CO-CATALYST
Journal of Science Education Research and Theories

Vol. 1, No. 1, May 2023
e-ISSN: xxxx – xxxx
DOI: 10.33830/cocatalyst.vol1iss1.4891

Development of Conceptual Understanding Student Tests to The Basic Physics Subject: a Rasch Model Analysis

Riza Andriani[1*], Widya[1], Nurul Fadieny[1], Muttakin[2], Niki Dian Permana[3]

[1)] Physics Education FKIP Malikussaleh University, Lhokseumawe, Aceh, Indonesia
[2)] Chemistry Education FKIP Malikussaleh University, Lhokseumawe, Aceh, Indonesia
[3)] Natural Sciences Educations, Faculty of Education and Teacher Training, State Islamic University of Sultan Syarif Kasim, Pekanbaru, Riau, Indonesia

E-mail*: *rizaandriani@unimal.ac.id*

Article Info

Article History:
Received: March 5th, 2023
Revised: May, 26th , 2023
Accepted: May 30th, 2023

Keywords:
Conceptual understanding;
physics;
Rasch model;
test

ABSTRACT

The use of tests with unknown reliability and validity to measure student achievement was still widely practiced. This study examined the quality tests of conceptual understanding for college students in the basic concept of physics subject. This research was an R&D that used the 4D model. The test consisted of 2 types, namely multiple-choice and true-false tests, which were equipped with reasons. Development stages: 1) defined test grids; 2) designed tests; 3) developed tests; 4) expert validated of the content and construct; and 5) tested the validity, reliability, and level of difficulty. Analysis used the Rasch model assisted by Winstep software. The test results show that 18 multiple-choice and 17 true-false items fit the Rasch model in a range outfit MNSQ of 0.7-1.33 and ZSTD of -0,7-1,9. The reliability test is based on consistency in the range of 0.8-0.91 with good and excellent categories. The difficulty level shows three categories easy, difficult, and very difficult. And based on the results, 15 items are selected for each type of test. This selection is made by considering the similarity of competency outcomes measured by each test item. Both tests can be used for broader data collection to determine the best tests that represent students' conceptual understanding.

How to Cite:
Andriani, R., Widya, Fadieny, N., Muttakin, & Permana, N. D. (2023). Development of Conceptual Understanding Student Tests to The Basic Physics Subject; a Rasch Model. *Co-Catalyst: Journal of Science Education Research and Theories,* 1 (1): 43-54.

INTRODUCTION

Education is a person's conscious effort to change himself regarding abilities, attitudes, and knowledge. This self-change is obtained after a person receives an educational process in an educational institution, both formal and non-formal. It is necessary to carry out a test of knowledge, attitudes, or skills to measure the quality of its changes, with test indicators referring to learning outcomes in the aspects of knowledge, attitudes, or skills contained in the curriculum (Kemendikbudristek, 2022; Reotutar et al., 2020). Achievement of learning objectives in knowledge is carried out through cognitive tests or conceptual understanding tests that refer to the cognitive process dimensions mentioned by Bloom. Tests are implemented for students by working on several questions within a certain period, the results are in the form of a knowledge score, and this score will be used as a benchmark to determine the achievement of learning (Azizah et al., 2020; Novitasari et al., 2021; Sujarwanto, 2019). The results of this test provide information about the level of students' conceptual

understanding. This result can also be used to find out which concepts still need to be mastered by students so that improvements can be made in the following learning process to increase student achievement (Sindelar, 2011; Wiliam & Leahy, 2016).

Covid-19, which has hit the world since March 2020, has caused a significant change in the order of life in the world community, including the educational process. The educational process is inevitably forced to use distance learning systems assisted by social media (WhatsApp), video conferences (zoom meeting, google meet, Microsoft Teams, etc.), and e-learning (google classroom, Edmodo, etc.). This process affects academic achievement in the learning process and assessment, learning outcomes, and student perception (Hidalgo-Camacho et al., 2021). Online learning does not suit students' learning styles and harms students' mental destudent (Rohmani & Andriani, 2021). Students become lazier because during online learning, the teacher or lecturer tends to give many assignments, and lots of learning achievements are measured based on these assignments, which are difficult to control for their validity; the probability that the students themselves do not do the assignment is greater (Khan & Jawaid, 2020; Mega Susanti & Soleman Ritonga, 2021; Muhammadiyah et al., 2021; Yulianto & Majid Mujtahid, 2021). This invalidity is also magnified by the opportunity for the student to use search engines (google, youtube, etc.) in doing assignments. Based on this assumption, assignment is inappropriate to indicate student learning achievement because it cannot recognize and detect student conceptual understanding.

However, not all learning processes use assignments as a determinant of student achievement. Giving several questions in various forms such as objective tests, essays, case analysis, project creation, and others is also carried out both offline (in class) or online (google document, Edmodo, Quiziz, Moodle, etc.) (Dewa et al., 2020; Haruna et al., 2021; Hidayati & Aslam, 2021; Khusna et al., 2021; Nikat et al., 2022; Rozal et al., 2021; Somahhida & Makruf, 2022; Wiyoko & Hidayat, 2020). Distribution of questionnaires about the types of tests and Mechanism of developing course test questions at the University of Sultan Syarif Kasim Riau It is known that each lecturer has his type of test. There needs to be a standardized assessment instrument in the course. Course taught by different lecturers in different classes uses different types of tests (the test instruments level for one subject to another are not the same). The test was also only tested after being used to assess student achievement. The existence of injustice in the assessment process received by students and the doubtfulness of the validity of the test indicates that the test used cannot measure what it should measure.

Invalid test instruments will cause test results to be biased, and data cannot be used optimally to improve student learning outcomes, track student abilities in learning, track concepts that students need to understand adequately, or even student misconceptions. The test results cannot be used to conclude student learning completeness. The inability of the test to show relevant results will undoubtedly affect the competence owned by students in a course subject. The competence of this subject will significantly determine the qualifications of students in the field of study, even though it will be very much needed by students when they apply their knowledge in society after graduating from college. Therefore, to obtain truly relevant test results, correctly measure what should be measured, and indeed show the quality of student understanding, a test instrument is needed with the following criteria: valid, reliable,

efficient, objective, and suited to learning achievement indicators (Millard & Chavez, 2012).

The development of the type of test that will be used to measure student learning outcomes is adjusted to the material, needs, efficiency, and readiness of the teacher. Consideration of this criterion, then test type modification for physics concepts can be done. Several forms of tests that can be developed to test students' conceptual understanding are: 1) multiple-choice tests: the most common test is done out of consideration of simplicity and high objectivity (Dyahesita et al., 2019; Mutmainna et al., 2018); 2) Reasoned multiple choice: a test accompanied by reasons why the participant chose that answer (Cahyaningrum & Hidayat, 2018; Samaduri, 2022); 3) true-false test: a test that leads participants to be able to evaluate statements based on concepts they have understood (Couch et al., 2018; McAllister & Guidice, 2012; Michel et al., 2009; S. Khan, 2001) dan 4) essay test: an open test, the answer can be independent depending on the participant. Considering the ease of correction, true-false and multiple-choice tests are widely chosen by people worldwide compared to essay tests (Chandratilake et al., 2011).

In addition, considering that the test is not only applied offline but can also be applied online, the essay test is not suitable to be developed because it will be difficult to maintain objectivity in completing the test by the test taker. It is challenging to control test takers to work independently; refrain from using the help of friends or Google search engines during the test. This condition will cause a bias in the measured results. So, the development of true-false and multiple-choice tests is deemed necessary to support the implementation of offline and online tests and minimize the risk of cheating and invalidity of test results (Gudiño Paredes et al., 2021; Rowe, 2004). The development of these two types of tests will provide better results and descriptions of 1) the quality of learning received by students, 2) the quality of students' understanding of the physics lecture material they take, 3) physics concepts that lead to misconceptions, and 4) equality of assessment for each class even though taught by different lecturers. Moreover, the existence of these two types of tests does not have to be done offline; even online tests will provide valid and reliable results. So that in its implementation, it is more flexible according to needs.

METHODS

The research and development (R&D) method in this study applies a modified 4D model by stages: 1) definition of test indicators: refers to learning outcomes in the basic concept of physics course; 2) designing tests: designing types of tests based on criteria according to the level of dimensions processes cognitive in bloom taxonomy; 3) developing tests: developing test items based on indicators and learning outcomes; 4) expert validation in content and construct of the test: to see the suitability of the items with the question indicators; 5) test the validity, reliability, and level of difficulty; which was tested on 50 students of Department of PGMI (Elementary Teacher Education) at UIN Sultan Syarif Kasim Riau, 25 students from class 4A and 25 students from class 4B.

The instrument indicators and the number of items for each indicator can be seen in Table 1.

Table 1. Instrument Indicators and the Number of Items

No.	Subject	Indicator	Number of Items	
			Multiple Choice	True-False
1	Kinematics motion	Analyze the motion of objects: distance, displacement, time, speed, and acceleration	1	1
2	Dynamics Motion	Analyze the effect of force on the object's motion	2	2
3	Matter and properties	Analyze matter, its properties, and changes	2	2
4	Work and Energy	Analyze work and energy in everyday life	2	2
5	Temperature and Heat	Analyze phenomena related to heat and heat transfer in everyday life	2	2
6	Fluid Static	Analyze the static fluid characteristics and properties	2	2
7	Fluid Dynamic	Analyze fluid dynamic properties and application in everyday life	1	1
8	Vibration and Waves	Analyze the nature and behavior of vibration-waves	2	2
9	Sound	Analyze the nature of Sound and its application in life	2	2
10	Light and Optics	Analyze the concept of light and its behavior on mirrors and lenses	1	1
11	Electricity and Magnetism	Analyze electrical and magnetic phenomena	2	2
12	Earth and Solar System	Analyze the structure of the Earth, solar system, and celestial bodies	1	1

The research instruments are: 1) expert validation sheets: adjusting the indicators with the items designed; and 2) test instruments: 2 test packages, namely reasoned multiple-choice and true-false tests, each consisting of 20 questions. Analysis of validity, reliability, and difficulty level of test based on fit criteria with the Rasch model, a modern test that considers the results based on the response of item and person in a measurable distribution (Sumintono & Widhiarso, 2014). The Rasch model can also help detect the difficulty level of questions (Boone, 2016). Those with higher abilities will have a greater chance of answering the questions correctly; the opposite also applies. Rasch analysis was performed using Winstep software. Outfit MNSQ, Outfit ZSTD, and Pt Measure Corr values are used to determine the validity of the items: the suitability of the items with the Rasch model. The criteria can be seen in Table 2.

Table 2. Item Test Fit Criteria for Rasch Model

Outfit MNSQ	Outfit ZSTD	Pt Measure Corr
$\mathbf{0,5-1,5}$	$-2,0-2,0$	$0,4-0,85$

Item reliability and Cronbach's alpha are used to determine instrument reliability. The difficulty level of the questions is seen from the value of the measuring item. Criteria for difficulty level, reliability, and Cronbach's alpha can be seen in Table 3.

Table 3. Criteria for Difficulty Level, Person/Item Reliability, and Cronbach's Alpha

Difficulty Level		**Person/Item Reliability**		**Cronbach's Alpha**
$\boldsymbol{b > 1}$	*Very Difficult*	$0{,}71 - 0{,}80$	*Fair*	$> 0{,}8$ *Good*
$\boldsymbol{0{,}5 \leq b < 1}$	*Difficult*	$0{,}81 - 0{,}90$	*Good*	
$\boldsymbol{-0{,}5 \leq b < 0{,}5}$	*Moderate*	$0{,}91 - 0{,}94$	*Very Good*	
$\boldsymbol{-1 \leq b < -0{,}5}$	*Easy*	$> 0{,}94$	*Excellent*	
$\boldsymbol{b \leq -1}$	*Very Easy*			

RESULT AND DISCUSSION

The development of multiple choice and true-false tests begins with defining the concept to determine the test indicators. Every question is developed based on the learning curriculum, which is described as learning outcomes of the Basic Concept of Physics course (CPMK) in the Department of PGMI (Primary School Teacher Education), Faculty of Education and Teacher Training, Islamic State University of Sultan Syarif Kasim Riau. Then this CPMK outlined to be more specific becomes sub-CPMK which targets certain physics concepts. Ten main topics are outlined into ten sub-CPMK that measure C2-C5 levels in the revised Bloom taxonomy: understanding, applying, analyzing, and evaluating (Krathwohl, 2002). All of this revision Bloom's cognitive dimensions are verbs that show knowledge is created through a thought process and does not just appear out of thin air. A person's mind has a complex process so that he knows, understands, applies, analyzes, evaluates, and finds something. This level does not always show mastery of the dimensions above, so the dimensions below have been passed and mastered automatically.

The number of items developed for each type of test is 20 items. Both types of tests measure the same learning outcomes at equivalent cognitive levels. The difference between the two types of tests is only in the response given by the participant. Multiple choice tests with four answer options have a 25% chance of participants providing the correct answer. True-false tests with two possible responses have a 50% chance that the participant will answer correctly.

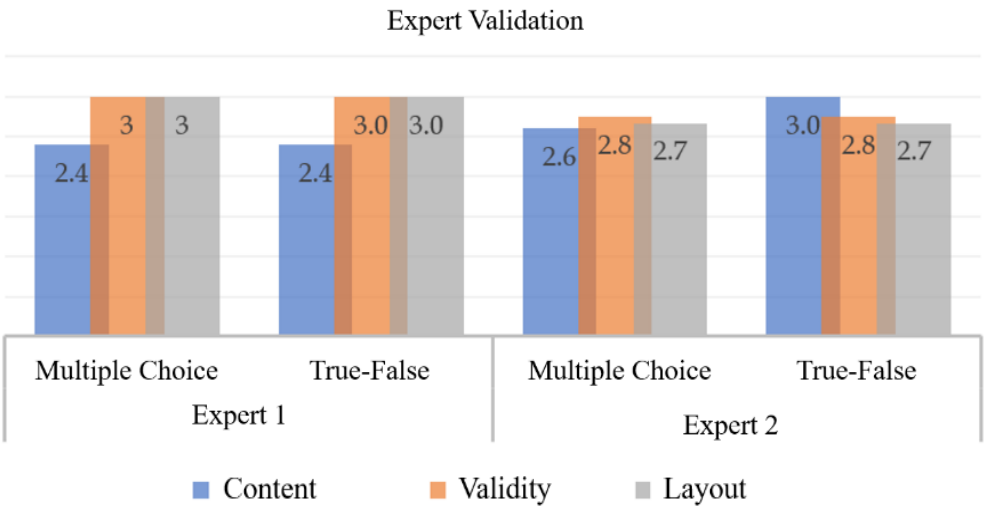

Figure 1. Expert Validation of the Content and Construct of the Test Instrument

After the instrument indicator formulation, experts carried out content validation by adjusting the questions with indicators, language suitability, and layout—content validation by two lecturers at the Department of Natural Sains Education UIN Sultan Syarif Kasim Riau. Analysis of expert validation results, the instruments developed have fulfilled the feasibility of the layout's contents, validity, and feasibility. The instrument was declared feasible to be tested to determine its validity, reliability, and difficulty level. The results of expert validation can be seen in Figure 1.

Before analyzing the suitability of the test results with the Rasch model, the instrument must fulfill the criteria of unidimensionality and local independence assumption. Unidimensionality shows that the test only measures one ability because modern tests use item response analysis, so it cannot be executed if it measures more than one ability. The results of the unidimensionality on both types of tests can be seen in Figure 2.

```
      Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                              -- Empirical --     Modeled
Total raw variance in observations     =        24.0 100.0%          100.0%
  Raw variance explained by measures   =         4.0  56.5%           57.2%
    Raw variance explained by persons  =          .4   1.8%            1.8%
    Raw Variance explained by items    =         3.5   4.8%            5.4%
  Raw unexplained variance (total)     =        20.0  43.5% 100.0%   82.8%
    Unexplned variance in 1st contrast =         3.4   4.4%   7.2%
    Unexplned variance in 2nd contrast =         3.2   3.4%   6.0%
    Unexplned variance in 3rd contrast =         2.5   4.4%   2.4%
    Unexplned variance in 4th contrast =         1.7   7.0%   8.4%
```

(a)

```
     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                              -- Empirical --    Modeled
Total raw variance in observations     =      24.3 100.0%         100.0%
  Raw variance explained by measures   =       4.3  67.6%          81.8%
    Raw variance explained by persons  =        .5   2.2%           2.3%
    Raw Variance explained by items    =       3.7  15.4%          16.5%
  Raw unexplained variance (total)     =      20.0  32.4% 100.0%   40.2%
    Unexplned variance in 1st contrast =       3.6  14.8%  17.9%
    Unexplned variance in 2nd contrast =       3.2  10.1%  15.9%
    Unexplned variance in 3rd contrast =       2.5   9.2%  12.4%
    Unexplned variance in 4th contrast =       1.9   7.8%   9.4%
    Unexplned variance in 5th contrast =       1.7   7.1%   8.6%
```

(b)

Figure 2. Unidimensionality Test; a) Multiple-Choice and b) True-False

A test is declared eligible for unidimensionality criteria when the raw variance value is greater than 20%, if more than 40% is good, and if above 60% is special. Based on this table, the values of raw variance for both types of tests are 56.5% and 67.6%, so both are eligible based on unidimensionality criteria. Unexplained variance indicates uniformity. The value should be at most 15%. Uniformity is in a good category if it is below 10% (Sumintono & Widhiarso, 2014). Both types of tests with unexplained variance below 10% indicate variations in test taker responses to each kind of test, both true-false and multiple-choice tests.

Local independent assumption explains that the response given by respondents to one item does not influence the response to other items. If this assumption does not meet, it can cause bias in the measurement result, as items are interrelated. This assumption is analyzed by examining the value on the diagonal of the variance-covariance matrix. If this value is close to 0 with the number 0,00… it means there is no local independence in the test instrument (Christensen & Bedrick, 1997; Klein Entink et al., 2009; Yu & Bien, 2017). The diagonal value of the variance-covariance matrix for both types of tests is almost close to 0, with the highest value being 0.011, so it can be said that the local independence criteria have been met, and there is no relation between the responses of one item and another.

The fit test of the instrument items to the Rasch model is based on the value of outfit MNSQ, Outfit ZSTD, dan Pt Measure Corr according to the acceptance criteria in Table 1. An item is considered fit if it meets at least two of these three criteria (Sumintono & Widhiarso, 2014). Test on True-False questions found 17 items that fit the Rasch model with a range of outfit MNSQ value 0,78-1,27 and outfit ZSTD value -0,7-1,6. Test on Multiple-Choice question found 18 items that fit the Rasch model with a range of outfit MNSQ value 0,77-1,33 and outfit ZSTD value -0,9-1,9. The result of the fit test of both instruments to the Rasch model can be seen in Table 4.

Table 1. Validity Test of True-False Questions and Multiple-Choice Questions

True-False					**Multiple Choice**				
No Butir	Outfit MNSQ	Outfit ZSTD	Pt Measure Corr	Decision	No Butir	Outfit MNSQ	OUTFIT ZSTD	Pt Measure Corr	Decision
1	1.27	1.6	-0.03	Not Fit	1	1.33	1.9	-0.06	Not Fit

7	1.12	1.0	0.15	Not Fit	8	1.21	1.6	0.10	Not Fit
13	1.09	0.4	0.31	Fit	7	1.08	0.5	0.23	Fit
2	1.03	0.3	0.35	Fit	13	1.09	0.4	0.27	Fit
8	1.05	0.4	0.28	Not Fit	16	1.07	0.3	0.26	Fit
11	1.03	0.3	0.31	Fit	14	1.03	0.5	0.32	Fit
5	1.02	0.2	0.4	Fit	3	1.03	0.3	0.32	Fit
12	1.01	0.3	0.39	Fit	9	1.01	0.3	0.42	Fit
10	0.99	0.0	0.45	Fit	12	1.00	0.1	0.39	Fit
19	0.97	0.0	0.43	Fit	10	0.86	0.1	0.36	Fit
17	0.82	-0.2	0.43	Fit	19	0.99	0.1	0.44	Fit
4	0.97	-0.1	0.53	Fit	17	0.78	0.1	0.44	Fit
16	0.95	-0.3	0.52	Fit	11	0.65	0.0	0.48	Fit
14	0.98	0.1	0.58	Fit	5	0.72	-0.2	0.43	Fit
3	0.95	-0.4	0.61	Fit	2	0.92	0.1	0.54	Fit
9	0.90	-0.2	0.56	Fit	4	0.95	-0.4	0.55	Fit
6	0.78	-0.3	0.67	Fit	18	0.87	-0.3	0.45	Fit
15	0.78	-0.1	0.77	Fit	6	0.87	-0.3	0.56	Fit
20	0.92	-0.5	0.72	Fit	15	0.76	-0.8	0.69	Fit
18	0.78	-0.7	0.7	Fit	20	0.77	-0.9	0.71	Fit

The reliability test for both instruments based on Cronbach's alpha shows a value of 0,8 for person reliability dan 0,91 for item reliability. This value means that the level of consistency of the test taker in completing the test is in the good category, and the consistency of items in measuring the test taker's ability is in the excellent category (Sumintono & Widhiarso, 2014).

The analysis of item difficulty level is used to determine whether the test question can distinguish the abilities of each test taker, differentiating those with higher and lower abilities. This analysis also identifies how likely it is for test takers with higher abilities to answer more straightforward questions correctly. The item difficulty level is determined using the Rasch Model by observing the value of item measure output and comparing them with the difficulty level criteria in Table 2 (Darmana et al., 2021; Purnama & Alfarisa, 2020).

The difficulty level for multiple choice questions falls within a range of the logit value from 2,28 to -1,44. The higher the logit value, the more difficult the question, and the fewer test takers can answer it correctly. The highest logit value for the multiple-choice test was 2,28 for item 15, with only one test taker answering correctly. The lower logit value was -1,44 for item number 20, with 20 test takers answering correctly. The difficulty level for true-false questions appeared to be higher, with a logit value of 2,46, and three items were considered extremely difficult: items 15, 12, and 6. Only one item, item number 20, was considered easy, with 14 test takers answering correctly. Items deemed too easy or difficult were discarded and not used further. The Logit value for the difficulty level of each item, both for the multiple-choice and true-false questions, can be seen in Figure 3.

NUMBER	SCORE	COUNT	MEASURE
15	1	25	2.28
12	2	25	1.26
6	3	25	.90
13	3	25	.90
16	3	25	.90
17	4	25	.88
18	5	25	.65
14	6	25	.52
9	6	25	.52
19	8	25	.22
10	9	25	.03
2	10	25	-.57
1	10	25	-.57
4	10	25	-.57
3	11	25	-.87
5	11	25	-.85
11	11	25	-.85
7	14	25	-1.27
8	14	25	-1.27
20	15	25	-1.44

(a)

NUMBER	SCORE	COUNT	MEASURE
15	1	25	2.46
12	1	25	2.46
6	2	25	1.54
13	3	25	.90
16	3	25	.90
17	3	25	.90
18	3	25	.90
14	5	25	.75
9	6	25	.52
19	6	25	.52
10	7	25	.37
2	8	25	.37
1	8	25	.37
4	9	25	-.27
3	9	25	-.27
5	10	25	-.35
11	10	25	-.35
7	11	25	-.56
8	13	25	-.86
20	14	25	-1.02

(b)

Figure 3. Level of Difficulty Instrument a) Multiple Choice and b) True-False

Analysis of validity, reliability, and difficulty level based on the fit to the Rasch model for both types of tests is the adjusted to learning outcome in the physics concept course in the Department of PGMI (Elementary Teacher Education) at UIN Sultan Syarif Kasim Riau. This consideration is made to obtain two types of tests that measure the same level of learning outcomes, or in other words, to find test instruments that measure the same learning outcome in different test formats. Therefore 15 multiple-choice items and 15 true-false items were selected that met the criteria for validity, reliability, and varying difficulty levels. These 30 items can be further used to 1) measure students' learning outcomes and conceptual understanding of physics; 2) identify the strengths and weaknesses of each test type; 3) identify how student performance differs when both types of test are applied to assess their understanding of physics concept; 4) find out which one of both type of test much difficult according to student based on their achievements; and 5) track the concepts that students have not yet fully mastered (Chandratilake et al., 2011; Couch et al., 2018; McAllister & Guidice, 2012; Cahyaningrum & Hidayat, 2018). Giving tests with various forms of procedures and feedback has a positive effect on student achievement (Phelps, 2012). Hence, the validity of the test instrument is a must so that the results obtained can be used as a reference for improving the quality of the following learning.

CONCLUSION

Based on the results of this research and discussion, the developed multiple-choice and true-false tests have met the criteria for validity and reliability. Validity tests on the multiple choice test found 18 items that fit the Rasch model with outfit MNSQ of 0,78-1,27 and outfit ZSTD of -0,7-1,6. Validity test on true-false test found 17 items that fit the Rasch model with outfit MNSQ of 0,77-1,33 and outfit ZSTD of -0,9-

1,9. Analysis of the difficulty level of the question showed varying results ranging from easy, moderate, difficult, and very difficult. Based on the test results, 15 items of each test type were retained. The selection of these 30 items was based on the similarity of learning outcomes measured by each test item. Further research can be conducted using this test to determine students' conceptual understanding of physics as measured by different tests. It will give more knowledge on the effectiveness of the assessment for measuring learning outcomes, which is one of both types of tests much more difficult according to students' responses and can be used as a reference for improving the quality of the following learning.

REFERENCES

Azizah, Z., Reyza, M., Taqwa, A., Ibnu, D., & Assalam, T. (2020). Analisis Pemahaman Konsep Fisika Peserta Didik menggunakan Instumen berbantuan Quizizz. *Jurnal Pendidikan Sains & Matematika*, *8*(2).

Boone, W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How? *CBE—Life Sciences Education*, *15*(4), rm4. https://doi.org/10.1187/cbe.16-04-0148

Cahyaningrum, R., & Hidayat, A. (2018). Analisis Pemahaman Konsep Fisika Mahasiswa pada Materi Induksi Elektromagnetik. *Jurnal Pendidikan*, *3*(10), 1383–1390. http://journal.um.ac.id/index.php/jptpp/

Chandratilake, M., Davis, M., & Ponnamperuma, G. (2011). Medical Education Assessment of medical knowledge: The pros and cons of using true/false multiple-choice questions. *The National Medical Journal of India*, *24*(4), 225–228.

Christensen, R., & Bedrick, E. J. (1997). Testing the Independence Assumption in Linear Models. *Journal of the American Statistical Association*, *92*(439), 1006. https://doi.org/10.2307/2965565

Couch, B. A., Hubbard, J. K., & Brassil, C. E. (2018). Multiple–True–False Questions Reveal the Limits of the Multiple–Choice Format for Detecting Students with Incomplete Understandings. *BioScience*, *68*(6), 455–463. https://doi.org/10.1093/biosci/biy037

Darmana, A., Sutiani, A., Nasution, H. A., Ismanisa*, I., & Nurhaswinda, N. (2021). Analysis of Rasch Model for the Validation of Chemistry National Exam Instruments. *Jurnal Pendidikan Sains Indonesia*, *9*(3), 329–345. https://doi.org/10.24815/jpsi.v9i3.19618

Dewa, E., Mikin, M. U. J., & Pandango, O. (2020). Pengaruh Pembelajaran DAring Berbantuan Laboratorium Virtual terhadap Minat dan Hasil Belajar Kognitif Fisika. *Jurnal Riset Teknologi Dan Inovasi Pendidikan (JARTIKA)*, *3*(2), 351–359.

Dyahesita, Q., Wahuni, A., & Suyudi, A. (2019). Analisis Pemahaman Konsep Fisika Siswa pada Pokok Bahasan Fluida Statis. *Jurnal Ilmu Fisika Dan Pembelajarannya)*, *3*(2), 68–75.

Gudiño Paredes, S., Jasso Peña, F. de J., & de La Fuente Alcazar, J. M. (2021). Remotely proctored exams: Integrity assurance in online education? *Distance Education*, *42*(2), 200–218. https://doi.org/10.1080/01587919.2021.1910495.

Haruna, N. A., Setiawan, D. G. E., & Odja, A. H. (2021). Penerapan E-Learning Menggunakan Media Edmodo dalam Pembelajaran Fisika Berbasis Nilai Karakter untuk Meningkatkan Hasil Belajar pada Konsep Usaha dan Energi. *Physics*

Education Research Journal, *3*(1), 65–74. https://doi.org/10.21580/perj.2021.3.1.6737

Hidayati, I. D., & Aslam, A. (2021). Efektivitas Media Pembelajaran Aplikasi Quizizz Secara Daring Terhadap Perkembangan Kognitif Siswa. *Jurnal Pedagogi Dan Pembelajaran*, *4*(2), 251. https://doi.org/10.23887/jp2.v4i2.37038

Kemendikbudristek. (2022). *Pembelajaran dan Asesmen*. Badan Standar, Kurikulum, dan Asesmen Pendidikan.

Khan, R. A., & Jawaid, M. (2020). Technology Enhanced Assessment (TEA) in COVID-19 Pandemic. *Pakistan Journal of Medical Sciences*, *36*(COVID19-S4). https://doi.org/10.12669/pjms.36.COVID19-S4.2795.

Khusna, A. A., Utami, R. E., & Nursyahidah, F. (2021). Kesalahan Siswa dalam Menyelesaikan Soal Sistem Persamaan Linear Dua Variabel Tipe HOTS Ditinjau dari Gaya Kognitif di Masa Pandemi Covid-19. *Jurnal Tadris Matematika*, *4*(1), 77–94. https://doi.org/10.21274/jtm.2021.4.1.77-94

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A Multivariate Multilevel Approach to the Modeling of Accuracy and Speed of Test Takers. *Psychometrika*, *74*(1), 21–48. https://doi.org/10.1007/s11336-008-9075-y

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2.

McAllister, D., & Guidice, R. M. (2012). This is only a test: a machine-graded improvement to the multiple-choice and true-false examination. *Teaching in Higher Education*, *17*(2), 193–207. https://doi.org/10.1080/13562517.2011.611868.

Mega Susanti, T., & Soleman Ritonga, P. (2021). Perbedaan Hasil Belajar Saat Terjadinya Pandemi Covid-19 Ditinjau dari Kemandirian Siswa pada Pelajaran Kimia. *Jurnal Pendidikan Dasar*, *8*(1), 1282–1290.

Michel, N., Cater, J. J., & Varela, O. (2009). Active versus passive teaching styles: An empirical study of student learning outcomes. *Human Resource Development Quarterly*, *20*(4), 397–418. https://doi.org/10.1002/hrdq.20025.

Muhammadiyah, M., Khurriyati, Y., Setiawan, F., & Binti Mirnawati, L. (2021). Dampak Pembelajaran Daring terhadap Hasil Belajar Siswa. *Jurnal Ilmiah Pendidikan Dasar*, *8*(1), 91–104.

Mutmainna, D., Mania, S., & Sriyanti, A. (2018). Pengembangan Instrumen Tes Diagnostik Pilihan Ganda Dua Tingkat untuk Mengidentifikasi Pemahaman Konsep Matematika. *MaPan: Jurnal Matematika Dan Pembelajaran*, *6*(1), 56–69. https://doi.org/10.24252/mapan.2018v6n1a6.

Nikat, R. F., Algiranto, A., Loupatty, M., & Henukh, A. (2022). Pemahaman Konsep Dinamika dan Kinematika Berdasarkan Conceptual Knowledge Melalui Aplikasi Game Quizizz. *Jurnal Pendidikan Sains Indonesia*, *10*(2), 218–230. https://doi.org/10.24815/jpsi.v10i2.23418.

Novitasari, D., Widyaningsih, S. W., & Sebayang, S. R. Br. (2021). Analisis Pemahaman Konsep Fisika Peserta Didik Kelas X IPA di SMA Negeri 1 Manokwari melalui Pembelajaran Online. *Silampari Jurnal Pendidikan Ilmu Fisika*, *3*(1), 39–57. https://doi.org/10.31540/sjpif.v3i1.1238.

Phelps, R. P. (2012). The Effect of Testing on Student Achievement, 1910–2010. *International Journal of Testing*, *12*(1), 21–43. https://doi.org/10.1080/15305058.2011.602920.

Primestike, I. N., & Salsabila, Q. (2021). *Efektivitas Pembelajaran Daring Di Masa Pandemi Covid-19*.

Purnama, D. N., & Alfarisa, F. (2020). Karakteristik Butir Soal Try Out Teori Kejuruan Akuntansi SMK berdasarkan Teori Tes Klasik dan Teori Respon Butir . *Jurnal Pendidikan Akuntansi Indonesia*, *18*(1), 36–46. https://doi.org/10.21831/jpai.v18i1.31457.

Reotutar, M. A. C., Tactay, N. T., & Ridwan, M. (2020). Achievement Test of Education Students in Assessment of Student Learning. *Budapest International Research and Critics in Linguistics and Education (BirLE) Journal*, *3*(4), 1742–1749. https://doi.org/10.33258/birle.v3i4.1330.

Rohmani, N., & Andriani, R. (2021). Correlation between academic self-efficacy and burnout originating from distance learning among nursing students in Indonesia during the coronavirus disease 2019 pandemic. *Journal of Educational Evaluation for Health Professions*, *18*, 9. https://doi.org/10.3352/jeehp.2021.18.9.

Rowe, N. C. (2004). Cheating in Online Student Assessment: Beyond Plagiarism. *Online Journal of Distance Learning Administration*, *7*(2), 1–10.

Rozal, E., Ananda, R., Zb, A., Fauziddin, M., & Sulman, F. (2021). The Effect of Project-Based Learning through YouTube Presentations on English Learning Outcomes in Physics. *AL-ISHLAH: Jurnal Pendidikan*, *13*(3), 1924–1933. https://doi.org/10.35445/alishlah.v13i3.1241.

S. Khan, D. A. D. J. K., K. (2001). Formative self-assessment using multiple true-false questions on the Internet: feedback according to confidence about correct knowledge. *Medical Teacher*, *23*(2), 158–163. https://doi.org/10.1080/01421590031075.

Samaduri, A. (2022). Analisis Pemahaman Konsep Siswa yang Diukur menggunakan Tes Pilihan Ganda Beralasan pada Mata Pelajaran Biologi. *Jurnal Pendidikan Glasser*, *6*(1), 109. https://doi.org/10.32529/glasser.v6i1.1466.

Sindelar, N. W. (2011). *Using Test Data for Student Achievement: Answers to "No Child Left Behind."* (Second). Rowman & Littlefield Education.

Somahhida, N. G., & Makruf, I. (2022). Multiple Choice Objective Test Arabic Subject: Analysis & Implementation of the Edmodo Application/ Tes Objektif Pilihan Ganda Mata Pelajaran Bahasa Arab: Analisis & Implementasi pada Aplikasi Edmodo. *Journal of Arabic Teaching, Linguistics, and Literature*, *3*(2), 162–176.

Sujarwanto, E. (2019). Pemahaman Konsep dan Kemampuan Penyelesaian Masalah dalam Pembelajaran Fisika. *Diffraction*, *1*(1), 22–33.

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi Model Rasch untuk Penelitian Ilmu-ilmu Sosial (Edisi Revisi)*. Trim Komunikata Publishing House.

Wiliam, D., & Leahy, S. (2016). *Embedding Formative Assessment*. Hawker Brownlow Education.

Wiyoko, T., & Hidayat, P. W. (2020). Analisis Miskonsepsi Mahasiswa PGSD dengan Metode Certainty of Response Index (CRI) melalui Fitur Quis Edmodo. *Jurnal Muara Pendidikan*, *5*(2), 680–688. https://doi.org/10.52060/mp.v5i2.375

Yu, G., & Bien, J. (2017). Learning Local Dependence in Ordered Data. *Journal of Machine Learning Research*, *18*, 1–60. http://jmlr.org/papers/v18/16-198.html.

Yulianto, D., & Majid Mujtahid, N. (2021). Online Assessment during Covid-19 Pandemic: EFL Teachers' Perspectives and Their Practices. *JET (Journal of English Teaching)*, *7*(2), 229–242. https://doi.org/10.33541/jet.v7i2.2770.