Prosiding Semianr Nasional Fisika & International Physics Conference (2024) 3 (1): 60-70
http://proceedings.upi.edu/index.php/sinafi
ISSN 2986-9544 (Online)

Item analysis of the conceptual understanding test on momentum and impulse using the Rasch model

Dena Tresna Ningsih, Muhammad Zahran, Irma Rahma Suwarma, Hera Novia

Artikel ini telah dipresentasikan pada kegiatan Seminar Nasional Fisika (Sinafi X) & International Physics Conference (IPC)
Universitas Pendidikan Indonesia, Bandung, Indonesia
9 November 2024

Abstract

This study aims to analyze the item quality of a conceptual understanding test on momentum and impulse. The analysis includes validity, reliability, difficulty level, and item discrimination using the Rasch model. This descriptive quantitative study involved 60 twelfth-grade science students from a high school in Bandung Regency. The instrument used was a 25-item multiple-choice conceptual understanding test on momentum and impulse, scored dichotomously. Student responses were analyzed using MINISTEPS software. The results showed item reliability of 0.82 (logit reliability) and 0.85 (item reliability). Cronbach's alpha coefficient also indicated good reliability at 0.85. Of the 25 items, 4 items did not meet the specified quality criteria. However, overall, most of the items on the conceptual understanding test of momentum and impulse met the quality criteria. These results indicate that the developed research instrument can be effectively used to measure students' conceptual understanding.

Keywords: Rasch Model · Conceptual Understanding · Momentum and Impulse

INTRODUCTION

In the ever-evolving world of education, the quality of assessment instruments is paramount in accurately measuring student learning outcomes (Ningsih et al., 2024). This is because a well-constructed assessment instrument should effectively cover all learning objectives taught by the teacher, encompassing both multiple-choice and essay-type questions (Muluki, 2020). As one of the most widely used assessment tools, tests necessitate rigorous quality evaluation to ensure their validity and reliability. Question item analysis is a critical step in this process. Through this analysis, we can identify and eliminate flawed questions, resulting in a more valid and reliable instrument for measuring student competence (Ningsih et al., 2024)

A test instrument is said to be of high quality if it has good validity and reliability. The higher the validity and reliability value of an instrument, the more accurate the data obtained from a study (Susdelina, 2018). This is in line with the opinion of Hayati & Lailatussadah (2016) who stated that validity and reliability are important indicators in determining the quality of a research instrument. Validity is a test conducted to measure the extent of the accuracy of a test, while reliability measures the extent to which the results of an instrument can be trusted (Syahrul, 2010; Zulpan & Rusli, 2020)

✉ Dena Tresna Ningsih
denatresna@upi.edu

Universitas Pendidikan Indonesia. Bandung, Indonesia.

How to Cite: Ningsih, D.T., Zahran, M., Suwarma, I.R., & Novia, H. (2024). Item Analysis of The Conceptual Understanding Test on Momentum and Impulse Using the Rasch Model. *Prosiding Seminar Nasional Fisika & International Physics Conference*, *3*(1), 60-70. https://proceedings.upi.edu/index.php/sinafi/

Question item analysis is a crucial step in the development of valid and reliable assessment instruments. One of the increasingly popular methods in question item analysis is the Rasch Model. The Rasch model, a modern measurement model, offers a more in-depth approach to analyzing question items (Safitri et al., 2024). The Rasch model is a modern assessment theory that can classify the calculation of items and persons in a distribution map (Rozeha et al., 2007).This model is part of the grain response theory (Thissen et al., 2001).

Ministeps, a software that implements the Rasch Model, has become a popular tool among educators and researchers. Ministep is a limited version of Winsteps. This program can analyze the dichotomous shape test (object) or polytome (description). Ministeps is a software used to analyze quantitative data. Its main function is to test the quality of question items in a test or other assessment instrument. By using a statistical model called the Rasch Model, Ministeps can provide more in-depth and accurate information about the characteristics of question items and the test-taker's abilities. By objectively measuring the latent ability of test takers, Ministeps provides more detailed information about the quality of question items, such as difficulty level and discriminating power (Rusilowati, 2018).

The Rasch model is able to analyze better than CTT in measuring reliability, difficulty level and discriminating power (Hardianti et al., 2021). The Rasch Model has advantages when compared to classical theory, namely: 1) It provides a linear scale with equal intervals. 2) It can predict missing data. 3) It can provide a more accurate estimate of student ability. 4) It can detect model misfit. 5) It provides replicable measurements. (Sumintono & Widhiarso, 2015; Taufiq et al., 2021).

Question item analysis using the Rasch Model and Ministeps software provides a number of benefits for the world of education. First, this analysis allows us to identify question items that are not functioning properly, so that improvements or eliminations can be made. Second, this analysis can also be used to measure the relative difficulty of each question item, so that a better test can be prepared. Third, by using Ministeps, we can obtain more accurate information about students' abilities, which can be used as a basis for making pedagogical decisions. Thus, the analysis of question items using Ministeps contributes to improving the quality of learning and assessment (Suseno & Susongko, 2021)

By using Ministeps software, educators and researchers can implement the Rasch Model in a practical way, allowing them to gain a more comprehensive understanding of the quality of the assessment instruments they use (Wibowo, A., & Cholifah, T. N., 2018).

Previous research by Rahman et al. (2021) has successfully developed an instrument to measure high-level thinking skills (HOTS) on certain materials. However, the study focuses more on high-level cognitive aspects. This study aims to complement the research by developing and analyzing instruments that measure students' conceptual understanding of momentum and impulse materials. Thus, this study pays special attention to the cognitive foundations underlying higher-order thinking skills. In addition, the use of the Rasch Model in question item analysis allows for a more in-depth evaluation of the quality of the instrument. Therefore, this study aims to develop momentum and impulse instruments that can accurately measure student understanding, so that they can be used to improve the learning process.

RESEARCH METHODS

This study uses a quantitative descriptive research method with a One-Shot design. The population in this study is all students of grade XI MIPA in one of the high schools in Bandung Regency. The research sample consisted of 60 students in grade XI MIPA who were selected using the purposive sampling technique. The test instrument was used to obtain data on the results of students' understanding of concepts in the form of 25 multiple-choice questions with a correct score of 1 and a false score of 0. So that the data obtained is dichotomous data. This test tool contains comprehension skills in accordance with the Taxonomy of Anderson and Krathwohl which is divided into seven dimensions of cognitive processes, namely interpreting, exemplifying. classify, summarize, conclude, compare, and explain. (Siregar & Nara, 2015).

Data analysis was carried out using Ministeps software. From the output of the Ministeps software, several question item parameters that correspond to the Rasch Model are obtained.

The criteria for good question items in this study are as follows:

- MNSQ Outfit is in the range of 0.5 to 1.5.
- ZSTD Outfit is in the range of -2.0 to 2.0.
- Correlation of items with total scores is in the range of 0.4 to 0.85.

In addition, the value of Cronbach's alpha coefficient indicates the overall reliability level of the instrument (Sumintono & Widhiarso, 2015).

RESULTS AND DISCUSSION

The results of the item analysis of the concept comprehension test conducted at one of the Bandung Regency high schools showed that the quality of the instrument could be seen from the value of validity, reliability, difficulty of the questions, and differentiation.

Item Validity

Validity is a measure of the extent to which an instrument measures what should be measured (Arikunto, 2010). Validity refers to the degree to which a measurement instrument accurately measures what it is intended to measure. An instrument is considered valid if it truly measures the construct it is designed to assess (Sugiyono, 2004) in (Arsi & Herianto, 2021). In other words, a valid test actually measures what it wants to measure. The validity test of the construct was analyzed using Rasch modeling (*unidimensionality*). Unidimensionality was analyzed using rasch analysis software in the form of *Ministep* version 5.6.1 which was seen from the *raw variance value explained by measure* in the output part of *table* 23*: Dimensionality* items which were then interpreted based on the following criteria.

Table 1. Unidimensionality Instrument Value Criteria

Nilai Raw variance explained by measure (%)	Criterion
$20 < Rve \leq 40$	Fulfilled
$40 < Rve \leq 60$	Appropriate
$Rve > 60$	Special

(Sumintono & Widhiarso, 2015)

The results of the unidimensionality of the instrument obtained are shown in the following figure 1.

```
TABLE 23.0 Data Hasil Uji Coba 25 Soal (Analisis ZOU125WS.TXT  Apr 17 2024 10:49
INPUT: 60 Person  25 Item  REPORTED: 60 Person  25 Item  2 CATS MINISTEP 5.6.3.0
--------------------------------------------------------------------------------

     Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = Item information units
                                      Eigenvalue   Observed   Expected
Total raw variance in observations     =      36.3300 100.0%         100.0%
  Raw variance explained by measures   =      11.3300  31.2%          29.5%
    Raw variance explained by persons  =       7.4427  20.5%          19.4%
    Raw Variance explained by items    =       3.8873  10.7%          10.1%
  Raw unexplained variance (total)     =      25.0000  68.8% 100.0%   70.5%
    Unexplned variance in 1st contrast =       3.6944  10.2%  14.8%
    Unexplned variance in 2nd contrast =       2.6140   7.2%  10.5%
    Unexplned variance in 3rd contrast =       2.4539   6.8%   9.8%
    Unexplned variance in 4th contrast =       1.9531   5.4%   7.8%
    Unexplned variance in 5th contrast =       1.6752   4.6%   6.7%
```

Figure 1. Output item dimensionality table

Based on Figure 1 the raw variance explained by measure obtained from the trial is 31.2%, then based on the criteria for the unidimensionality value of the instrument, the value meets the criteria of "met", which means that the instrument used can measure one variable without being affected by other variables. The value of unexplained variance in 1st contrast obtained from the trial was 10.2% which means that the quantity of unidimensionality instrument was fairly good because the unexplained variance in 1st contrast value obtained was less than 15%.

The validity test of each question item is obtained from the output of the ministep tables in table 10: item fit order. Fit items are seen from the value of the average square outfit (MNSQ), the value of Z-Standard outfits (ZSTD) and the value of point measure correlation (Pt Measure Corr). The results of each criterion are then interpreted based on the fit-statistical value criteria according to (Sumintono & Widhiarso, 2015), as shown in table 2 and table 3 below.

Table 2. Criteria for MNSQ, ZSTD, and Pt Measure Corr outfits

Indicator	Accepted values
Outfit MNSQ	$0{,}5 < MNSQ < 1{,}5$
Outfit ZSTD	$-2{,}0 < ZSTD < +2{,}0$
Pt Measure Corr	$0{,}4 < Pt\ Measure\ Corr < 0{,}85$

(Sumintono & Widhiarso, 2015)

Table 3. Interpretation of Question Item Quality

Criterion	Interpretation
The three indicators are met	Perfect fit
Two out of three indicators met	Appropriate
One in three indicators met	Less Suitable
All indicators are not met	Not Suitable

(Sumintono & Widhiarso, 2015)

The results of the validity test of the instrument obtained are shown in the following figure 2.

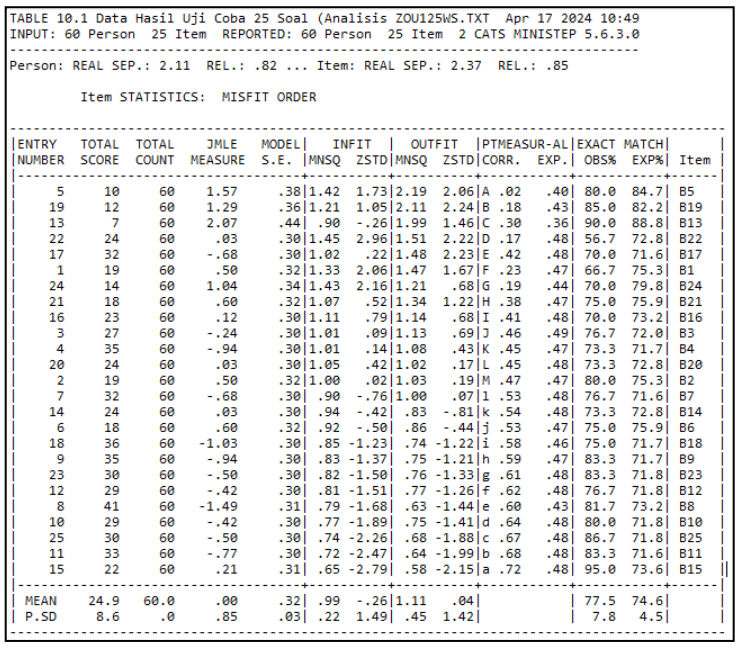

TABLE 10.1 Data Hasil Uji Coba 25 Soal (Analisis ZOU125WS.TXT Apr 17 2024 10:49
INPUT: 60 Person 25 Item REPORTED: 60 Person 25 Item 2 CATS MINISTEP 5.6.3.0

Person: REAL SEP.: 2.11 REL.: .82 ... Item: REAL SEP.: 2.37 REL.: .85

Item STATISTICS: MISFIT 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
5	10	60	1.57	.38	1.42	1.73	2.19	2.06	A .02	.40	80.0	84.7	B5
19	12	60	1.29	.36	1.21	1.05	2.11	2.24	B .18	.43	85.0	82.2	B19
13	7	60	2.07	.44	.90	-.26	1.99	1.46	C .30	.36	90.0	88.8	B13
22	24	60	.03	.30	1.45	2.96	1.51	2.22	D .17	.48	56.7	72.8	B22
17	32	60	-.68	.30	1.02	.22	1.48	2.23	E .42	.48	70.0	71.6	B17
1	19	60	.50	.32	1.33	2.06	1.47	1.67	F .23	.47	66.7	75.3	B1
24	14	60	1.04	.34	1.43	2.16	1.21	.68	G .19	.44	70.0	79.8	B24
21	18	60	.60	.32	1.07	.52	1.34	1.22	H .38	.47	75.0	75.9	B21
16	23	60	.12	.30	1.11	.79	1.14	.68	I .41	.48	70.0	73.2	B16
3	27	60	-.24	.30	1.01	.09	1.13	.69	J .46	.49	76.7	72.0	B3
4	35	60	-.94	.30	1.01	.14	1.08	.43	K .45	.47	73.3	71.7	B4
20	24	60	.03	.30	1.05	.42	1.02	.17	L .45	.48	73.3	72.8	B20
2	19	60	.50	.32	1.00	.02	1.03	.19	M .47	.47	80.0	75.3	B2
7	32	60	-.68	.30	.90	-.76	1.00	.07	l .53	.48	76.7	71.6	B7
14	24	60	.03	.30	.94	-.42	.83	-.81	k .54	.48	73.3	72.8	B14
6	18	60	.60	.32	.92	-.50	.86	-.44	j .53	.47	75.0	75.9	B6
18	36	60	-1.03	.30	.85	-1.23	.74	-1.22	i .58	.46	75.0	71.7	B18
9	35	60	-.94	.30	.83	-1.37	.75	-1.21	h .59	.47	83.3	71.7	B9
23	30	60	-.50	.30	.82	-1.50	.76	-1.33	g .61	.48	83.3	71.8	B23
12	29	60	-.42	.30	.81	-1.51	.77	-1.26	f .62	.48	76.7	71.8	B12
8	41	60	-1.49	.31	.79	-1.68	.63	-1.44	e .60	.43	81.7	73.2	B8
10	29	60	-.42	.30	.77	-1.89	.75	-1.41	d .64	.48	80.0	71.8	B10
25	30	60	-.50	.30	.74	-2.26	.68	-1.88	c .67	.48	86.7	71.8	B25
11	33	60	-.77	.30	.72	-2.47	.64	-1.99	b .68	.48	83.3	71.6	B11
15	22	60	.21	.31	.65	-2.79	.58	-2.15	a .72	.48	95.0	73.6	B15
MEAN	24.9	60.0	.00	.32	.99	-.26	1.11	.04			77.5	74.6	
P.SD	8.6	.0	.85	.03	.22	1.49	.45	1.42			7.8	4.5	

Figure 2. Output of the fit order item table

Based on figure 2, information was obtained regarding the values of *MNSQ*, *ZSTD*, and *Pt Measure Corr*. To determine the suitability of the question items, the interpretation of each question item is shown in the following table 4.

Table 4. Results of Question Item Quality Interpretation

No Question	MNSQ Score	ZSTD Score	Skor Pt Measure Corr	Value Criteria Met	Interpretation	Information
B1	1,47	1,67	0,23	2 Criteria	Appropriate	Used
B2	1,03	0,19	0,47	3 Criteria	Perfect Fit	Used
B3	1,13	0,69	0,46	3 Criteria	Perfect Fit	Used
B4	1,08	0,43	0,45	3 Criteria	Perfect Fit	Used
B5	2,19	2,06	0,02	0 Criteria	Not Suitable	Not Used
B6	0,86	-0,44	0,53	3 Criteria	Perfect Fit	Used
B7	1	0,07	0,53	3 Criteria	Perfect Fit	Used
B8	0,63	-1,44	0,6	3 Criteria	Perfect Fit	Used
B9	0,75	-1,22	0,58	3 Criteria	Perfect Fit	Used
B10	0,75	-1,41	0,64	3 Criteria	Perfect Fit	Used
B11	0,64	-1,99	0,68	3 Criteria	Perfect Fit	Used
B12	0,77	-1,26	0,62	3 Criteria	Perfect Fit	Used
B13	1,99	1,46	0,3	1 Criteria	Less Suitable	Not Used
B14	0,83	-0,81	0,54	3 Criteria	Perfect Fit	Used
B15	0,58	-2,15	0,72	2 Criteria	Appropriate	Used
B16	1,14	0,68	0,41	3 Criteria	Perfect Fit	Used
B17	1,48	2,23	0,42	2 Criteria	Appropriate	Used

No Question	MNSQ Score	ZSTD Score	Skor Pt Measure Corr	Value Criteria Met	Interpretation	Information
B18	0,74	-1,22	0,59	3 Criteria	Perfect Fit	Used
B19	2,11	2,24	0,18	0 Criteria	Not Suitable	Not Used
B20	1.02	0,17	0,45	3 Criteria	Perfect Fit	Used
B21	1,34	1,22	0,38	2 Criteria	Appropriate	Used
B22	1,51	2,22	0,17	0 Criteria	Not Suitable	Not Used
B23	0,76	-1,33	0,61	3 Criteria	Perfect Fit	Used
B24	1,21	0,68	0,19	2 Criteria	Appropriate	Used
B25	0,68	-1,88	0,67	3 Criteria	Perfect Fit	Used

Based on table 4, information was obtained regarding the interpretation of the suitability of the instrument test test questions given to 60 respondents and the results were obtained that there were 21 questions out of 25 questions that were tested that met the question functionality criteria, so that the 21 questions could be said to be good and could be used. While 1 question item only meets 1 criterion and the other 3 questions do not meet the criteria, so it is better not to use the question. Thus, in this study, the researcher only used 21 questions that met the item-fit criteria.

The analysis of the validity of the question items in Table 4 shows that more detailed results regarding valid and invalid question items for each aspect of concept understanding can be found in Table 5.

Table 5. Result valid and invalid question items for each aspect of concept understanding

No	Aspect of Concept Understanding	Valid	Invalid
1	Interpreting	B1, B2, B3	
2	Exemplifying	B4, B6, B7	B5
3	Classify	B8, B9, B10	
4	Summarize	B11, B12, B14	B13
5	Conclude	B15, B16, B17	
6	Compare	B18, B20, B21	B19
7	Explain	B23, B24, B25	B22

Based on Table 5, there are 3 valid questions each to measure the seven aspects of concept understanding. However, there is one question that is invalid in each aspect of exemplising, summarizing, comparing, and explaining. Thus, there are a total of 4 questions that are invalid and cannot be used in further analysis.

Item Reliability

Reliability is the determination of a test when it is applied to the same subject. (Arikunto, 2012). According to (Sugiyono, 2018) a reliable instrument is an instrument that, when used several times to measure the same object, will produce the same data. The Reliability Test is an index test that shows the extent to which a measuring device can be trusted or relied upon. This shows the extent to which the measurement results remain consistent when performed twice or more for the same symptom, using the same measuring instrument. A measuring instrument is said to be reliable if it produces the same result even though it is measured many times (Amanda et al., 2019). A reliable or reliable test is a test that produces a score in an orderly manner,

relatively unchanged even though it is administered in different situations. Reliability was analyzed using rasch model analysis software in the form of Ministep version 5.6.1 which was seen from the values of person reliability (p), item reliability (r), and Cronbach alpha (KR-20) in the output section of table 3.1: summary statistics.

Table 6. Interpretation of Person Reliability, Item Reliability, and Cronbach Alpha

Statistics	Index value	Interpretation
Item and person reliability	< 0.67	Low
	0.67 – 0.80	Enough
	0.81 – 0.90	Good
	0.91 – 0.94	Excellent
	> 0.94	Very good
Cronbach alpha (KR-20)	< 0.50	Low
	0.50 – 0.60	Keep
	0.61 – 0.70	Good
	0.71 – 0.80	Tall
	> 0.80	Very High

(Sumintono & Widhiarso, 2015)

The results of the validity test of the instrument obtained are shown in the following figure 3.

TABLE 3.1 Data Hasil Uji Coba 25 Soal (Analisis) ZOU125WS.TXT Apr 17 2024 10:4
INPUT: 60 Person 25 Item REPORTED: 60 Person 25 Item 2 CATS MINISTEP 5.6.3.

SUMMARY OF 60 MEASURED Person

	TOTAL			MODEL	INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	10.4	25.0	-.48	.50	.99	-.11	1.11	.06
SEM	.7	.0	.16	.01	.02	.13	.07	.16
P.SD	5.5	.0	1.23	.09	.19	1.00	.55	1.21
S.SD	5.6	.0	1.24	.10	.19	1.01	.55	1.22
MAX.	23.0	25.0	2.75	1.03	1.42	2.15	3.58	2.99
MIN.	1.0	25.0	-3.46	.43	.65	-2.10	.58	-2.06

REAL RMSE	.53	TRUE SD	1.11	SEPARATION	2.11	Person RELIABILITY .82
MODEL RMSE	.51	TRUE SD	1.12	SEPARATION	2.20	Person RELIABILITY .83
S.E. OF Person MEAN = .16						

Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .85 SEM = 2.12
STANDARDIZED (50 ITEM) RELIABILITY = .91

SUMMARY OF 25 MEASURED Item

	TOTAL			MODEL	INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	24.9	60.0	.00	.32	.99	-.26	1.11	.04
SEM	1.8	.0	.17	.01	.05	.30	.09	.29
P.SD	8.6	.0	.85	.03	.22	1.49	.45	1.42
S.SD	8.8	.0	.87	.03	.23	1.52	.46	1.45
MAX.	41.0	60.0	2.07	.44	1.45	2.96	2.19	2.24
MIN.	7.0	60.0	-1.49	.30	.65	-2.79	.58	-2.15

REAL RMSE	.33	TRUE SD	.79	SEPARATION	2.37	Item RELIABILITY .85
MODEL RMSE	.32	TRUE SD	.79	SEPARATION	2.50	Item RELIABILITY .86
S.E. OF Item MEAN = .17						

Item RAW SCORE-TO-MEASURE CORRELATION = -1.00
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000

Figure 3. Output table summary statistic

Based on Figure 3, information was obtained about the value of person reliability, item reliability, and Cronbach's alpha. These categories need to be interpreted in order to know the reliability of people and items. The interpretation of each question item is presented in table 6.

Table 7. Results of Interpretation of Question Item Reliability

	Average Logit (SD)	Separation	Reliability	Interp.	Alpha Cronbach	Interp.
Person	-0,48 (1,23)	2,11	0,82	Good	0,85	Very High
Item	0,00 (0,85)	2,37	0,85	Good		

Based on table 7, the *person reliability* value is 0.82 with a "good" interpretation. Meanwhile, the value of the *reliability* item obtained is 0.85 with a "good" interpretation. Then for the *value of Cronbach's alpha* (KR-20) obtained is 0.85 with a "very high" interpretation. Thus, based on the results of the analysis, it can be concluded that this instrument is reliable to be used as an instrument in this study.

Item Difficulty Level

The difficulty level is a number that expresses the difficulty level of a question item (Junika et al., 2020). According to (Bagiyono, 2017), quality questions have a good level of difficulty, namely having a balance in the comparison between easy, medium, and difficult questions. The difficulty level of a question is the opportunity to answer a form of a question at a certain level of ability which is usually expressed in the form of an index (Kadir, 2015). The greater the difficulty index (obtained from calculations), the easier the problem will be. The difficulty level of the questions is used to find out whether the question items used are in the easy, medium, or difficult categories. The determination of the difficulty level was carried out using *Rasch modeling analysis software* in the form of *Ministep* version 5.6.1. The difficulty level of the question items can be reviewed from the measure value (ME) and standard deviation (SD) by comparing the logit value of ME on each item and the SD score (Sumintono & Widhiarso, 2015). The level of difficulty of each question item can be interpreted based on the criteria in Table 8 below.

Table 8. Interpretation of Difficulty Level

Difficulty Level Criteria	Interpretation
ME < −1SD	Easy
-1SD ≤ ME ≤ +1SD	Medium
ME > +1SD	Difficult

The results of the difficulty level analysis from the results of the instrument test using Ministep software version 5.6.1 *on the output of the item measure table* obtained a standard deviation (SD) value of 0.32, then interpreted based on the difficulty level criteria in the following table 9.

Table 9. Results of Interpretation of Question Item Difficulty Level

Item Number Question	Measure (ME)	Standard Deviation (SD)	Criterion	Interpretation
B1	0,50	0,32	0,50 > 0,32	Difficult
B2	0,50	0,32	0,50 > 0,32	Difficult

Item Number Question	Measure (ME)	Standard Deviation (SD)	Criterion	Interpretation
B3	-0,24	0,30	-0,32 ≤ -0,24 ≤ 0,32	Medium
B4	-0,94	0,30	-0,94 < -0,32	Easy
B5	1,57	0,38	1,57 > 0,32	Difficult
B6	0,60	0,32	0,60 > 0,32	Difficult
B7	-0,68	0,30	-0,68 < -0,32	Easy
B8	-0,94	0,30	-0,94 < -0,32	Easy
B9	-1,49	0,31	-1,49 < -0,32	Easy
B10	-0,42	0,30	-0,42 < -0,32	Easy
B11	-0,77	0,30	-0,77 < -0,32	Easy
B12	-0,42	0,30	-0,42 < -0,32	Easy
B13	2,07	0,44	2,07 > 0,32	Difficult
B14	0,03	0,30	-0,32 ≤ 0,03 ≤ 0,32	Medium
B15	0,21	0,31	-0,32 ≤ 0,21 ≤ 0,32	Medium
B16	0,12	0,30	-0,32 ≤ 0,12 ≤ 0,32	Medium
B17	-0,68	0,30	-0,68 < -0,32	Easy
B18	-1,03	0,30	-1,03 < -0,32	Easy
B19	1,29	0,36	1,29 > 0,32	Difficult
B20	0,03	0,30	-0,32 ≤ 0,03 ≤ 0,32	Medium
B21	0,60	0,32	0,60 > 0,32	Difficult
B22	0,03	0,30	-0,32 ≤ 0,03 ≤ 0,32	Medium
B23	-0,50	0,30	-0,50 < -0,32	Easy
B24	1,04	0,34	1,04 > 0,32	Difficult
B25	-0,50	0,30	-0,50 < -0,32	Easy

Based on table 9, it can be seen that the question items are spread into easy, medium and difficult categories. The level of difficulty can be further analyzed by calculating the frequency and percentage for each interpretation of the difficulty level of the question item shown in the following table 10.

Table 10. Frequency and percentage of difficulty of question items

Interpretation	Frequency	Percentage (%)
Easy	11	44
Medium	6	24
Difficult	8	32

Based on table 10, information was obtained that the largest frequency was shown at the level of difficulty with an "easy" interpretation, which was as many as 11 questions with a percentage of 44%. There were 6 questions with a "medium" interpretation with a percentage of 24% and 8 questions with a "difficult" interpretation with a percentage of 32%. This shows that the difficulty level of the instrument is quite well distributed.

Complementing the previous research conducted by Rahman et al. (2021), this study explores students' understanding of the concept of momentum and impulse material as an important foundation before measuring higher-order thinking skills (HOTS). Thus, this study makes a significant contribution to the mapping of students' cognitive development in understanding abstract physics concepts.

The scope of this research is limited to physics learning, especially the concepts of Momentum and Impulse at the high school level. Therefore, it is necessary to conduct further

research with various materials and levels of education to identify students' ability to understand concepts more comprehensively.

Several recommendations can be put forward for further research development. The instruments that have been developed in this study can be tested on a wider scale, involving students from different schools and backgrounds. In addition, it is necessary to conduct research involving other physics materials and different levels of education.

The results of this study can be a valuable reference for educators in designing effective learning activities to improve students' ability to understand concepts, as well as making an important contribution to the development of evaluation instruments that can measure students' higher-level thinking skills more accurately, especially in understanding the concepts of momentum and impulse.

CONCLUSION

Based on the results of the study, it shows that the research instrument has excellent reliability. The logit reliability value of 0.82 and item reliability of 0.85 indicates that this instrument is consistent in measuring the understanding of the concept of momentum and impulse on the logit scale, and each individual question item has high consistency. Cronbach's alpha coefficient of 0.85 also confirms the instrument's overall internal reliability. Of the 25 questions tested, 4 questions were found that did not meet the quality criteria that had been set. Further analysis shows that questions B8 and B18 have low differentiation, while questions B5 and B13 have too high a level of difficulty. This indicates that the question items need to be revised or deleted in future research.

REFERENCES

Amanda, L., Yanuar, F., & Devianto, D. (2019). Uji validitas dan reliabilitas tingkat partisipasi politik masyarakat kota Padang. *Jurnal Matematika UNAND*, *8*(1), 179–188.

Arikunto, S. (2010). *Research Design Pendekatan Metode Kualitatif*. Al Fabet.

Arikunto, S. (2012). *Prosedur Penelitian: Suatu Pendekatan Praktik Edisi Revisi VI*. PT Rineka Cipta.

Arsi, A., & Herianto, H. (2021). *Langkah-langkah Uji Validitas Dan Reliabilitas Instrumen Dengan Menggunakan SPSS*.

Bagiyono. (2017). Analisis Tingkat Kesukaran dan Daya Pembeda Soal Ujian Pelatihan Radiografi Tingkat 1. *Widyanuklida*, *16*(1), 1–12.

Hardianti, H., Liliawati, W., & Tayubi, Y. R. (2021). Karakteristik tes kemampuan berpikir kritis siswa SMA pada materi momentum dan impuls: Perbandingan classical theory test (CTT) dan model Rasch. *Wahana Pendidikan Fisika*, *8*(1), 21–28.

Hayati, S., & Lailatussadah, L. (2016). Validitas dan reliabilitas instrumen pengetahuan pembelajaran aktif, kreatif, dan menyenangkan (PAKEM) menggunakan model Rasch. *Jurnal Ilmiah Didaktika*, *16*(2), 169–179.

Junika, N., Izzati, N., & Tambunan, L. (2020). Pengembangan soal statistika model PISA untuk melatih kemampuan literasi statistika siswa. *Mosharafa: Jurnal Pendidikan Matematika*, *9*(3), 499–510.

Kadir, A. (2015). Menyusun dan menganalisis tes hasil belajar. *Jurnal Kajian Ilmu Kependidikan*, *8*(2), 70–81.

Muluki, A. (2020). Analisis kualitas butir tes semester ganjil mata pelajaran IPA Kelas IV MI Radhiatul Adawiyah. *Jurnal Ilmiah Sekolah Dasar*, *4*(1), 86–96.

Ningsih, I., Srinanda, S., & Widyanti, E. (2024). Pemeriksaan dan Panskoran Tes. *Harmoni: Jurnal Ilmu Komunikasi Dan Sosial*, *2*(3), 211–219.

Rahman, A., Rusnayati, H., & Muslim, M. (2021). 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)*, *11*(2), 127–137.

Rozeha, A., Zaharim, A., & Masodi, M. S. (2007). Application of Rasch Measurement in Evaluation of Learning Outcomes: A Case Study in Electrical Engineering. *Regional Conference on Engineering Mathematics, Mechanics, Manufacturing & Architecture 2007 (EM3ARC)*.

Rusilowati, A. (2018). Asesmen literasi sains: Analisis Karakteristik instrumen dan kemampuan siswa menggunakan teori tes modern rasch model. *Prosiding Seminar Nasional Fisika Universitas Riau Ke-3*.

Safitri, I., Lestarani, D., Imtikhanah, R., Akbarini, N., Sari, M., Fitrah, M., & Hapsan, A. (2024). Teori Pengukuran Dan Evaluasi. *CV Ruang Tentor*.

Siregar, N., & Nara, H. (2015). *Belajar dan pembelajaran*. Penerbit Ghalia Indonesia.

Sugiyono, S. (2018). *Metode Penelitian Pendidikan Pendekatan Kualitatif, Kuantitatif dan R & D*. Alfabeta.

Sumintono, B., & Widhiarso, W. (2015). Aplikasi pemodelan rasch pada asesmen pendidikan. *Trim Komunikata*.

Suseno, E., & Susongko, P. (2021). Mengukur Validitas Tes. *Pemeral Edukreatif*.

Syahrul, S. (2010). Pengembangan Model Asesmen Kompetensi Siswa SMK dalam Konteks Pembelajaran Berbasis Kerja di Industri. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *14*(2), 246–268.

Taufiq, A., Yudha, E. S., Md, Y. H., & Suryana, D. (2021). Examining the Supervision Work Alliance Scale: A Rasch Model Approach. *The Open Psychology Journal*, *14*(1), 179–184. https://doi.org/10.2174/1874350102114010179

Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item response theory for items scored in more than two categories. In *Test scoring.* (pp. 141–186). Lawrence Erlbaum Associates Publishers.

Zulpan, Z., & Rusli, A. (2020). Validitas dan reliabilitas instrumen penilaian membaca short functional text pada siswa SMP KELAS VIII. *Jurnal Pendidikan Guru*, *1*(1). https://doi.org/10.47783/jurpendigu.v1i1.66