JISAE
Journal of Indonesian Student Assessment and Evaluation

https://doi.org/10.21009/JISAE

JISAE (Journal of Indonesian Student Assessment and Evaluation)
ISSN : P-ISSN: 2442-4919 | E-ISSN: 2597-8934
Website : http://journal.unj.ac.id/unj/index.php/jisae

Vol 8 No 1 (2022)

THE APPLICATION OF THE RASCH MODEL TO DEVELOP A TWO-TIER MULTIPLE-CHOICE TEST TO MEASURE HIGHER-ORDER THINKING SKILLS ON MOTION AND FORCE

Swastika Rhea Zivanka[1],
Universitas Tanjungpura, Pontianak, Indonesia

Haratua Tiur Maria Silitonga[2]
Universitas Tanjungpura, Pontianak, Indonesia

Hamdani[3]
Universitas Tanjungpura, Pontianak, Indonesia

Address for Correspondence:
[1]swastika.rhea9@student.untan.ac.id
[2]haratua.tiur.maria@fkip.untan.ac.id
[3]hamdani@fkip.untan.ac.id

ABSTRACT
In Indonesia, students' higher-order thinking skills are considered low, and students still struggle to solve HOTS problems. One of the reasons is that students are unfamiliar with HOTS questions. This research aims to produce a two-tier multiple-choice test to measure students' higher-order thinking skills on motion and force. The method used in this research is 4D by Thiagarajan. The test developed was validated by experts and was tested on 250 8th grade students in Pontianak. Analysis of the test used the Rasch Model to determine item fit, reliability, and difficulty levels using Winstep. Based on the qualitative and quantitative analysis, the test developed is feasible to measure students' higher-order thinking skills with the content validity value score 0,84 and construct validity explained by the test 64,71%. The item reliability is excellent, with a score of 0,94, but students have low consistency in answering the test. The test consisted of 2 complicated items, six difficult items, five easy items, and two specific items.

Keywords: HOTS; Two-Tier Multiple Choice Test; Rasch Model; Motion and Force

INTRODUCTION

In the 21st Century, human resources are required to have skills, those are including 1) critical thinking, 2) creativity, 3) collaboration, and 4) communications (Mukhtar & Haniin, 2019). Critical thinking is highly associated with higher-order thinking skills; hence students must be able to pick, interpret, and evaluate different kinds of information that are relevant, credible, and valid to solve problems creatively based on the information given that has to be considered (Afandi, Sajidan, Akhyar, & Suryani, 2018; Mislia, Indartono, & Mallisa, 2019; Miterianifa, Ashadi, Saputro, & Suciati, 2021). Critical and creative thinking skills can be improved through work experience in solving HOTS questions, which will impact problem-solving abilities habit (Widana, 2017). Indonesia is currently implementing the HOTS-oriented curriculum to develop these skills, the 2013 curriculum revision. According to Permendikbud RI Number 37 of 2018, in the knowledge core competence stated that students are expected to have to comprehend and apply knowledge based on their curiosity, and in the skill, core competence stated that students are expected to be able to process, present, and reason in concreate and abstract realm based on what they had learned. This shows that learning in the 2013 curriculum, students are expected to have conceptual comprehension and develop

higher order thinking skill so they can apply what they had learned at school to their daily lives.

The Bloomian Taxonomy revised by Anderson thinking skills or cognitive process is divided into six skills, remember; understand; apply; analyze; evaluate, and create (Krathwohl, 2002). Those six skills are then divided into two categories: Low Order Thinking Skills (LOTS) and Higher Order Thinking Skills (HOTS). Anderson and Krathwohl define HOTS as an analysis, evaluation, creation process. Higher-order thinking is a skill that requires a person to think logically and critically to understand a fact, concluding then linking it to facts in a new way to use it to solve a problem creatively (Thomas & Thorne, 2009). Hence, conceptual comprehension is not enough to solve HOTS problems; students also have to be able to connect between concepts logically and creatively.

According to PISA 2018, Indonesia placed 62 out of 79 countries participating in the science field with a score below the OECD average. The students can only solve problems with low complexity (OECD, 2019). Based on the PISA results, students' general ability is inadequate in integrating information, generalizing case by case to formulate a general solution, formulating real-world problems to the scientific concept, and investigating. The National Examination results also show the lack of students' higher-order thinking skills. According to the National Examination 2018 and 2019 results in science, students are still struggling to solve questions that require them to analyze and questions that have indirect information such as tables, pictures, and graphs (Puspendik, 2018, 2019). Hence students' higher-order thinking skill needs to be improved.

Teachers' evaluation process is one of the necessary factors in students' higher-order thinking skills development (Bhattacharya & Mohalik, 2021). Low physics learning achievement can be caused by the learning process or an inaccurate model assessment (Istiyono, Mardapi, & Suparno, 2013). Students' lack of practice and experience in solving problems that test students' higher-order thinking skills can also be the reason. Putri and Raharjo (2017) state that one way to improve students' HOTS is to provide and familiarize students with problem-solving, creative thinking, and critical thinking. Providing training or assessment in the form of HOTS questions can improve students' thinking abilities and learning motivation (Brookhart, 2010).

Regulation of the Minister of Education and Culture of the Republic of Indonesia No. 104 of 2014 concerning Assessment of Learning Outcomes by Educators in Primary and Secondary Education states that the Targets of Assessment of Learning Outcomes by Educators on knowledge competence include the level of ability to know, understand, apply, analyze, and evaluate factual knowledge, conceptual knowledge, procedural knowledge, and metacognitive knowledge. The use of HOTS questions to assess learning outcomes is expected to encourage students to think broadly about the subject matter to improve students' higher-order thinking skills. Nevertheless, in reality, in the field, the use of tests categorized as HOTS is lacking, even on a national scale.

Based on the results of pre-research in several public junior high schools in Pontianak, it was found that educators very rarely used HOTS questions in the assessment process, both informative and summative assessments. The test instruments used in the assessment primarily only measure C1 to C3 cognitive level, which is to measure students' comprehension and ability to apply. This shows that the test questions used are still in the LOTS category. Items that measure HOTS are only limited to measuring the ability of students to analyze, and there are no items that measure their skill in evaluating and creating. The form of the test used in schools is multiple choice four choices.

The lack of use of HOTS level questions was also found in the National Examination. In the results of the analysis of items used in the National Examination, the use of HOTS questions is still low. The majority of items used in the National Examination only measure

students' low-order thinking skills. There are only a few items that measure students' higher-order thinking skills, and the items are limited to analysis skill (C4) only (Afifah, 2020; Iffa, Fakhruddin, & Yennita, 2017; Ukhtia, 2020; Wijaya, Eresti, Despa, & Walid, 2019).

The HOTS test is usually given in the form of an essay or multiple choice. The two-tier multiple-choice test developed to better measure an ability is the two-tier multiple-choice test. The two-tier multiple-choice test developed by Treagust (2007) can be used to measure students' ability and find out students' misconceptions. This two-tier test was developed to reduce the shortcomings of the ordinary multiple-choice test model. Namely, in the ordinary multiple-choice test model, test takers are only asked to answer the given problem without considering the reasons why the test taker chose the answer. In the first tier, HOTS questions are presented, and in the second tier, the reasons are presented to determine students' understanding of the material and reduce the lucky guess factor. In multiple-choice one tier, students are only asked to work on questions without being asked why they chose the choices presented. Cullinane (2011) states that the inclusion of reasons at the second level of the two-tier multiple-choice question form can be used to improve higher-order thinking skills and see the ability of students to give reasons.

Rasch Model was developed to produce an objective measurement, where the measurement is sample dependent rather than test-dependent scoring (Novinda, Silitonga, & Hamdani, 2019; Safihin, 2019). The total of correct answers depends on the subject being measured, which is descriptive and applies to all subjects. On Rasch Model, there is a probabilistic model, which is that subjects who have higher abilities than the rest should have a greater chance of answering one item correctly and the other way around (Bond & Fox, 2020).

METHOD

The method used in this research is research and development. To develop the test, this research used the development procedure by Thiagarajan (1974), 4D (define, design, develop and disseminate). The test was written with higher-order thinking test development method by Widana (2017), which consist of six steps, (1) analyze Core Competency that can be developed to HOTS questions; (2) create HOTS items' blueprint; (3) Write down the items based on the analysis done; (4) determine scoring guidelines; (5) perform qualitative analysis; and (6) perform quantitative analysis.

The subjects of this study are 8th-grade students in Pontianak. This research was conducted in 3 different Junior High Schools in Pontianak, selected by the average National Examination score, SMP Negeri 10 Pontianak, SMP Negeri 11 Pontianak, and SMP Negeri 18 Pontianak. There are 15 items developed in this study to measure students' higher-order thinking skills on motion and forces. The test developed were validated by experts. The test was then given to 250 8th grade students from the selected schools. The data collected were analyzed using Rasch Model with Winstep. The analysis was done to observe the test characteristics: item fit, person reliability, item reliability, and the difficulty of the items.

RESULTS AND DISCUSSION

This research produced a two-tier, multiple-choice test instrument to measure students' higher-order thinking skills on motion and force. Higher-order thinking skills (HOTS) test instrument on this research is based on Bloomian Taxonomy revised by Anderson and Krathwohl, which consists of analysis(C4), evaluation (C5), and creation (C6). The test instrument developed has 15 items with three answer choices on each tier. The first tier of the item is the HOTS question and the second tier of the item is the arguments following the first tier. The answer collected then were quantitatively analyzed using Rasch Model with Winstep.

On the define stage, a preliminary study was done to gather information about students' higher-order thinking skills, their difficulty in solving HOTS problems, and the use of HOTS questions at school. Based on the evaluation documents, most of the items used to evaluate students only measure students' low order thinking skills or comprehension skills.

a) Test Development Results

The items developed were analyzed qualitatively based on the construction, materials, and the use of language (Widana, 2017). The test developed was validated by two experts in physics education and three science teachers on the development stage to obtain content validity. The validation results based on the Aiken V score are shown in Table 1.

Table 1 : Content Validity

Item Number	Aiken V' score	Category	Validity
1	0,88	Very High	Valid
2	0,86	Very High	Valid
3	0,85	Very High	Valid
4	0,86	Very High	Valid
5	0,81	Very High	Valid
6	0,89	Very High	Valid
7	0,84	Very High	Valid
8	0,83	Very High	Valid
9	0,85	Very High	Valid
10	0,87	Very High	Valid
11	0,83	Very High	Valid
12	0,82	Very High	Valid
13	0,83	Very High	Valid
14	0,82	Very High	Valid
15	0,84	Very High	Valid

Based on expert validation, the test instrument developed has an average coefficient of validity of Aiken V' of 0.84, so it is valid with perfect criteria. Items validated by experts are assessed in material, construction, and language. In terms of material, the test instrument developed was valid in the very high category with an average Aiken V' coefficient of 0.83. This shows that the material tested on the test instrument follows the essential competencies; the question indicators refer to higher-order thinking skills. The questions developed by the questions indicators can measure higher-order thinking skills according to their cognitive dimensions. In addition, the construction of the test instrument is also valid with a very high category with an average Aiken V' coefficient of 0.83. This shows that the formulation of the subject matter is excellent and clear, the stimulation used is clear and functioning, and the answer choices on both tiers are homogeneous and logical. Linguistically, the test instrument is also valid with a very high category with the average Aiken V' coefficient of 0.90. This shows that the language on the test instrument is communicative, following the rules of the Indonesian language, and does not cause double interpretation.

After the items developed were revised based on the experts' comments and suggestions, the items' readability was tested on a few 8th-grade students. The test was done to discover if the language used on each item can be understood and do not cause any misconception, also to test if the time given is enough to finish the whole items developed. The results show that the items developed are feasible to be tested on a bigger scale. The test instrument then was tested on 250 eighth-grade students from three state junior high schools in Pontianak. The test was scored by Afnia (2020) with four criteria as follows in Table 2.

Table 2: Scoring Guidelines

Criteria	Score
Both tiers incorrect	0
First-tier correct and second-tier incorrect	1
First-tier correct and second-tier incorrect	2
Both tiers correct	3

b) Testing Results

The test results were quantitatively analyzed using Rasch Model with Winstep to discover item fit, person reliability, item reliability, and items difficulty. The items developed were tested on 250 eighth-grade students from three state junior high schools in Pontianak.

Item Fit

Item fit shows whether the item developed function normally in measuring. The score of Outfit MNSQ (0,5<MNSQ<1,5); Outfit ZSTD(-2,0<ZSTD<+2,0); and Pt. Measure Correlation (0,4<Pt. Measure Corr<0,85) are the criteria of item fit that should be met (Sumintono & Widhiarsono, 2015). An item can be categorized as fit or valid if met at least two of three criteria. The result of the item fit analysis is shown in Table 3.

Table 3 : Item Fit

Item Number	Outfit MNSQ	Outfit ZSTD	Pt. Measure Corr	Validity
1	0,94	-0,63	0,43	Valid
2	1,14	2,23	0,43	Valid
3	1,22	3,29	0,4	Valid
4	0,55	-8,25	0,44	Valid
5	0,92	-1,25	0,51	Valid
6	0,95	-0,71	0,48	Valid
7	1,04	0,65	0,47	Valid
8	0,99	-0,13	0,24	Valid
9	0,97	-0,18	0,37	Valid
10	1,14	2,23	0,43	Valid
11	1,11	1,17	0,21	Valid
12	1,04	0,30	0,22	Valid
13	0,82	-3,29	0,41	Valid
14	1,09	0,97	0,23	Valid
15	1,12	1,35	0,07	Valid

Some items do not meet Outfit ZSTD, and Point-Measure Correlation criteria based on the item fit analysis. Outfit ZSTD shows if the item fits the Rasch Model. A negative value of the Outfit ZSTD shows that the data is overfitting to the Rasch Model or it has too slight variance in students' response than the Rasch Model and is closer to the Guttman-style response string where all of the students with high ability answer the item correctly and the students with low ability answer the item incorrectly (Susac, Planinic, Klemencic, & Milin Sipus, 2018). Whereas, a positive value of the Outfit ZSTD shows that the data is underfitting or it has too much variance in students' response than the Rasch

Model, which means that the students answer the item unpredictability (Bond & Fox, 2020; Lailiyah, Supriyati, & Komarudin, 2018). Point-Measure Correlation describes how an item correlates with the test as a whole. If the point-measure correlation value is 1,0, it shows a perfect correlation between the item responses and the estimated Rasch measure of the test takers where all students with low ability answer the item incorrectly and all students with high ability answer the item correctly. If the value is 0,00, it shows that the item does not correlate with the rest of the items; whether students answer correctly or incorrectly is random and do not have anything to do with their ability (Planinic, Boone, Susac, & Ivanjek, 2019; Smiley, 2015).

However, all of the items developed met at least two of the three criteria needed to be met. It can be concluded that all of the items developed are fit. This shows that all items are valid and can be used without needing any items to be removed or replaced.

Reliability

Rasch model informs about the person dan item reliability. The person and item reliability of this study is shown in Picture 1. Person reliability pictures respondents' consistency in answering the test, which means that respondents will reproduce the sequence of order if they are given another test measuring the same construct (Chan, Ismail, & Sumintono, 2014; Sumintono & Widhiarsono, 2015). Based on the analysis, the personal reliability of this research is categorized as low, with a score of 0,46. This shows that the consistency of the students' answers is weak. The value of the person separation is 0,93 or below 1,0, which means that the test developed could not distinguish students' ability well enough (Pratama & Husnayaini, 2020). Adding items to the test or testing students with a more extreme ability (high and low) can help to increase person reliability (Chan et al., 2014).

On Rasch Model, item reliability pictured the items' quality of the test instrument. Based on the analysis, the item reliability on this research is categorized as excellent, with a score of 0,98. The high value of item reliability shows that the test developed is sufficient and can measure students' higher-order thinking skills in motion and forces (Erfan, Maulyda, Ermiana, Hidayati, & Widodo, 2020).

Picture 1: Person and Item Reliability

```
TABLE 3.1 Analisis Data.xlsx                    ZOU832WS.TXT  Dec 16 2021 10:36
INPUT: 250 PERSON  15 ITEM  REPORTED: 250 PERSON  15 ITEM  4 CATS WINSTEPS 5.1.4.0
-------------------------------------------------------------------------------

     SUMMARY OF 250 MEASURED PERSON
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      21.4      14.9        -.06     .23      1.00    .02   1.00    .04 |
|  SEM        .4        .0         .02     .00       .02    .06    .02    .05 |
| P.SD       6.4        .6         .33     .04       .24    .93    .30    .86 |
| S.SD       6.4        .6         .33     .04       .24    .93    .30    .86 |
| MAX.      39.0      15.0        1.01     .82      2.07   3.69   2.74   4.06 |
| MIN.       1.0       7.0       -1.54     .22       .43  -2.59    .40  -1.98 ||
|-----------------------------------------------------------------------------|
| REAL RMSE    .24 TRUE SD     .23  SEPARATION   .93  PERSON RELIABILITY  .46 |
|MODEL RMSE    .23 TRUE SD     .24  SEPARATION  1.01  PERSON RELIABILITY  .51 |
| S.E. OF PERSON MEAN = .02                                                   |
-------------------------------------------------------------------------------
PERSON RAW SCORE-TO-MEASURE CORRELATION = .99 (approximate due to missing data)
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .48  SEM = 4.59 (approximate due to missing data)
STANDARDIZED (50 ITEM) RELIABILITY = .77

     SUMMARY OF 15 MEASURED ITEM
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     356.7     248.5         .00     .06      1.00   -.19   1.00   -.14 |
|  SEM       32.1        .3         .10     .00       .04    .83    .04    .71 |
| P.SD     119.9       1.0         .38     .01       .15   3.10    .16   2.65 |
| S.SD     124.1       1.1         .39     .01       .16   3.21    .16   2.74 |
| MAX.     600.0     250.0         .87     .08      1.21   3.61   1.22   3.29 |
| MIN.     107.0     246.0        -.78     .05       .55  -9.90    .55  -8.25 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .06 TRUE SD     .37  SEPARATION  6.44  ITEM   RELIABILITY  .98 |
|MODEL RMSE    .06 TRUE SD     .37  SEPARATION  6.60  ITEM   RELIABILITY  .98 |
| S.E. OF ITEM MEAN = .10                                                     |
-------------------------------------------------------------------------------
ITEM RAW SCORE-TO-MEASURE CORRELATION = -1.00 (approximate due to missing data)
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000
```

Wright-Map

The difficulty level states how difficult the items are by students' responses. The higher the level of difficulty, the lower the opportunity for students to answer the questions correctly. The items' difficulty level is classified by the standard deviation (SD) value and the items' logit value (Hamdu, Fuadi, Yulianto, & Akhirani, 2020). If the logit value is less than the minus of SD value (logit < -SD), the item is categorized as very easy; if the logit value is between the range of minus SD value to zero (-SD – 0), the item is categorized as easy; if the logit value is in the range of zero to SD value (0 – SD), the item is categorized as brutal; if the logit value is greater than the SD value (logit > SD), the item is categorized as very difficult (Palimbong, Mujasam, & Allo, 2019).

The level of difficulty in the developed test instrument has difficulty in the straightforward, easy, challenging, and very difficult categories, with two tough questions, six difficult questions, five easy questions, and two straightforward questions, which are shown in Table 4.

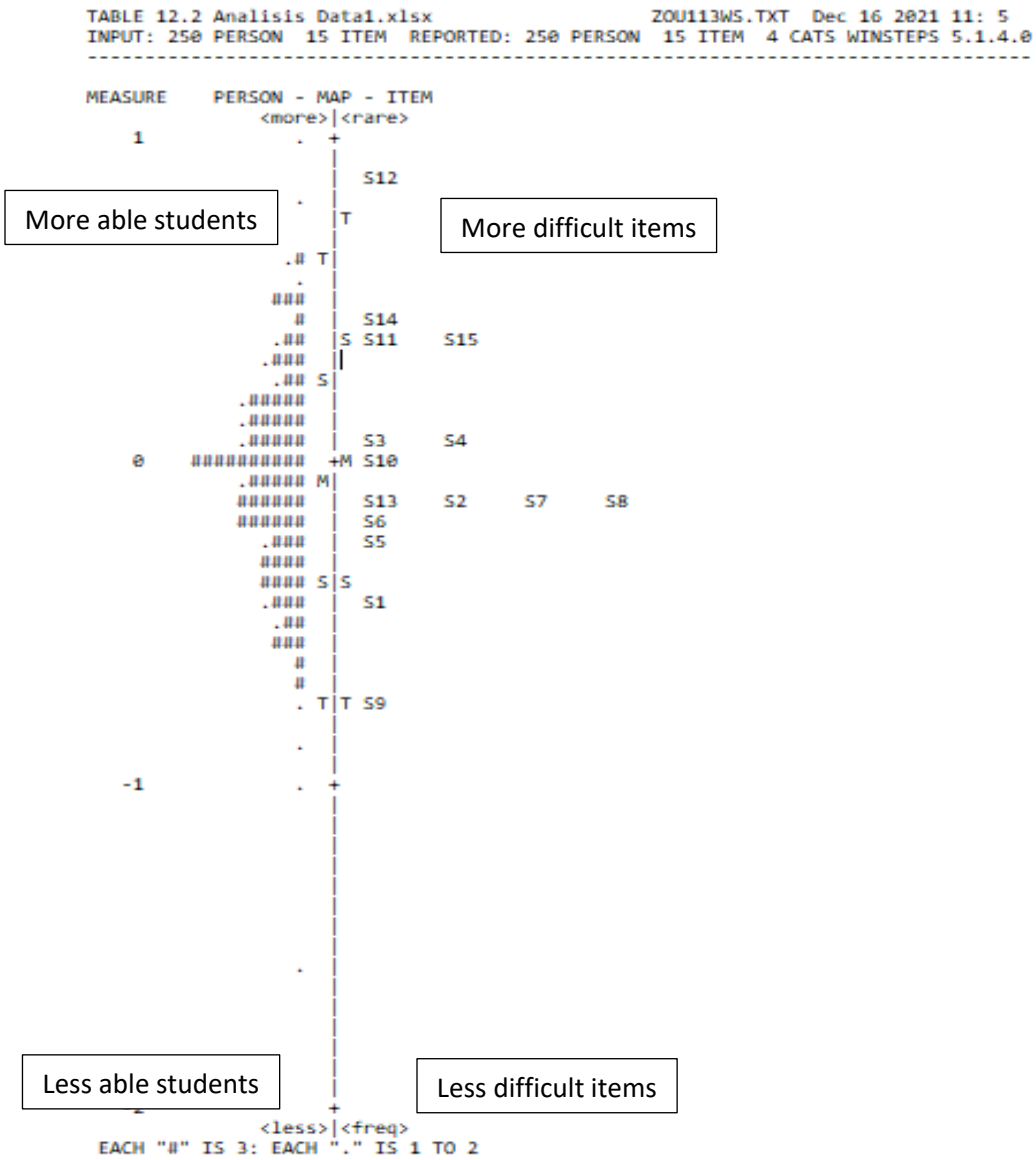
Table 4: Item Difficulty

Measure	Category	Item Number	Total	Percentage
>+0,38	Very Difficult	12, 14	2 items	13%
0-0,38	Difficult	11,15, 4, 3, 10, 7	6 items	40%
-0,38 – 0	Easy	13, 8, 2, 6, 5	5 items	33%
<-0,38	Very Easy	1, 9	2 items	13%

The width of the distribution of test items should match the population's ability so that the ability can be well measured (Planinic et al., 2019). Based on the results of the analysis

of the difficulty level of the items, all the items developed were within the range of the students' abilities. The distribution of students' abilities is shown in Picture 2. The wright-map compares students and the items developed, placing the difficulty of the items on the same measurement scale with the students' ability (Azura, Samsudin, & Utari, 2020). The wright-map shows that there are students that have ability greater than what the test measured, so there is no item that can differentiate their ability with the rest of the students tested, and there are also students that have ability lower than what the test measured, so there is no item that can discriminate them with the rest of the students. The most difficult question for students to work on is item number 12, this is indicated by the logit value of 0.89 which is the highest logit value compared to other items. In addition, the easiest question for students to work on is item number 9 with a logit value of -0.78 which is the lowest logit value of the other items.

Picture 2: Wright-map



Construct Validity

The construct validity of this research was conducted using SPSS with exploratory factor analysis. Construct validity shows how well the indicators describe the construct based on the measurement (Djamba & Neuman, 2014). The items ' dimensions or factors can be explained with the exploratory factor analysis method. The factor analysis using SPSS shows that the developed test instrument has a percentage of variance that the test of 64.71% can explain through 5 factors.

CONCLUSION

1. The test instrument developed is valid based on the validation results. The results of expert validation are in the very high category with a value of 0.84, and the results of construct validation of the developed test instrument have a percentage of variance that the test of 64.71% can explain.
2. The test instrument developed is valid and reliable, and the difficulty level was known according to the Rasch model. All items met at least two of the three criteria for item suitability; Person reliability in the weak category with a value of 0.46; item reliability in the particular category with a value of 0.98; and the test instrument consists of 2 tough questions, six difficult questions, five easy questions, and two straightforward questions.

REFERENCE

Afandi, A., Sajidan, S., Akhyar, M., & Suryani, N. (2018). Pre-Service Science Teachers' Perception About High Order Thinking Skills (HOTS) in the 21st Century. *International Journal of Pedagogy and Teacher Education*, *2*(1), 107. https://doi.org/10.20961/ijpte.v2i1.18254

Afifah, N. (2020). *ANALISIS SOAL UJIAN NASIONAL MATA PELAJARAN IPA SMP/MTs TAHUN AJARAN 2018/2019 BERBASIS HIGHER ORDER THINKING SKILL (HOTS)*.

Afnia, P. N., & Istiyono, E. (2020). Development of Two-tier Multiple Choice Instrument to Measure Higher Order Thinking Skills. *Advances in Social Science, Education and Humanities Research*, *397*(Icliqe 2019), 1038–1045. https://doi.org/10.2991/aisteel-19.2019.94

Azura, A., Samsudin, A., & Utari, S. (2020). ANALISIS PETA WRIGHT KETERAMPILAN BERPIKIR LEVEL LOTS DAN HOTS. *Wahana Pendidikan Fisika*, *5*(1), 76–83.

Bhattacharya, D., & Mohalik, R. (2021). Factors Influencing Students' Higher Order Thinking Skills Development. *Education India Journal : A Quarterly Refereed Journal of Dialogues on Education*, *10*(1), 349–361.

Bond, T. G., & Fox, C. M. (2020). Applying the Rasch model : Fundamental Measurement in the Human Sciences. In *Applying the Rasch Model* (Fourth Edi). New York: Routledge. https://doi.org/10.4324/9781410614575

Brookhart, S. M. (2010). How To Assess Higher-Order Thinking Skills In Your Classroom. In *Journal of Education* (Vol. 88). Alexandria. https://doi.org/10.1177/002205741808801819

Chan, S. W., Ismail, Z., & Sumintono, B. (2014). A Rasch Model Analysis on Secondary Students' Statistical Reasoning Ability in Descriptive Statistics. *Procedia - Social and Behavioral Sciences*, *129*, 133–139. https://doi.org/10.1016/j.sbspro.2014.03.658

Chandrasegaran, A. L., Treagust, D. F., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, *8*(3), 293–307.

https://doi.org/10.1039/B7RP90006F
Cullinane, A. (2011). Two-tier Multiple Choice Questions (MCQs) - How effective are they. *International Journal of Science & Technology Education*, *7*(1), 611–624.
Djamba, Y. K., & Neuman, W. L. (2014). Social Research Methods: Qualitative and Quantitative Approaches. In *Teaching Sociology* (Seventh Ed, Vol. 30). London: Pearson Education. https://doi.org/10.2307/3211488
Erfan, M., Maulyda, M. A., Ermiana, I., Hidayati, V. R., & Widodo, A. (2020). Validity and reliability of cognitive tests study and elementary curriculum development using Rasch model. *Psychology, Evaluation, and Technology in Educational Research*, *3*(1), 26–33. https://doi.org/10.33292/petier.v3i1.51
Hamdu, G., Fuadi, F. N., Yulianto, A., & Akhirani, Y. S. (2020). Items Quality Analysis Using Rasch Model To Measure Elementary School Students' Critical Thinking Skill On Stem Learning. *JPI (Jurnal Pendidikan Indonesia)*, *9*(1), 61. https://doi.org/10.23887/jpi-undiksha.v9i1.20884
Iffa, U., Fakhruddin, & Yennita. (2017). ANALISIS HIGHER ORDER THINKING SKILLS (HOTS) SISWA SMP N 1 SALO DALAM MENYELESAIKAN SOAL UJIAN NASIONAL IPA FISIKA TINGKAT SMP/MTs. *Jurnal Online Mahasiswa Fakultas Keguruan Dan Ilmu Pendidikan*, *4*(1), 1–9.
Istiyono, E., Mardapi, D., & Suparno, S. (2013). Pengembangan Tes Kemampuan Berpikir Tingkat Tinggi Fisika (PysTHOTS) Peserta Didik SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *17*(1), 108–126. Retrieved from https://journal.uny.ac.id/index.php/jpep/article/view/1364/1133
Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy : An Overview. *Theory in Practice*, *41*(4), 352. Retrieved from http://books.google.com/books?id=JPkXAQAAMAAJ&pgis=1
Lailiyah, L., Supriyati, Y., & Komarudin, K. (2018). Analysis of Measures Items in Development of Instruments Self-Assessment (Rasch Modeling Application). *Jisae: Journal of Indonesian Student Assesment and Evaluation*, *4*(1), 1–9. https://doi.org/10.21009/jisae.041.01
Mislia, T. S., Indartono, S., & Mallisa, V. (2019). Improving Critical Thinking among Junior High School Students through Assessment of Higher Level Thinking Skills. *Advances in Social Science, Education and Humanities Research*, *323*(ICoSSCE 2018), 326–333. https://doi.org/10.2991/icossce-icsmc-18.2019.58
Miterianifa, M., Ashadi, A., Saputro, S., & Suciati, S. (2021). Higher Order Thinking Skills in the 21st Century: Critical Thinking. *Proceedings of the First International Conference on Social Science, Humanities, Education and Society Development*. EAI. https://doi.org/10.4108/eai.30-11-2020.2303766
Mukhtar, M., & Haniin, K. (2019). *Modul Penyusunan Soal Keterampilan Berpikir Tingkat Tinggi (Higher Order Thinking Skills) Fisika* (S. Hadi & J. Abdi, eds.). Direktorat Pembinaan Sekolah Menengah Atas.
Novinda, M. R. R., Silitonga, H. T. M., & Hamdani. (2019). Pengembangan tes pilihan ganda menggunakan model Rasch materi gerak lurus kelas X Pontianak. *Jurnal Pendidikan Dan Pembelajaran*, *8*(6), 1–11. Retrieved from https://jurnal.untan.ac.id/index.php/jpdpb/article/view/33452
OECD. (2019). PISA 2018 Results. Combined Executive Summaries. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699. Retrieved from www.oecd.org/about/publishing/corrigenda.htm.
Palimbong, J., Mujasam, M., & Allo, A. Y. T. (2019). Item Analysis Using Rasch Model in Semester Final Exam Evaluation Study Subject in Physics Class X TKJ SMK Negeri 2 Manokwari. *Kasuari: Physics Education Journal (KPEJ)*, *1*(1), 43–51.

https://doi.org/10.37891/kpej.v1i1.40
Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research*, *15*(2), 20111. https://doi.org/10.1103/PhysRevPhysEducRes.15.020111
Pratama, D., & Husnayaini, I. (2020). Applying Rasch Model To Measure Students` Reading Comprehension. *JISAE: Journal of Indonesian Student Assessment and Evaluation*, *6*(2), 203–209. https://doi.org/10.21009/jisae.v6i2.14920
Puspendik. (2018). *Ringkasan Eksekutif Hasil Ujian Nasional 2018 SMP/MTs*. Jakarta.
Puspendik. (2019). *Ringkasan Eksekutif Hasil Ujian Nasional 2019 SMP/MTs*. Jakarta.
Putri, B. A. Y., & Raharjo, R. (2017). Emprical Validity Questions of High Order Thinking (HOT) Evaluation Instrument based on Computer Based Test (CBT) at Sensory System Sub Topic of Student Class XI Senior High School. *Berkala Ilmiah Pendidikan Biologi*, *6*(3), 353–359.
Safihin, M. (2019). Pengembangan Tes Menggunakan Model Rasch Materi Gaya Untuk SMA. *Jurnal Pendidikan Dan Pembelajaran*, *8*(6), 1–11. Retrieved from http://jurnal.untan.ac.id/index.php/jpdpb/article/view/33424/75676581548
Smiley, J. (2015). Classical test theory or Rasch- A personal account from a novice user. *Shiken*, *19*(1), 16–29.
Sumintono, B., & Widhiarsono, W. (2015). *Aplikasi Permodelan Rasch pada Assessment Pendidikan*. Cimahi: Trim Kominikata.
Susac, A., Planinic, M., Klemencic, D., & Milin Sipus, Z. (2018). Using the Rasch model to analyze the test of understanding of vectors. *Physical Review Physics Education Research*, *14*(2), 23101. https://doi.org/10.1103/PhysRevPhysEducRes.14.023101
Thomas, A., & Thorne, G. (2009). How to increase higher order thinking. *Center for Development and Learning*, 264. Retrieved from https://eric.ed.gov/?id=ED421544
Ukhtia. (2020). *Analisis Soal Ujian Nasional (UN) Tahun 2017/2018 dan 2018/2019 Mata Pelajaran IPA Terpadu SMP Berdasarkan Tahap Kognitif dan Tingkat Berpikir*. Banda Aceh.
Widana, I. W. (2017). Higher Order Thinking Skills Assessment (Hots). *Jisae: Journal of Indonesian Student Assesment and Evaluation*, *3*(1), 32–44. https://doi.org/10.21009/jisae.031.04
Wijaya, A., Eresti, A., Despa, D., & Walid, A. (2019). Analisis Butir Soal Persiapan Ujian Nasional Ipa Smp/Mts Tahun 2018 Sampai Dengan 2019 Berdasarkan Taksonomi Bloom. *LENSA (Lentera Sains): Jurnal Pendidikan IPA*, *9*(2), 57–63. https://doi.org/10.24929/lensa.v9i2.78