Development Two Tier Multiple Choice for Measure Students' Higher-Order Thinking Skills in Physics Learning

Misykah Aulia Anwar[1*], Muslim[1], Endi Suhendi[1]

1 Universitas Pendidikan Indonesia, Bandung, Indonesia

* E-mail: misykahaulia10@gmail.com

ABSTRACT

The two-tier multiple-choice question comprises two levels: the initial level centers on understanding the materials' notion. Conversely, the second tier elucidates the rationale for the correctness of the response provided in the first tier. Higher-order thinking skills can be assessed using a two-tier multiple-choice format. This study aims to create a set of two-tier multiple-choice questions that can effectively assess students' advanced cognitive abilities in physics. The research development employed the ADDIE development model, encompassing the five phases of analysis, design, development, implementation, and evaluation. The two-tier multiple-choice test comprises 21 questions assessing the capacity to analyze, evaluate, and generate knowledge on sound waves. The participant group comprised 99 eleventh-grade students from Senior High School in Bandung. The content validity was assessed using the Aiken formula, while the empirical validity and reliability were analyzed using the Rasch model through the Winstep Application. The two-tier multiple-choice assessment method is applicable and viable for evaluating higher-order thinking skills (HOTS) in sound wave content.

KEYWORDS

higher order thinking skills; physics learning; two tier multiple choice

1. INTRODUCTION

Human existence is changing rapidly as we enter the 21st century, particularly in the technology sphere, where numerous technologies can replace many human jobs, causing some old jobs to vanish and be replaced by new ones. The solution to the current educational challenges is higher-order thinking skills (HOTS)(Tanudjaya & Doorman, 2020). HOTS incorporate the capacity to analyze (C4), assess (C5), and create or be creative (C6) (Suprapto et al., 2020). When someone practices HOTS, they take new knowledge, retain it, and expand it to search for connections, leading to achieving goals or discovering answers after experiencing perplexity (Hubers, 2022). Based on the comes about of the PISA national report in 2018, a few Indonesian understudies were as it were able to choose the most excellent logical clarification for information displayed in a common setting, whereas the rest were at a lower level (Pusat Penilaian Pendidikan Balitbang Kemendikbud, 2019).

HOTS and physics are connected to moved forward learning results (Kahar et al., 2021). Physics will aid students in developing their critical thinking abilities, as they will need to figure out HOTS. Teachers should also encourage students to use educational materials like the HOTS, which the HOTS instrument can enhance (Widyaningsih et al., 2021).

HOTS can be measured by reasoning multiple-choice questions, which has become common as two-tier multiple choice (TTMC) (Istiyono et al., 2020). The first tier of the TTMC focuses on concepts, whereas the second level (tier II) explains why the level I answer is correct. The second tier will help enhance HOTS since it involves more complex thinking and won't directly ask about the concepts being tested (Andriyatno et al., 2023). TTMC level 1 study presents problem-based questions related to physics concepts.

HOTS training activities benefit students, including increasing their readiness to face a developing and challenging era and improving their ability to socialize with the community. Students who successfully use HOTS can be seen from their explanations and decisions to solve problems or choose existing options (Ramadhan et al., 2019). Therefore, developing two-tier multiple-choice questions is done by presenting problem-based questions to find solutions.

2. METHODOLOGY

Research and development (R&D) using the ADDIE model is the methodology used in this study. The phases of ADDIE include (1) Analysis, (2) Design, (3) Development, (4) Implementation, and (5) Evaluation, as seen in Figure 1 (Widyaningsih et al., 2021).

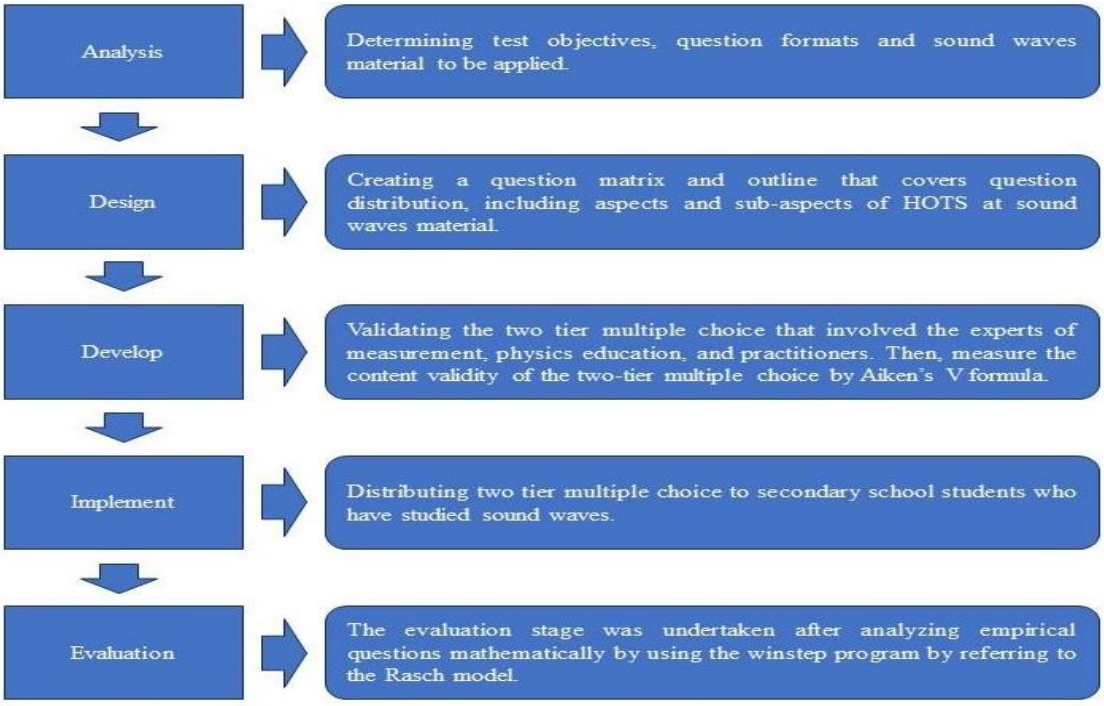


Figure 1. Stage of the ADDIE model in designing two-tier multiple-choice

Population and sample used. This study's population consisted of eleventh-grade high school students in Bandung. Purposive sampling was employed to select the sample, which involved research on sound waves. There were 99 senior high school students who participated in this study.

Data Collection Techniques. Expert validation data was collected using a questionnaire. The questionnaires were distributed to 7 assessment experts, HOTS, and practitioners. Then, the two-tier multiple-choice test data was collected through a Google Form.

Tools or Instruments Used. The two-tier multiple-choice construction evaluation was gathered using a Likert scale questionnaire. The questionnaire assessed the suitability of the questions in measuring HOTS indicators (Aviory & Susetyawati, 2021). The rating scale is 1 to 5, where 1 represents very unsuitable, 2 represents unsuitable, 3 represents enough, 4 represents suitable, and 5 represents very suitable.

Data Analysis Methods.

Two-tier multiple choice is measured for content and empirical validity. The validity of content results was calculated using the Aiken V formula. The validity of the content analysis method applied the Aiken formula as shown below:

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

where V is the index of validator agreement about the items' validity, S is the score of the validator's assessment subtracting the lowest score of the assessment, n is the number of validators who evaluate the instrument, and c is the total of categories that can be selected by the validator (Yusuf & Widyaningsih, 2022). The item is valid if the V Aiken index value equals or exceeds 0,8 (Aiken, 1980). Then, empirical validity was applied to the Rasch model using Winstep software. The empirical validity of items is assessed from the value of OUTFIT MNSQ, OUTFIT ZSTD, and PT-MEASURE CORR at item: measure table (Sumintono & Widhiarso, 2015). The requirements for items to be valid follow:

a. The passed Oufit mean square (MNSQ): 0,5 < MNSQ < 1,5 ;

b. The passed Oufit Z-standard (ZSTD): -2,0 < ZSTD < +2,0 ;

c. The passed Point Measure Correlation (PT MEAN CORR): 0,4 < PT MEAN CORR < 0,85.

The reliability of two-tier multiple-choice assessments is evaluated using Cronbach's alpha. The reliability criteria, as indicated in Table 1, are then used to understand the two-tier multiple choice's reliability. The reliability of items and people can be measured separately, with person and item reliability in the summary statistics table. The criteria of item and person reliability are interpreted in Table 2 (Sumintono & Widhiarso, 2015).

Table 1. The criteria of reliability measured by Cronbach alpha

Cronbach alpha value	Category
< 0,5	Poor
0,5-0,6	Bad
0,6-0,7	Enough
0,7-0,8	Good
> 0,8	Very good

Table 2. The criteria of item and person reliability

Item reliability/Person reliability	Category
< 0,67	Weak
0,67-0,8	Enough
0,8-0,9	Good
0,91-0,94	Very Good
> 0,94	Excellent

The logit person's measure value can determine the mean score of all students working on the items given. The person-measure value can be seen in the summary statistics table menu in the Winstep output tables. If the person measure value is smaller than the logit value of 0.0, the student's ability tendency is smaller than the question difficulty level (Sumintono & Widhiarso, 2015).

Question difficulty level is known from the logit measure value on the item measure table. Question difficulty levels are grouped based on information on the average logit value with a standard deviation value (SD). The question difficulty level is categorised based on Table 3 (Sumintono & Widhiarso, 2015).

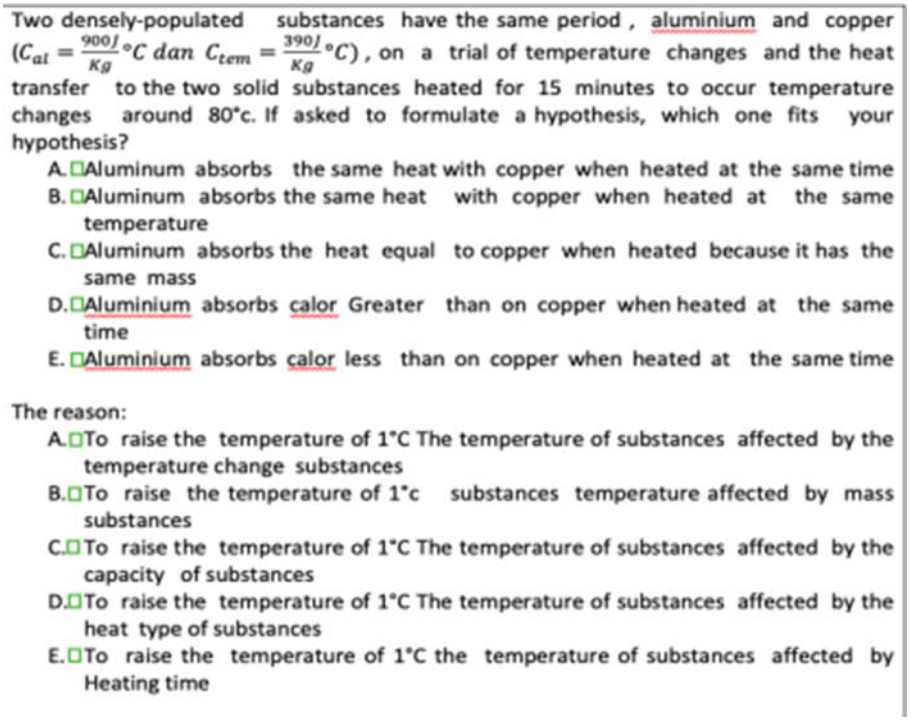
Table 3. The level of question difficulty

Measure Value (logit)	Category
logit < -1SD	Very easy
- 1SD $\leq$ *logit* $\leq$ 0,0	Easy
0,0 $<$ *logit* $\leq$ +SD	Difficult
logit > +1SD	Very difficult

3. RESULTS AND DISCUSSION

3.1. Development of Two-Tier Multiple Choice for Measure HOTS

The two-tier multiple-choice instrument employed to assess higher-order thinking skills in prior studies is constructed as follows: (1) Tier 1 questions are the primary physics-related questions that can be answered by applying the skills of analysis (C4), evaluation (C5), and creation (C6); (2) tier 2 questions ask students to select the rationale behind their tier 1 answer selection; (3) tier 1 answer selections are the solutions to the problems provided; and (4) tier 2 answer selections are the rationales behind the students' tier 1 answer selections (Istiyono et al., 2020). The form of the two-tier multiple-choice developed by Istiyono can be seen in Figure 2.



Two densely-populated substances have the same period , aluminium and copper ($C_{al} = \frac{900J}{Kg}°C$ *dan* $C_{tem} = \frac{390J}{Kg}°C$) , on a trial of temperature changes and the heat transfer to the two solid substances heated for 15 minutes to occur temperature changes around 80°c. If asked to formulate a hypothesis, which one fits your hypothesis?

A. Aluminum absorbs the same heat with copper when heated at the same time
B. Aluminum absorbs the same heat with copper when heated at the same temperature
C. Aluminum absorbs the heat equal to copper when heated because it has the same mass
D. Aluminium absorbs calor Greater than on copper when heated at the same time
E. Aluminium absorbs calor less than on copper when heated at the same time

The reason:

A. To raise the temperature of 1°C The temperature of substances affected by the temperature change substances
B. To raise the temperature of 1°c substances temperature affected by mass substances
C. To raise the temperature of 1°C The temperature of substances affected by the capacity of substances
D. To raise the temperature of 1°C The temperature of substances affected by the heat type of substances
E. To raise the temperature of 1°C the temperature of substances affected by Heating time

Figure 2. Two-tier multiple choice
(Source: Istiyono, 2020)

The two-tier multiple-choice design referred to Istiyono's research (2020) using different physics materials, such as sound waves. The two-tier question should be compatible with HOTS indicators (Obeidat & Saleh, 2022; Suprapto et al., 2020). The distribution of two-tier multiple-choice questions that suit the HOTS category can be seen in Table 4. After completing the two-tier multiple-choice design, a validation sheet must be created to assess the instrument's content validity. The validation sheet was prepared to measure the suitability of the questions in measuring HOTS indicators(Aviory & Susetyawati, 2021).

Table 4. Distribution of HOTS indicators on two-tier multiple-choice

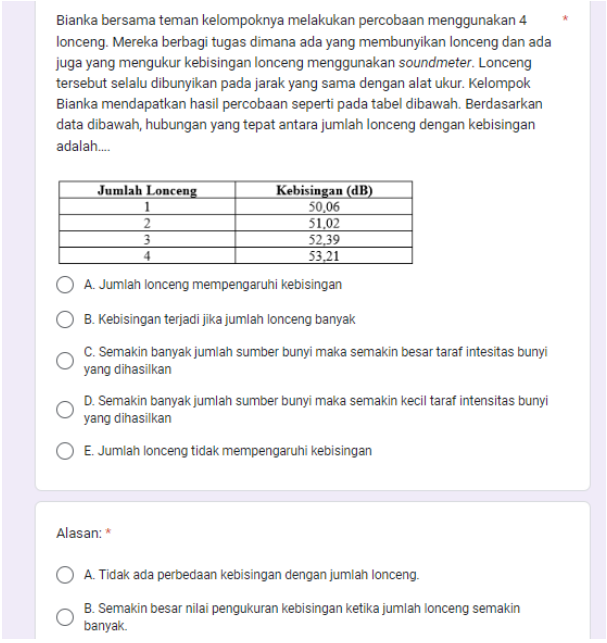
HOTS indicator	Item Number
Analyze	1, 2, 7, 10, 13, 16, 19, 22
Evaluate	3, 4, 8, 11, 14, 17, 20, 23
Create	5, 6, 9, 12, 15, 18, 21, 24

3.2. The Validity of Content fot Two-Tier Multiplier Choice

The validation results were calculated using the Aiken V formula, and it was found that 21 questions were valid because the Aiken V index value was equal to or greater than 0.8 (Aiken, 1980). In addition to providing scores, the experts also provided suggestions that became the basis for improving the questions and answer choices in the two-tier multiple choice.

3.3. Implementation of Two-Tier Multiple Choice for Measure HOTS

After improving the two-tier multiple-choice questions according to expert suggestions, they were made in a Google Form, as shown in Figure 3, to be tested on secondary students. This study involved several secondary schools where students had studied sound waves. The questions were distributed to students during class via a Google Form link. The activity of testing the two-tier multiple-choice questions can be seen in Figure 4. A total of 99 students answered the two-tier multiple-choice questions.


Bianka bersama teman kelompoknya melakukan percobaan menggunakan 4 lonceng. Mereka berbagi tugas dimana ada yang membunyikan lonceng dan ada juga yang mengukur kebisingan lonceng menggunakan *soundmeter*. Lonceng tersebut selalu dibunyikan pada jarak yang sama dengan alat ukur. Kelompok Bianka mendapatkan hasil percobaan seperti pada tabel dibawah. Berdasarkan data dibawah, hubungan yang tepat antara jumlah lonceng dengan kebisingan adalah.... *

Jumlah Lonceng	Kebisingan (dB)
1	50,06
2	51,02
3	52,39
4	53,21

- A. Jumlah lonceng mempengaruhi kebisingan
- B. Kebisingan terjadi jika jumlah lonceng banyak
- C. Semakin banyak jumlah sumber bunyi maka semakin besar taraf intesitas bunyi yang dihasilkan
- D. Semakin banyak jumlah sumber bunyi maka semakin kecil taraf intensitas bunyi yang dihasilkan
- E. Jumlah lonceng tidak mempengaruhi kebisingan

Alasan: *

- A. Tidak ada perbedaan kebisingan dengan jumlah lonceng.
- B. Semakin besar nilai pengukuran kebisingan ketika jumlah lonceng semakin banyak.

Figure 3. Google Form for distributing the two-tier multiple-choice (Source: personal documents)

Figure 4. Students work TTMC by filling out Google Forms on their mobile phones. (Source: personal documents)

3.4. The Empirical Validity of Two-Tier Multiple Choice

The two-tier multiple-choice responses provided by students are converted into Excel data for Rasch model analysis with the Winstep tool. After that, the tables provided in the application can be used to analyze the two-tier multiple-choice questions. In this section, the validity, reliability, and degree of difficulty of the questions on the designed two-tier multiple-choice test will be analyzed.

3.4.1 Item Validity

Item validity can be determined by looking at the item measure table in Figure 5 below. First, the outfit MNSQ value in Table 5 is compared with the accepted value of 0.5 < MNSQ < 1.5. Based on the analysis, all questions fulfill the accepted outfit MNSQ value. Second, the outfit ZSTD value in Table 5 is compared with the accepted value of -2,0 < ZSTD <+2,0. After the scores were analyzed, it was found that two questions didn't pass the acceptable outfit ZSTD value. Question number 9 doesn't fulfill the accepted value because it has an outfit ZSTD value greater than 2, namely a value of 2.48. In contrast, question number 17 doesn't fulfill the accepted value because it has an outfit MNSQ value smaller than -2, namely -2.91. Last, the point measure correlation (PTMEASUR CORR) value in Figure 5 is compared with the accepted value of 0,4 < PTMEASUR CORR < 0,85. The two-tier multiple choice questions that have fulfilled this value are questions number 4,5,6,7,9,10.

INPUT: 99 Person 21 Item REPORTED: 99 Person 21 Item 2 CATS WINSTEPS 5.4.1.0

Person: REAL SEP.: 1.17 REL.: .58 ... Item: REAL SEP.: 3.70 REL.: .93

Item STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
18	3	99	3.33	.59	1.07	.30	1.31	.63	.03	.15	97.0	97.0	N18
17	21	99	1.02	.26	1.19	1.30	1.58	2.48	.01	.32	78.8	79.4	N17
15	24	99	.83	.25	1.11	.85	1.19	1.07	.18	.33	77.8	76.9	N15
21	27	99	.65	.24	1.09	.77	1.15	.96	.22	.34	70.7	74.6	N21
13	28	99	.59	.24	1.02	.26	1.00	.03	.32	.34	73.7	73.9	N13
19	30	99	.48	.23	1.08	.81	1.18	1.28	.22	.34	75.8	72.5	N19
7	35	99	.22	.22	.83	-1.98	.82	-1.57	.54	.35	77.8	69.4	N7
12	35	99	.22	.22	1.06	.65	1.12	1.03	.27	.35	69.7	69.4	N12
16	36	99	.17	.22	1.02	.29	1.02	.24	.33	.35	66.7	68.9	N16
14	37	99	.12	.22	1.05	.62	1.06	.61	.29	.35	63.6	68.3	N14
8	45	99	-.26	.22	1.17	2.21	1.17	1.82	.17	.35	53.5	65.5	N8
20	46	99	-.31	.22	1.08	1.11	1.07	.75	.27	.35	60.6	65.2	N20
2	48	99	-.40	.22	1.01	.14	1.02	.29	.34	.35	62.6	64.9	N2
11	48	99	-.40	.22	.98	-.34	.95	-.55	.39	.35	62.6	64.9	N11
3	51	99	-.54	.21	.98	-.33	.96	-.37	.38	.35	63.6	64.7	N3
5	53	99	-.63	.22	.91	-1.34	1.01	.13	.43	.35	73.7	64.6	N5
9	53	99	-.63	.22	.80	-3.12	.75	-2.91	.59	.35	73.7	64.6	N9
10	53	99	-.63	.22	.89	-1.64	.88	-1.37	.48	.35	71.7	64.6	N10
6	56	99	-.77	.22	.85	-2.27	.84	-1.75	.52	.35	76.8	65.2	N6
1	64	99	-1.16	.22	1.02	.23	.95	-.41	.33	.33	63.6	68.4	N1
4	77	99	-1.88	.25	.79	-1.64	.64	-1.91	.55	.28	79.8	78.0	N4
MEAN	41.4	99.0	.00	.24	1.00	-.15	1.03	.02			71.1	70.5	
P.SD	16.1	.0	1.01	.08	.11	1.32	.20	1.31			9.0	7.6	

Figure 5. Item Measure of Two Tier Multiple Choice
(Source: Winstep Application)

Two-tier multiple-choice question items can be valid if it has fulfilled two categories among the accepted outfit MNSQ, outfit ZSTD, and point measure correlation value (Sumintono & Widhiarso, 2014). Based on the analysis results, all questions have met both criteria, so the two-tier multiple choice is valid. Empirical validity testing is carried out to determine the reliability level of the assessment instrument developed (Dewi et al., 2020). Therefore, the Two-tier multiple choice is valid for measuring students' HOTS.

3.4.2 Reliability

Item reliability indicates the quality of the items in the instrument(Sumintono & Widhiarso, 2015). The item reliability value can be seen in Figure 6. The figure shows that the value is 0.93. Furthermore, the value is compared with Table 2. The value of 0.93 is included in very good. Based on this result, the items in two-tier multiple-choice questions are very good for testing HOTS.

SUMMARY OF 21 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	41.4	99.0	.00	.24	1.00	-.15	1.03	.02
SEM	3.6	.0	.23	.02	.03	.30	.04	.29
P.SD	16.1	.0	1.01	.08	.11	1.32	.20	1.31
S.SD	16.5	.0	1.03	.08	.12	1.35	.20	1.34
MAX.	77.0	99.0	3.33	.59	1.19	2.21	1.58	2.48
MIN.	3.0	99.0	-1.88	.21	.79	-3.12	.64	-2.91

REAL RMSE	.26	TRUE SD	.97	SEPARATION 3.70	Item	RELIABILITY .93
MODEL RMSE	.26	TRUE SD	.98	SEPARATION 3.80	Item	RELIABILITY .94
S.E. OF Item MEAN = .23						

Item RAW SCORE-TO-MEASURE CORRELATION = -.96 (approximate due to missing data)

Figure 6. Item Reliability
(Source: Winstep Application)

Person reliability provides information about the consistency of student answers (Sumintono & Widhiarso, 2015). The person reliability value can be seen in Figure 7. The figure shows that the value is 0,58. Furthermore, the value is compared with Table 2. The value of 0,58 is included in weak. Based on this result, students' consistency in answering two-tier multiple-choice questions is weak.

SUMMARY OF 99 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	8.8	21.0	-.46	.50	1.00	-.04	1.03	.06
SEM	.3	.0	.08	.00	.02	.08	.04	.08
P.SD	3.3	.0	.81	.04	.16	.80	.42	.81
S.SD	3.3	.0	.81	.04	.16	.81	.42	.82
MAX.	18.0	21.0	2.09	.69	1.32	1.80	3.74	3.91
MIN.	3.0	21.0	-2.07	.47	.72	-2.12	.59	-1.13

REAL RMSE	.52	TRUE SD	.61	SEPARATION 1.17	Person	RELIABILITY .58
MODEL RMSE	.51	TRUE SD	.63	SEPARATION 1.24	Person	RELIABILITY .61
S.E. OF Person MEAN = .08						

Person RAW SCORE-TO-MEASURE CORRELATION = 1.00 (approximate due to missing data)
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .61 SEM = 2.06 (approximate due to missing data)
STANDARDIZED (50 ITEM) RELIABILITY = .79

Figure 7. Person Reliability
(Source: Winstep Application)

Last, the reliability of two-tier multiple-choice can be known from the Cronbach alpha value in Figure 7 (Sumintono & Widhiarso, 2015). The Cronbach alpha value is 0.61. The Cronbach alpha value is compared with the categories in Table 1 so that the value is included enough. Based on these results, two-tier multiple-choice is acceptable to measure students' HOTS in sound wave material (Azizah et al., 2021).

3.4.3 The Question Difficulty Level

The function of the question difficulty level is to understand the difficulty level of each question and design a balance between the question and the student's abilities (Siregar et al., 2023). The standard deviation value (SD) should have been known to determine the question difficulty level. From Table 5, the standard deviation value (SD) is 1,01. After that, the category of question difficulty level can be established, as shown in Table 5.

Table 5. The level of question difficulty

Measure Value (*logit)*	Category
logit < -1,01	Very easy
- 1,01 ≤ *logit* ≤ 0,0	Easy
0,0 < $logit$ ≤ +1,01	Difficult
logit > +1,01	Very difficult

Table 5 provides each item's measure value (logit) in a two-tier multiple-choice test. So, that logit is compared with the categories in Table 5 to obtain the data processing as in Table 6.

Table 6. The question difficulty level in two-tier multiple-choice

Number Question	Measure Value (*logit*)	Category
1	-1,16	Very easy
2	-0,40	Easy
3	-0,54	Easy
4	-1,18	Very easy
5	-0,63	Easy
6	-0,77	Easy
7	0,22	Difficult
8	-0,26	Easy
9	-0,63	Easy
10	-0,63	Easy
11	-0,40	Easy
12	0,22	Difficult
13	0,59	Difficult
14	0,12	Difficult
15	0,83	Difficult
16	0,17	Difficult
17	1,02	Very Difficult
18	3,33	Very Difficult
19	0,48	Difficult
20	-0,31	Easy
21	0,65	Difficult

Table 6 shows that two-tier multiple-choice questions have difficulty levels ranging from very easy to very difficult. Questions with a very easy level are in numbers 1 and 4. Easy questions are in numbers 2, 3, 5, 6, 8, 9, 10, 11, and 20. Difficult questions are 7, 12, 13, 14, 15, 16, and 21. Finally, questions with a very difficult level are questions 17 and 18. The distribution of question difficulty levels for two-tier multiple choice can be seen in Figure 8. From the graph, the two-tier multiple choice is dominated by questions with easy and difficult levels that are almost the same proportion. In addition, the two-tier multiple-choice questions are very easy and very difficult to answer correctly in the same proportion.

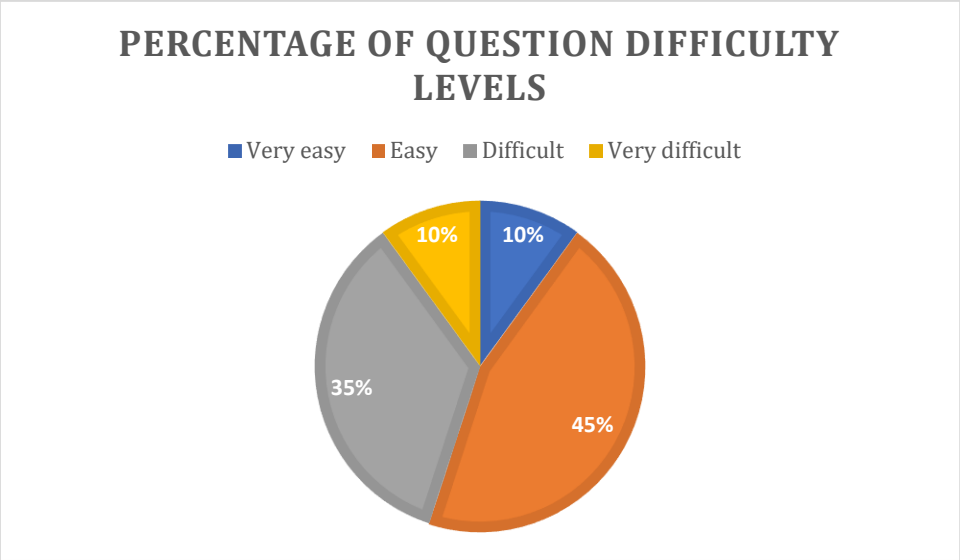


Figure 8. Percentage of question difficulty levels in two-tier-multiple-choice (Source: personal documents with excel)

4. CONCLUSION

Based on this research, the conclusions are (1) HOTS can be measured with two-tier multiple-choice; (2) two-tier multiple-choice is valid as long as it fulfills two accepted values among the outfit MNSQ, outfit ZSTD, and point measure correlation values; (3) two-tier multiple-choice is reliable because it meets the accepted Cronbach alpha value; (4) two-tier multiple-choice is built with a variety of questions with varying difficulty levels (ranging from very easy to very difficult). In addition, students still need to practice with HOTS questions to have strong consistency in their answers.

REFERENCES

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, *40(4)*, 955–959.

Andriyatno, I., Zulfiani, Z., & Mardiati, Y. (2023). Higher Order Thinking Skills: Student Profile Using Two-Tier Multiple Choice Instrument. *International Journal of STEM Education for Sustainability*, *3*(1), 111–124. https://doi.org/10.53889/ijses.v3i1.79

Aviory, K., & Susetyawati, M. M. . (2021). KUALITAS SOAL HOTS (HIGH ORDER THINKING SKILL) PADA PESERTA DIDIK SMP KELAS VII. *AKSIOMA*, *10*(2), 639–647.

Azizah, A., Wahyuningsih, S., Kusumasari, V., Asmianto, A., & Setiawan, D. (2021). Validity and reliability of mathematical instruments in online learning using the Rasch measurement model at UM lab school. *AIP Conference Proceedings*, *2330*(March). https://doi.org/10.1063/5.0043356

Dewi, N. P., Rahmi, Y. L., Alberida, H., & Darussyamsu, R. (2020). Validitas dan reliabilitas instrumen penilaian kemampuan berpikir tingkat tinggi tentang materi hereditas untuk peserta didik SMA/MA [The validity and reliability of the high-order thinking ability assessment instrument on heredity material for senior high school students]. *Jurnal Eksakta Pendidikan (Jep)*, *4*(2), 138.

Hubers, M. D. (2022). Using an Evidence-Informed Approach to Improve Students' Higher Order Thinking Skills. *Education Sciences*, *12*(11). https://doi.org/10.3390/educsci12110834

Istiyono, E., Dwandaru, W. S. B., Setiawan, R., & Megawati, I. (2020). Developing of computerized adaptive testing to measure physics higher order thinking skills of senior high school students and its feasibility of use. *European Journal of Educational Research*, *9*(1), 91–101. https://doi.org/10.12973/eu-jer.9.1.91

Kahar, M. S., Syahputra, R., Arsyad, R. Bin, Nursetiawan, N., & Mujiarto, M. (2021). Design of Student Worksheets Oriented to Higher Order Thinking Skills (HOTS) in Physics Learning. *Eurasian Journal of Educational Research*, *2021*(96), 14–29. https://doi.org/10.14689/ejer.2021.96.2

Obeidat, F. A. A., & Saleh, S. (2022). The Relationship of Fluid Intelligence Level with Higher-order Thinking Skills in Work and Energy among Sixth-grade Students in Jordan. *Journal of Curriculum and Teaching*, *11*(4), 224–234. https://doi.org/10.5430/jct.v11n4p224

Pusat Penilaian Pendidikan Balitbang Kemendikbud. (2019). Pendidikan di Indonesia Belajar dari Hasil PISA 2018. *Pusat Penilaian Pendidikan Balitbang Kemendikbud*, *021*, 1–206.

Ramadhan, S., Mardapi, D., Prasetyo, Z. K., & Utomo, H. B. (2019). The development of an instrument to measure the higher order thinking skill in physics. *European Journal of Educational Research*, *8*(3), 743–751. https://doi.org/10.12973/eu-jer.8.3.743

Siregar, P. S., Hatika, R. G., & Hayadi, B. H. (2023). Multiple Choice Question Difficulty Level Classification with Multi Class Confusion Matrix in the Online Question Bank of Education Gallery. *Journal of Applied Data Sciences*, *4*(4), 392–406. https://doi.org/10.47738/jads.v4i4.132

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial* (1st ed.). Trim Komunikata.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan* (1st ed.). Trim Komunikata.

Suprapto, E., Saryanto, S., Sumiharsono, R., & Ramadhan, S. (2020). The Analysis of Instrument Quality to Measure the Students' Higher Order Thinking Skill in Physics Learning. *Journal of Turkish Science Education*, *17*(4), 520–527. https://doi.org/10.36681/tused.2020.42

Tanudjaya, C. P., & Doorman, M. (2020). Examining higher order thinking in Indonesian lower secondary mathematics classrooms. *Journal on Mathematics Education*, *11*(2), 277–300. https://doi.org/10.22342/jme.11.2.11000.277-300

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. https://doi.org/10.29333/iji.2021.1444a

Yusuf, I., & Widyaningsih, S. W. (2022). Higher Order Thinking Skills Oriented Student Worksheet of E-learning Model in Electric Circuit Topic. *TEM Journal*, *11*(2), 564–573. https://doi.org/10.18421/TEM112-10