from: https://www.kemdikbud.go.id/main/blog/2017/06/pendidikan-karakter-dorong-tumbuhnya-kompetensi-siswa-abad-21#.

[2] Anderson LW and Krathwohl D. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman; 2001.

[3] Kementerian Pendidikan dan Kebudayaan. *Bahan Uji Publik Kurikulum 2013*. Jakarta: Kementerian Pendidikan dan Kebudayaan; 2013. Available from: https://www.kemdikbud.go.id/kemdikbud/dokumen/Paparan/Paparan%20Wamendik.pdf.

[4] Istiyono E, Mardapi D, and Suparno. Pengembangan Tes Kemampuan Berpikir Tingkat Tinggi Fisika (PysTHOTS) Peserta Didik SMA. *Jurnal Penelitian dan Evaluasi Pendidikan*. 2014; **18**(1): 1-12. DOI: https://doi.org/10.21831/pep.v18i1.2120.

[5] Sanjaya W. *Kurikulum dan Pembelajaran*. Jakarta: Kencana Prenada Media Group; 2008.

[6] Kementerian Pendidikan dan Kebudayaan. *Laporan Hasil Ujian Nasional*. Jakarta: Kementerian Pendidikan dan Kebudayaan; 2019. Available from: https://puspendik.kemdikbud.go.id/hasil-un/.

[7] OECD. *PISA 2018 Results (Volume I): What Students Know and Can Do*. French: OECD Publishing; 2019. DOI: https://doi.org/10.1787/5f07c754-en.

[8] Arifin Z. *Evaluasi Pembelajaran.* Bandung: PT Remaja Rosdakarya; 2009.

[9] Malik A, Rosidin U, and Ertikanto C. Pengembagan Instrumen Asesmen HOTS Fisika SMA Menggunakan Model Inkuiri Terbimbing. *Jurnal Lentera Pendidikan Pusat Penelitian LPPM UM METRO.* 2018; **3**(1): 11-25. Available from: https://ojs.ummetro.ac.id/index.php/lentera/article/view/733.

[10] Anwar S. *Penilaian Berbasis Kompetensi*. Padang: UNP Press; 2009.

[11] Anisa. Perbandingan Penskoran Dikotomi dan Politomi dalam Teori Respon Butir untuk Pengembangan Bank Soal Matakuliah Matematika Dasar. *Jurnal Matematika Statistika & Komputasi*. 2013; **9**(2): 95-113. Available from: https://journal.unhas.ac.id/index.php/jmsk/article/view/3402.

[12] Lord FM. *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum; 1980.

[13] Naga DS. *Pengantar Teori Skor pada Pengukuran Pendidikan*. Jakarta: Penerbit Gunadarma; 1992.

[14] Rakkapao S, Prasitpong S, and Arayathanitkul K. Analysis test of understanding of vectors with three-parameter logistic model of item response theory and item response curve technique. *Physical Review Physics Education Research.* 2016; **12**(2): 020135. DOI: https://doi.org/10.1103/PhysRevPhysEducRes.12.020135.

[15] Rahmat. *Panduan Analisis Butir tes*; 2010. Available from: http://gurupembaharu.com/home/download/panduan-analisis-butir-soal.pdf.

[16] Etkina E, Gitomer D, Iaconangelo C, Phelps G, Seeley L, and Vokos S. Design of An Assessment to Probe Teacher's Content Knowledge for Teaching: An Example from Energy in High School Physics. *Physical Review Physics Education Research*. 2018; **14**: 010127. DOI: https://doi.org/10.1103/PhysRevPhysEducRes.14.010127.

[17] Alnasraween MS, Al-mughrabi AM, Ammari RM, and Alkaramneh MS. Validity and Reliability of Eigh-Grade Digital Culture Test in Light of Item Response Theory.

Cypriot Journal of Education Science. 2021; **16**(4) 1816-1835. DOI: https://doi.org/10.18844/cjes.v16i4.6034.

[18] Sudjana N. *Penelitian dan Penilaian Pendidikan*. Bandung: Sinar Baru Algesindo; 2004.

[19] Arikunto S. *Prosedur Penelitian Suatu Pendekatan Praktik*. Jakarta: PT. Rineka Cipta; 2019.

[20] Sugiyono. *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. Bandung: Alfabeta, CV; 2017.

[21] Hambleton RK and Swaminathan H. *Item Response Theory Principles and Application*. Boston, MA: Kluwer Inc; 1985.

[22] Sumintono B and Widhiarso W. *Aplikasi Model Rasch untuk Penelitian Ilmu-ilmu Sosial*. Jakarta: Tim Komunikata Publishing House; 2013.

[23] Anderson LW and Krathwohl D. *Kerangka Landasan untuk Pembelajaran, Pengajaran, dan Asesmen: Revisi Taksonomi Pendidikan Bloom (Translated by Prihantoro)*. Yogyakarta: Pustaka Pelajar; 2010.

[24] Sumintono B and Widhiarso W. *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Trim Komunikata; 2015.

[25] Asbupel FMD and Sanova A. Pengembangan Instrumen Tes kemampuan Berpikir Tingkat Tinggi Kimia. *Repository Universitas Jambi.* 2018; 1–11. Available from: https://repository.unja.ac.id/id/eprint/3024.

[26] Arikunto S. *Dasar-Dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara; 2015.

[27] Setiyawan RA and Wijayanti PS. Analisis Kualitas Instrumen Untuk Mengukur Kemampuan Pemecahan Masalah Siswa Selama Pembelajaran Daring Di Masa Pandemi. *Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika*. 2020; **1**(2): 130–139. DOI: https://doi.org/10.46306/lb.v1i2.26.

[28] Hakim ML, Ramalis TR, and Muslim. Karakteristik Tes Hasil Belajar Ranah Kognitif Materi Elastisitas Menggunakan Analisis Item Response Theory. *Jurnal Penelitian Pembelajaran Fisika*. 2019; **10**(1): 22-32. DOI: https://doi.org/10.26877/jp2f.v10i1.3318.

[29] Saptawulan W. *Karakterisasi Tes Penalaran Ilmiah Materi Suhu dan Kalor Berdasarkan Teori Respon Butir*. Thesis. Bandung: Universitas Pendidikan Indonesia; 2018. Available from: http://repository.upi.edu/40844/.

[30] Retnawati H. *Teori Respon Butir dan Penerapannya.* Yogyakarta: Nuha Medika; 2014.