Advances in Social Science, Education and Humanities Research, volume 541
Proceedings of the 6th International Seminar on Science Education (ISSE 2020)

Implementation of Item Response Theory at Final Exam Test in Physics Learning: Rasch Model Study

Muh Asriadi AM [1,*] Samsul Hadi[2]

[1] *Master of Education Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta, Indonesia*
[2] *Department of Education Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta, Indonesia*
[*]*Corresponding author. Email asriadi190197@gmail.com*

ABSTRACT
A good test is one that has been validated and the accuracy of the gauge is tested. This article aims to provide a study of the implementation of item response theory with the Rasch model to analyze the quality of the final exam test items in physics. This is a survey conducted once and at a time. The subjects were the 131 grade XI students of MAN 1 Bone. Test of final exam and documentation of physics learning were used as data collection techniques. The data analysis method used was the item response analysis with the Rasch model. The results showed that the reliability value with Cronbach alpha (KR-20) which measures the interaction between the person and the item as a whole, namely $\alpha = 0.47$. The value of person reliability is 0.48 and the value of the item reliability is 0.88. This shows that the consistency of the answers from the test takers is still weak, but the quality of the items in the reliability aspect of the instrument is quite good. So that the questions final exam test in physics subject still needs improvement and consideration to be used again in the future.

Keywords: *Item response theory, Rasch model, Physic learning*

1. INTRODUCTION

The assessment of learning outcomes is used by educators on an ongoing basis to assess the achievement of students' competencies, to prepare learning outcomes reports, and to improve the learning process. The assessment will be useful for viewing the educational quality overall and the assessment will also provide important information for improving the learning process [1]. The test is one of the many forms of assessment of learning outcomes. The tests have been designed which are useful for some different purposes. they can be used for selection, evaluation such as when students are scored in class or the effectiveness of a teaching program is evaluated [2]. One form of assessment of learning outcomes is the end of the semester test, which in this article will discuss the final exam test on physics.

Tests are systemic procedures, containing samples of behavior and measuring behavior. The test is one way to estimate the level of human ability or skill indirectly, namely through a person's response to several stimuli or questions [3]. In this connection, the problem to be highlighted and studied is from the aspect of using the test which is designed in such a way as to raise questions, to what extent the test is following the ability of students who answer it. So that the review is directed at assessing the application of modern tests, namely the item response theory in assessing the learning outcomes of students with all the attributes and requirements they have.

Item response theory considers the test taker's behavior at the item level, not at the test level. Modeling at the item level creates a great deal of flexibility for application in test development, comparative study of item functions, adaptive computer testing, and score reporting [4]. Item response theory (IRT) rests on two basic postulates: (a) examiner performance on test items can be predicted (or explained) by a set of factors called traits, latent traits, or abilities; and (b) the relationship between test item performance and the set of traits underlying the item performance can be explained by a monotonically increasing function called the item character function or item characteristic curve (ICC). This function sets it as the trait level increases, the probability of correct response to an item increases [5]. In essence, IRT aims to overcome the weaknesses found in classical measurements. In a test situation, the

examinee's performance can be predicted (or explained) by determining the examinee's characteristics, known as traits, or the ability to estimate the examinee's score about these traits (called an "ability score") and by using the scores to predict or explain performance items and tests. When traits cannot be measured directly, they are referred to as latent traits or abilities. The item response model determines the relationship between the examiner's observable test performance and the unobservable traits or abilities assumed to underlie performance on the test [6]. This means that the same item for different test takers must be subject to the rules, or the same test-taker for different test items must also comply with the formula.

The two basic assumptions with the IRT model are unidimensional and local independence. The assumption of unidimensionality refers to the ability measured in a set of questions to be single. This means that a set of questions and/or test size (s) has only one latent trait (θ). Unidimensionality requires that our analytical procedures must include indicators of the extent to which people and goods conform to our concept of the ideal one-dimensional line [7]. However, in reality, this unidimensional assumption cannot be strictly adhered to because there are several influencing factors such as cognitive, personality, and factors related to testing administration. The most important thing in the unidimensional assumption is that there is one dominant component that affects the subject's performance. The second assumption is Local Independence which refers to the assumption that the responses are assumed to be independent at the level of individuals who have the same value, but the assumption does not generalize to the case of variation. For groups of individuals with a variety of traits assessed, responses to different test items are usually correlated because they are all related to the level of the individual trait [8]. The subject's response to an item does not affect the subject's response to other items. This means that the responses given to separate items in a test are mutually independent of the abilities given [9]. In simple terms, it can be concluded that local independence is fulfilled if the response of the subject/test taker to the item does not depend on the response of other items.

One form of item response theory modeling is the Rasch Model or One-Parameter Logistic Model. The One Parameter Logistic Model or called the Rasch Model (1 PLM) was brought by a mathematician in Denmark, George Rasch, who came up with a different approach to IRT in 1950. He used a logistic function to derive the ICC instead of the normal ogive function (although at the time he expressed his model differently), and his model contributed to simplifying the normal ogive model and computational complexity [10]. In the Rasch model, the probability of a randomly selected test taker at the ability level (θ) to get the correct answer on item (i) can be stated as follows.

$$P_i(\theta) = \frac{1}{1+e^{-D(\theta - b_i)}} \quad (1)$$

Information:
e: The exponential constant whose value is about 2.718
D: A scale factor whose value is 1.7 [4]

The results of the Rasch model show that the level of student success in working on the questions depends on the level of ability and difficulty of the questions. So that the probability of one success is the ability of the respondent to be reduced by the difficulty level of the item. One of the advantages of applying the Rasch model to analyze assessment data is its ability to consider and analyze assessment items and students as test-takers together and present the results together [11]. The Rasch model is a dichotomous assessment model that has only two categories, namely correct answers with a score of 1 and wrong answers with a score of 0. Rasch modeling uses both person scores per data and score per item data. These two scores form the basis for estimating the correct score indicating the level of individual ability as well as the level of difficulty of the test. The advantage of the Rasch Model compared to other models, in particular, the classical model of test theory, is the ability to predict missing data, based on systematic response patterns [12]. Thus the Rasch model can provide an objective measure of student ability that does not depend on the difficulty of the item in the assessment task.

Based on the description above, the purpose of writing this article is to provide a study of the implementation of item response theory with the Rasch model to analyze the quality of the final exam test items in physics.

2. RESEARCH METHODS

The type of research is a survey using a *Cross-Sectional Survey*. This type collects information from a sample drawn from a predetermined population. Besides, information is collected only at one time, although the time required to collect all the data can take from one day to several weeks or more [13]. Samples were taken from the results of the final semester exam for physics class XI IPA MAN 1 Bone-in 2020.

The data used in this study were documents in the form of final exam test scores totaling 131 people in the form of dichotomous data. For the data obtained to meet the characteristics of the Rasch Model, the data were analyzed using Winstep software. In this study, data analysis was in the form of summary statistics, information function, person item map, item difficulty level, and item fit level.

3. RESULT AND DISCUSSION

3.1. Summary Statistic of Final Exam Test

Winstep provides a complete analysis of the final test of the semester. To get information about the summary statistics of the test kits presented. To find out the value, it can be seen in the table 1 below.

Table 1. Summary statistic

Statistical Component	Value
Item Questions	30
Test Participants	131
Log-Likelihood Chi-Square	4001.73 with 3770 d.f. p = .0044
Average Measurement	-1.21
Cronbach alpha (KR-20) Person Raw Score "Test" reliability	0.47
Person Reliability	0.48
Item Reliability	0.88
Separation	2,65

Based on table 1, information is obtained that the amount of data provided by 131 test participants with 30 item questions is 3930 data. The resulting Chi-Square value is 4001.73 with 3770 degrees of freedom and p = .0044 where p <0.01. This shows that in general the measurements taken are very good and the results are significant. The results of this analysis contain two outputs, namely output for test-takers (person) and output for items. The output person explains in general whether or not the respondent is used. Likewise, with the output items, explaining whether in general the item items used as part of the test can be said to be fit or not. Referring to Figure 1 above, the average measure value obtained in the output person is -1.21 (μ <0.00). A negative score indicates that the student is having problems answering the question [14]. The mean value that is smaller than 0 indicates that the tendency of the test taker's ability is smaller than the difficulty level of the questions. The ideal value for INFIT and OUTFIT MNSQ is close to 1, while for INFIT and OUTFIT ZSTD, the ideal value is close to 0. For the person and item table, the mean values of INFIT and OUTFIT MNSQ and INFIT and OUTFIT ZSTD are close to ideal. The reliability value with Cronbach alpha (KR-20) which measures the interaction between the person and the item as a whole is α = 0.47. The value of person reliability is 0.48 and the value of the item reliability is 0.88 [15]. This shows that the consistency of the answers from the test takers is still weak, but the quality of the items in the reliability aspect of the instrument is quite good. There is also a separation value which indicates the quality of the test and the quality of the test takers. The greater the separation value the better because it can identify the broader subject group (capable or unable) and item group (difficult or easy). The formula that can be used to see the grouping more accurately is called the strata separation with the formula H = [(4 X separation) + 1] / 3[16]. Winstep output results like Figure 2, it is known that our item separation value is 2.65 then H = [(4 X 2.65) + 1] / 3, namely 3.8 or rounded to 4. This indicates that the respondent consists of only four groups. More separation indicates that the test is good. The higher the grain separation value, the better the measurement will be. In general, the final exam test item items have met the good criteria as evidenced by the reliability value that is close to 1. But what needs to be paid attention to is the poor or inconsistent patterns of student response responses. This is usually due to being unprepared for the test.

3.2. Information Function

Each measurement provides information about the measurement results. in the item response theory, the information function is one of the factors that affect the quality of an instrument. The information function will show you what the measurement was made for. The desired measurement information is not based on the individual being measured, but information on the measurement focus, especially on the relationship between the test and the individual [17]. The information function shows the reliability of the measurements made. The Rasch model emphasizes the separation coefficient (item separation). The higher the peak information that can be achieved, the higher the reliability value of the measurements made[18]. To see the measurement information function can be seen in the image below.

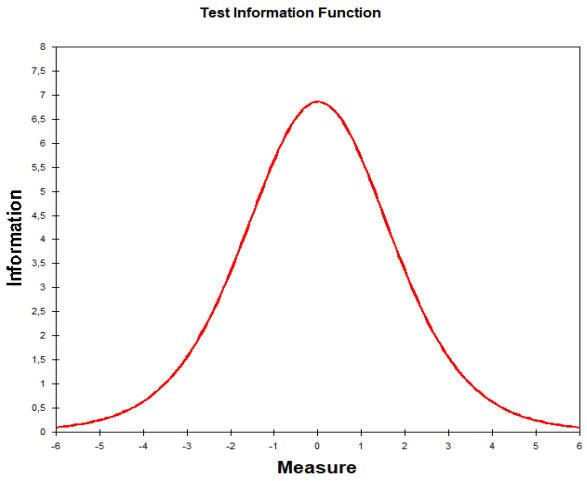

Figure 1 Graph of information function

Based on the picture above we can conclude that of the 30 questions presented to 131 test participants, it shows that the question items are suitable for determining the level of ability of moderate students. From this, it is obtained an illustration that for students with low or high abilities it will be difficult to describe their abilities when using this test question. This finding is an important note for teachers to improve the quality

of the physics test questions given by students to be able to describe the overall abilities of students with low, medium, and high abilities

3.3. Person Map Item

Rasch's analysis provides a person-item distribution map known as the Wright Map which is nothing but a person-item map. The Wright map describes the distribution of test-takers' abilities and the distribution of the difficulty levels of items with the same scale. provides an overview of the respondent's readiness by placing the difficulty level of the task on the same measurement scale as the respondent's ability [19]. The person-item map output results can be seen in the image below.

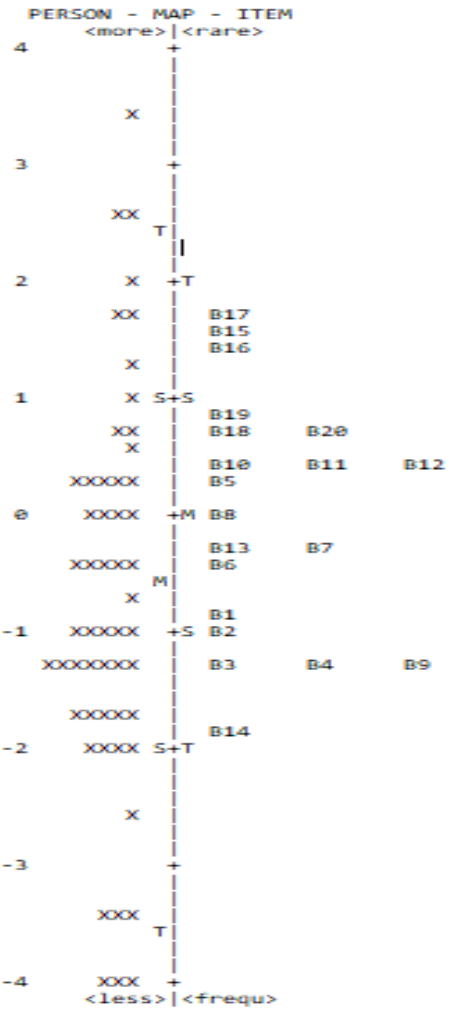

Figure 2 Person map item

In Figure 3 above, it shows that on the left side is the distribution of the subject's ability, while on the right side is the distribution of items. From this map, it can be seen that in general the questions in the test are equivalent when compared to the subject's ability. This means that all 30 items can be worked on by the test taker. There are no items that are the most difficult and the easiest. All questions are in the medium difficulty category. Besides, all questions are at intervals of -2 to 2, which means that the questions given to students do not have a significant difficulty level. So that the final exam test questions still need assessment and consideration to be used again in the future. The important thing to note from this analysis is that it is necessary to improve the quality of the questions by organizing each question that has been done by students and the results of the quality analysis have been known. In this case, it is suggested that the teacher make a question bank so that in the future when giving questions to students, they can generally measure the students' abilities with good quality questions.

3.4.Item Difficulty Level

One thing we need to pay attention to is the output of Rasch's analysis with this winstep. A high logit (measure) value indicates that the item has a high difficulty level. This correlates with the total score, where the small number of correct answers in the total score correlates with the higher measure value. The data measure of this item also has the same scale. The results of the analysis of the difficulty level of the questions can be seen in the table 1 below.

Table 2. Level of difficulty item

Item	Level of Difficulty	Description
1	-0.13	Easy
2	-0.37	Easy
3	-0.17	Easy
4	0.18	Difficult
5	0.50	Difficult
6	-0.29	Easy
7	0.13	Difficult
8	-1.40	Very Easy
9	-0.37	Easy
10	0.00	Easy
11	0.56	Difficult
12	-0.97	Easy
13	0.50	Difficult
14	1.15	Very Difficult
15	-0.52	Easy
16	0.33	Difficult
17	0.18	Difficult
18	0.33	Difficult
19	0.04	Difficult
20	1.24	Very Difficult
21	-0.21	Easy
22	-0.21	Easy
23	0.04	Difficult
24	1.06	Very Difficult
25	0.00	Easy
26	0.50	Difficult
27	-1.23	Very Easy
28	-0.76	Easy
29	-0.76	Easy
30	0.68	Difficult

Based on table 2, the difficulty level of the questions can be seen in the measure column where the results of the analysis have been sorted by Winstep based on the

level of difficulty. The item with the highest difficulty level is at the top, while the easiest item is at the bottom. So it is known that the most difficult item is the 20th item and the easiest is the 8th item. The category table for the difficulty level of the questions to understand the percentage of the difficulty of the questions. To see the percentage of questions can be seen in Table 2 below.

Table 3. Category level of difficulty item

Category	Frequency	Percentage (%)
Very Easy	2	6.67
Easy	13	43.33
Difficult	12	40
Very Difficult	3	10

Based on table 2 above, it is known that most of the item questions that are used as the final semester test are in the easy category with a percentage of 43.33%. So it needs improvement on the test, especially in terms of the difficulty level of the questions. An important note based on the results of the analysis is to improve the quality of the questions starting from the theoretical construct that underlies the making of the questions. The indication that can be seen from the results of this analysis is that the construction of the material in each item has not fully represented the competencies that must be achieved by each student. Therefore, it is suggested to the teacher to review the blueprint test and review its construction and adjust it to the competencies that must be achieved by students.

3.5. Item Fit Level

The suitability level of this item is used to see the accuracy of the item with the model or fit item. To find out the difficulty level of an item, it can be seen in the table 4 below.

Table 4. Item statistics: misfit order

Item	Infit		Outfit		Pt-Measure	
	Mnsq	Zstd	Mnsq	Zstd	Corr.	Exp.
24	1.01	0.1	1.42	1.4	A 0.13	0.23
18	1.04	0.3	1.34	1.8	B 0.12	0.24
30	1.02	0.3	1.15	0.7	C 0.18	0.23
1	1.11	1.2	1.15	1.1	D 0.07	0.25
3	1.01	0.1	1.15	1.2	E 0.21	0.25
21	1.09	1.1	1.05	0.5	F 0.13	0.25
26	1.02	0.2	1.09	0.5	G 0.19	0.24
14	1.08	0.4	1.01	0.1	H 0.16	0.23
11	1.01	0.1	1.06	0.4	I 0.21	0.24
23	1.02	0.3	1.06	0.5	J	0.24
					0.20	
8	1.06	1.2	1.04	0.5	K 0.18	0.26
25	1.05	0.5	1.06	0.1	L 0.18	0.24
20	.97	0.0	1.01	0.2	M 0.25	0.23
28	1.04	0.7	1.04	0.5	N 0.20	0.26
9	1.01	0.2	1.04	-0.2	O 0.24	0.25
19	.98	-0.1	.98	0.1	o 0.27	0.24
6	1.01	0.1	1.01	0.0	n 0.24	0.25
27	.99	-0.3	.99	-0.3	M 0.28	0.26
2	.98	-0.2	.98	-0.6	l 0.29	0.25
22	.98	-0.1	.93	-0.3	k 0.28	0.25
16	.98	-0.1	.96	-0.3	j 0.27	0.24
13	.98	-0.1	.94	-0.3	i 0.27	0.24
5	.98	-0.1	.92	-0.4	h 0.29	0.24
4	0.97	-0.2	.90	-0.6	g 0.30	0.24
7	0.97	-0.2	.90	-0.5	f 0.30	0.24
17	0.95	-0.4	.91	-0.1	e 0.31	0.24
29	0.96	-0.8	.97	-0.7	d 0.32	0.26
10	0.92	-0.8	.94	-1.2	c 0.39	0.24
15	0.91	-1.4	.86	-1.5	b 0.40	0.25
12	0.91	-2.1	.88	-1.6	a 0.40	0.26
Mean	1.00	0.0	1.02	0.0		
S.D	0.05	0.73	0.13	0.8		

Based on table 4 above, provides information on the value of outfit means-square, outfit z-standard, and point measure correlation are the criteria used to see the level of suitability of items. If the item does not meet the criteria for repair or replacement of the item. The guide for assessing item suitability criteria is the accepted Outfit Mean Square (MNSQ) value: 0.5 < MNSQ <1.5. Accepted Z-standard outfit value (ZSTD): -2.0 < ZSTD <+2.0 Accepted Point Measure Correlation value: 0.4 <pt measure corr < 0.85. In addition, the output has a Point Measure Correlation value which is classified as very good (> 0.40), good (0.30 - 0.39), sufficient (0.20 - 0.29), unable to discriminate (0.00 - 0.19), and requires examination of items (<0.00) [20]. Winstep has already sorted the items based on which items are not fit. Items that are not fit are usually placed at the top.

Besides, in the Mean Square (MNSQ) output and the Z-standard outfit (ZSTD), all items have met the fit criteria. The items displayed have a Point Measure Correlation value, 7 items are in a good category, 14 items are in the sufficient category, and 9 items are in

the incapable of discriminating category. So it is suggested that 9 item questions, namely items 1, 8, 14, 18, 21, 24, 25, 26, 30 must get improvement in terms of material substance. Suggestions for improvement on items that this item must be improved, namely the teacher to improve the material structure both from the characteristics of the material and the language used. The questions must also be able to avoid cultural bias and gender differences so that DIF does not occur in the physics questions made. Also, the use of language should be understood by every student who raises the questions so that there are no misconceptions when working on the questions.

4. CONCLUSIONS

The conclusion that can be drawn in this study is that the item items in the final exam test in physics are not very suitable for use by respondents. This can be seen in the alpha Cronbach reliability value which is still low even though in terms of quality it is quite good. Besides, several item questions must be corrected in terms of material construction so that the final semester test questions on physics can measure students' abilities correctly and fairly. The benefits that can be obtained from this study are obtained by analyzing the quality of a test instrument with the Rasch model which can be a guide for teachers to produce test instruments that have good quality in measuring student competence.

AUTHORS' CONTRIBUTIONS

The first author acts as the main writer and researcher, while the second author acts as a guide in improving the content of the article as a whole.

REFERENCES

[1] O. Ofianto, S. Suhartono, An Assessment Model of Historical Thinking Skills through The RASCH Model, Journal Research and Evaluation in Education, 1(1) (2015) 73-83 DOI: https://doi.org/10.21831/reid.v1i1.4899

[2] M.J. Allen, W.M. Yen, Introduction to Measurement Theory, Wadsworth.Inc, 1979.

[3] S.E. Ng, K.J. Yeo, A.B. Mohd Kosnin, Item Analysis for the Adapted Motivation Scale Using Rasch Model, International Journal of Evaluation and Research in Education (IJERE), 7(4) (2018), 264-269. DOI: https://doi.org/10.11591/ijere.v7i4.15376

[4] W.J. Van Der Linden, R.K. Hambleton, Handbook of Modern Item Response Theory, Springer New York, 1997.

[5] R.K. Hambleton, H. Swaminathan, H.J. Rogers, Fundamentals of Item Response Theory, Sage Publications, 1991.

[6] R.K. Hambleton, H. Swaminathan, Item Response Theory, Springer Netherlands, 1985.

[7] T.G. Bond, C.M. Fox, Applying the Rasch Model: Fundamental Measurement in The Human Sciences, Third Edition, Routledge Taylor and Francis Group, 2015.

[8] M.D. Reckase, Multidimensional Item Response Theory, Springer New York, 2009.

[9] I. Camminatiello, M. Gallo, T. Menini, The Rasch Model for Evaluating Italian Student Performance, Journal of Applied Quantative Methods, 5(2) (2018) 331–349.

[10] H. Habibi, J. Jumadi, M. Mundilarto, The Rasch-Rating Scale Model to Identify Learning Difficulties of Physics Students based on Self-Regulation Skills, International Journal of Evaluation and Research in Education, 8(4) (2019) 659. DOI: https://doi.org/10.11591/ijere.v8i4.20292

[11] M. Chan, R. Subramaniam, Validation of a Science Concept Inventory by Rasch Analysis, in: Editor: M.S. Khine, Rasch Measurement Application in Quantitative Educational Research, Springer, Singapore, 2020.

[12] I. Isnani, W.B. Utami, P. Susongko, H.T. Lestiani, Estimation of College Students' Ability on Real Analysis Course Using Rasch Model, Jounal of Research and Evaluation in Education, 5(2) (2019) 95–102. DOI: https://doi.org/10.21831/reid.v5i2.20924

[13] J.R. Fraenkel, N.E. Wallen, How to Design and Evaluate Research in Education, 7th ed. McGraw-Hill, 2009.

[14] H. Abdullah, N. Arsad, F.H. Hashim, N.A. Aziz, N. Amin, S.H. Ali, Evaluation of students' achievement in the final exam questions for microelectronic (kkkl3054) using the rasch model, in: Procedia Social and Behavioral Sciences, vol. 60, Elsevier, Amsterdam, 2012, pp. 119–123. DOI: https://doi.org/10.1016/j.sbspro.2012.09.356

[15] H. Othman, I. Asshaari, H. Bahaludin, Z.M. Nopiah, N.A. Ismail, Application of rasch measurement model in reliability and quality evaluation of examination paper for engineering mathematics courses, in: Procedia Social and Behavioral Sciences, vol. 60, Elsevier, Amsterdam, 2012, pp. 163–171. DOI: https://doi.org/10.1016/j.sbspro.2012.09.363

[16] S. Nuryanti, M. Masykuri, E. Susilowati, Analisis Iteman dan Model Rasch pada Pengembangan Instrumen Kemampuan Berpikir Kritis Peserta Didik Sekolah Menengah Kejuruan, Jurnal Inovasi Pendidikan IPA, 4(2) (2018) 224–233. DOI: https://doi.org/10.21831/jipi.v4i2.21442

[17] P.E. Larasati, Supahar, D.R.A. Yunanta, Validity and reliability estimation of assessment ability instrument for data literacy on high school physics material, in: Proceedings of The 5th International Seminar on Science Education, vol. 1440, IOP Publishing, Bristol, 2020, pp. 1-8. DOI: https://doi.org/10.1088/1742596/1440/1/012020

[18] U.D. Purnamasari, B. Kartowagiran, Application Rasch Model Using R Program in Analyze The Characteristics of Chemical Items, Jurnal Inovasi Pendidikan IPA, 5(2) (2019) 147–157. DOI: https://doi.org/0.21831/jipi.v5i2.24235

[19] A. Salman, A.A. Aziz, Evaluating user readiness towards digital society: A rasch measurement model analysis, in: Procedia Computer Science, vol. 65, Elsevier, Amsterdam, 2015, pp. 1154–1159. DOI: https://doi.org/10.1016/j.procs.2015.09.028

[20] W.J. Boone, J.R. Staver, M.S. Yale, Rasch Analysis in The Human Sciences, Springer Netherlands, 2014.