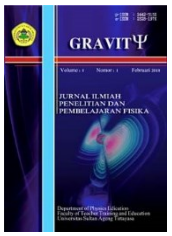
Gravity: Jurnal Ilmiah Penelitian dan Pembelajaran Fisika

http://jurnal.untirta.ac.id/index.php/Gravity

ISSN: 244-515x; e-ISSN: 2528-1976

Vol. 7, No. 2, August 2021, Page 21-28



Analysis of model fit and item parameter of work and energy test using item response theory

Yustiandi[1*], Duden Saepuzaman[2]

[1]*SMAN CMBBS, Indonesia*

[2]*Departement of Physics Education, Universitas Pendidikan Indonesia, Indonesia*

[*]*E-mail: yustiandi2@gmail.com*

(Received: 24 February 2021; Accepted: 19 July 2021; Published: 30 August 2021)

ABSTRACT

One of the important parts in assessing learning outcomes is using a good instrument that is analyzed using an appropriate analytical model and can measure students' abilities accurately. This study aims to determine the model fit and item parameter of work and energy test using item response theory. This research is a quantitative research that was carried out on the responses of 1177 high school students spread across Banten Province. The instrument is a set of work and energy tests consisting of 25 multiple choices. The data analysis used the item response theory approach with statistical methods ranging from determining the fit model to the item characteristics.. The analysis showed the students' responses to the fit energy and effort test with 9 items of 1 PL model, 17 items of 2 PL model, and 16 items of 3 PL model. Based on the percentage, the 2PL model is suitable than 1PL and 3PL. Further analysis determines the item parameter value by referring to the 2PL model, namely the item parameter difficulty level (b) and discrimination (a). The result shows that all items have difficulty in the range of 2.501 to 1.595, and the discrimination was in the range of 0.289 to 1.109. Based on this analysis, it can be concluded that all items in this test are the good item criteria

Keywords: item response theory, logistic parameter model, work and energy

DOI: 10.30870/gravity.v7i2.10563

INTRODUCTION

Today Indonesia's education is faced with quality challenges. Various programs are designed to produce quality education. A good program must be based on accurate data to produce optimal effects. This accurate data can be obtained through a good process. Mardapi (2017) shows that the quality of education can be improved through the quality of learning and quality of assessment. Teachers must be able to prepare learning materials developed based on the competencies and characters (Widya, Hamdi & ahmad, 2017). The right decision in the assessment system will be helpful for further decision-making.

Assessment of learning outcomes is carried out by providing tests that will assess students' abilities and determine completeness and achievement in certain fields of study (Gronlund, 1998). The more specific part of the assessment is measurement. Measurement is an activity to assign numbers to an individu-

al or individual characteristics according to certain rules (Griffin & Nix, 1991; Ebel & Frisbie, 1986).

In practice, the assessment is carried out using tests and non-tests. Generally, the most widely used assessment is a test. The test is a question given to the test to get answers from the test in the form of an oral or oral test or an action test or action test (Baskoro & Wihasko-ro, 2013). The test can also be viewed as part of a measurement of learning outcomes. The definition of measurement is an activity to dis-tinguish a person's characteristics or attributes (Oriondo & Antonio, 1998).

Using the test instrument for the assess-ment of learning outcomes is very important to ascertain the item parameters used. The pa-rameter of the test items used must meet the criteria for good items.

There are two approaches to estimating item parameters, namely classical test theory and item response theory. Classical test theory is seen to have weaknesses. According to Hambleton, Swaminathan, & Rogers (1991), the main drawback is that the characteristics of the examinees and the characteristics of the examinations cannot be separated, each of which can be interpreted only in other contexts. The test score itself only determines the ability of the examinee. When the test is difficult, the examinee will get a low score, and it can be concluded that the examinee's ability appears low.

On the other hand, when the test is easy, the examinee will get a big score and appear to have higher ability. In other words, the estimation of item parameters depends on the examinee and vice versa. Item characteristics will change when the examinee changes and the characteristics of the examinees change when the characteristics of the item change. Based on this explanation, there are limitations to the use of classical test theory because it will depend on the assessment subject.

Item response theory is a solution to overcoming weaknesses in classical test theory because item response theory has the concept of releasing the link between the Item and the test taker. The characteristics of the examinees will remain the same even though they work on the items with various characteristics, and vice versa, the characteristics of the test items will remain the same even though test-takers carry them out with different abilities. According to Hambleton et al., (1991), grain response theory rests on two basic postulates; (a) the ability of the test taker can be predicted (or explained) by a factor called trait, latent nature, or ability; and (b) the relationship between the abilities of the test taker and the characteristics of the test itself can be explained by a monotonically increasing function known as the Item characteristic function or item characteristic curve. This function explains that as the ability increases, the likelihood of the test taker answering correctly to an item increases. In Figure 1 we can see that the group of test-takers with high abilities will have a greater chance of answering correctly than the group with low abilities.

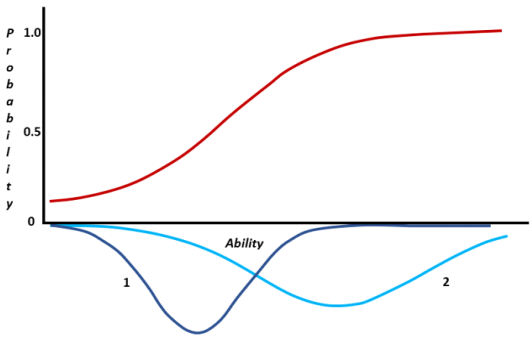


Figure 1. Item characteristic curves and ability distribution in the two groups of test-takers

In line with this, with the IRT analysis, the weaknesses of applying the cla ssical test theory can be resolved, namely: (1) the estimation of the test taker's ability does not depend on the characteristics of the test used; (2) estimated item parameters that do not depend on the ability of the teste; and (3) measurement error could be searched for each individual (Susongko, 2016).

The function of item response theory can be applied when the model used is compatible with the tested data (Hambleton et al., 1991). Stone & Zhang (2003) stated that grain estimation parameters could be

disturbed if the model used is not suitable.

Hambleton et al. (1991) describe several logistic models in item response theory, namely the one-parameter logistics model (1PL), the two-parameter logistic model (2PL), and the three-parameter logistic model (3PL). Each model has a certain number of grain parameters. Each parameter of an item will form an item response function.

The one-parameter logistic model (1PL) is an item response theory model with only one parameter: the level of difficulty. This model assumes that the test taker's ability is only affected by the difficulty level of the test items. An item is said to be good if it is in the range -2, which means easy to +2, which means difficult. The function of the PL model 1 can be seen in equation 1.

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

The two-parameter logistic model (2PL) has two parameters: the level of difficulty and discrimination, where the discrimination is in the range 0 and 2. In the grain characteristic curve, the discrimination is indicated by the slope of the curve. Items with high differing power have a steep curve. Grains with high differentiation power will better differentiate test takers who have a high ability from test-takers who have the low ability. The function of the 2PL model can be seen in equation 2.

$$Pi(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \quad (2)$$

The three-parameter logistic model (3PL) has three parameters: difficulty level, discrepancy, and pseudo-guessing. The pseudo -guessing parameter states the probability of a test taker with a low ability to answer a difficult question by guessing correctly. The value of pseudo guessing c ranges between 0 and 1. An item is good if the value of the parameter c is not more than 1/k, where k is the sum of selection. The function of the 3PL model can be seen in equation 3.

$$Pi(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \quad (3)$$

According to Retnawati (2014), two ways that we can use to prove the suitability of the model are statistical methods and graphical methods. The statistical method is done by calculating the chi-square value, comparing its value with a table, or looking at the probability value (significance). An item is said to be by the model if the results of the chi-squared calculation do not exceed the chi-squared value in the table or the sig> a value. While the analysis using the graph method is carried out by looking at the data distribution of the grain characteristic curve. Based on this curve, we can see the suitability of the data distribution compared to the model. The model is suitable if the distance from the point to the line is close (Retnawati, 2014).

RESEARCH METHODS

This research is quantitative research. Data obtained from student responses to the work and energy test instruments. The instrument used in this study was the Daily Physics Assessment of work and energy material. The test kit consists of 25 items in the form of multiple-choice and five choices. The test kits used previously were validated using Aiken validity. Respondents in the study were 1177 high school students spread across Banten province. The data collected was in the form of a dichotomy with a 1 if true and 0 if it was false.

The model of suitability analysis was carried out using statistical methods. After determining the appropriate model, the analysis determines the grain parameter values based on the appropriate model. The results of this item parameter analysis are seen from the output of BILOG MG 3.0 phase 2. The column "threshold" shows the difficulty level of item (b), "slope" shows the difference in power (a), and "asymptote" states the guessing parameter (c).

RESULTS AND DISCUSSION

Before the fit test stage of the appropriate or fit parameter model, the first thing to do is test the dimensional, whether unidimensional or multidimensional. Unidimensional means that each Item measures only one ability (Retnawati, 2014). Whereas multidimensional means that some or all items measure more than one dimension. The dimensional test in this study was proven through factor analysis using SPSS. Analysis factor was done by first doing a feasibility test analysis, namely the KMO-MSA test and the Barlett test. The KMO-MSA test aims to see the adequacy of the sample, while the Barlett test serves to prove the homogeneity of the data. Analysis factor can be continued if the Kaiser Meyer Olkin (KMO) -MSA value> 0.5 and Barlett's significant test <0.05 (Hair, JF, Black, WC, Babin, BJ, Anderson, RE, & Tatham, RL, 2009). Based on the response data in this study, the KMO-SMA and Barlett values were obtained as presented in table 1.

Table 1. KMO and bartlett's test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,938
Bartlett's Test of Sphericity	Approx. Chi-Square	4889,570
	df	300
	Sig.	,000

Based on Table 1, it can be seen that the KMO-MSA value is 0.938 and the significant Bartlett test is 0.000. It means that the sample used has met the sample adequacy requirements, and the data is homogeneous so that factor analysis can be carried out. The data processing results for factor analysis through SPSS can be seen in the eigenvalues section in Table 2.

Table 2. Eigenvalues

Component	Initial Eigenvalues		
	Total	% of Varians	Cumulative %
1	5.742	22.969	22.969
2	1.274	5.096	28.066
3	1.145	4.578	32.644
4	1.052	4.207	36.851

Based on table 2, the eigenvalues with more than one value indicate one factor. Based on these eigenvalues, the Work and Energy test instrument has three factors. These three factors can explain the 36, 851% variance. These eigenvalues can then be presented in the scree plot in Figure 3.

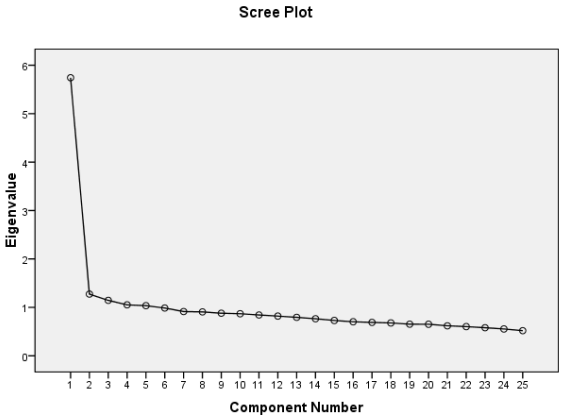


Figure 2. Scree plot analysis factor

The scree plot of the factor analysis shows a very sharp decrease between factor 1 and factor 2, and the Eigenvalue then begins to skew at a factor of 3 so that the scree plot almost forms a right angle. It shows that there is only 1 dominant factor in the work and energy material test.

Another test is local independence. This assumption of local independence will be fulfilled if the participant's answer to one Item does not affect the participant's answer to another item (Retnawati, 2014). According to De Mars (2010), local independence can also be detected by proving unidimensional assumptions. It can be interpreted that if the unidimensional assumptions are met, the local independence assumption is also fulfilled. In

this study, the unidimensional assumptions have been fulfilled so that the local independence test has also been fulfilled.

In this study, to determine the suitability or fitness of the logistic parameter model using statisticall analysis. Statistical analysis using Item fit has the power to detect measurement disturbances with a reasonable amount. The results show that, when the data fit the model, the distribution properties of the Item fit statistics, it is possible to construct a reasonable error rate (Smith, 1991). In this study, the suitability or fitness of the model was determined using statistical methods, namely by determining the chi-square for each Item on each logistic parameter. The technique of this method is to compare the calculated chi-square value with the chi-square table value in certain degrees of freedom. An item is deemed suitable to the logistic parameter model if thecalculated chi-square (χ^2) value does not exceed the table or critical (χ^2_{Crit}) chi-square value. The suitability of each Item in the 1PL, 2 PL, and 3 PL models is presented in Table 3.

Based on table 3, the number of items that fit the 1 PL model is 9 items, the 2 PL model is 17 items, and the 3 PL model is 16 items. If viewed from the percentage, the suitability with the 2PL model is the greatest compared to the 1PL and 3 PL. So it can be concluded based on this analysis that the analysis of the Work and Energy test instrument fits the 2PL parameter model. When the model fits the data, the model has shown conformity (Hattie, 1984).

Similar things can cause the number of items that do not fit the person or person fit. Meijer (1996) states that there are at least seven behaviors of test-takers when the test causes the items not to match the data. The seven behaviors, namely; a) sleep behavior, an examiner has difficulty starting a task, and after adapting, he does not check the answer; b) Guessing behavior (guessing), in which the examinee with low ability suddenly responds correctly to a complicated item; c) fraudulent behavior; d) Plodding or sluggish behavior, namely test takers who have not finished working on the problem; e) Alignment errors, occur to examinees who do not carefully respond to the answer sheets; f) too creative, that is, the examinee interprets the Item in an unusual or too creative way; g) lack of ability, occurs when the problem is measuring two different abilities. Further analysis, namely determining or estimating the difference between power parameters (a) and the level of difficulty (b) using the 2PL model. The results of this analysis produce parameter values for each Item that are presented in table 4.

Table 3. The suitability of each item in the PL, 2 PL, and 3 PL models

No.	1PL				2PL				3PL			
	χ^2	df	χ^2_{Crit}	Ket.	χ^2	df	χ^2_{Crit}	Ket.	χ^2	df	χ^2_{Crit}	Ket.
1	48,1	7	18,48	No fit	13,1	6	16,81	fit	28,8	7	18,48	No fit
2	90,8	6	16,81	No fit	19,3	7	18,48	No fit	34,1	7	18,48	No fit
3	116,9	8	20,09	No fit	21,3	9	21,67	fit	12,1	9	21,67	fit
4	43,5	6	16,81	No fit	4,3	6	16,81	fit	26,3	7	18,48	No fit
5	13,5	8	20,09	fit	14,1	9	21,67	fit	13,3	9	21,67	fit
6	59,2	8	20,09	No fit	12	8	20,09	fit	14,4	7	18,48	fit
7	19,6	8	20,09	fit	5,2	9	21,67	fit	6,2	9	21,67	fit
8	7,8	8	20,09	fit	7,8	9	21,67	fit	9,4	9	21,67	fit
9	13,8	7	18,48	fit	6,5	8	20,09	fit	14,7	8	20,09	fit

10	54,5	7	18,48	No fit	26,4	8	20,09	No fit	30,1	8	20,09	No fit
11	4,4	8	20,09	fit	8	9	21,67	fit	9,3	9	21,67	fit
12	185,7	8	20,09	No fit	38,1	9	21,67	No fit	33,4	9	21,67	No fit
13	23,1	8	20,09	No fit	8,6	9	21,67	fit	6,9	9	21,67	fit
14	77	7	18,48	No fit	38,4	9	21,67	No fit	10,8	8	20,09	fit
15	21,9	6	16,81	No fit	5	7	18,48	fit	6,7	7	18,48	fit
16	13,4	4	13,28	No fit	7,1	4	13,28	fit	42,8	5	15,09	No fit
17	6,6	7	18,48	fit	6,3	8	20,09	fit	27,5	8	20,09	No fit
18	29,6	8	20,09	No fit	23,7	9	21,67	No fit	3,6	9	21,67	fit
19	30,6	8	20,09	No fit	22	9	21,67	No fit	11,1	9	21,67	fit
20	8,4	8	20,09	fit	3	9	21,67	fit	4,7	9	21,67	fit
21	32,3	7	18,48	No fit	4,8	8	20,09	fit	3,1	8	20,09	fit
22	68,4	8	20,09	No fit	7,2	8	20,09	fit	4,7	8	20,09	fit
23	16,1	7	18,48	fit	22,8	8	20,09	No fit	29	8	20,09	No fit
24	34,4	8	20,09	No fit	24,3	9	21,67	No fit	34,1	9	21,67	No fit
25	8,9	8	20,09	fit	16,2	9	21,67	fit	4,7	9	21,67	fit
SUM		fit 1 PL		9		fit 2 PL		17		fit 3 PL		16

Table 4. Estimated parameters using the 2PL model

Item	a	b
1	1.109	-1.518
3	0.289	1.595
4	1.032	-0.837
5	0.810	0.025
6	1.060	-0.333
7	0.407	-1.021
8	0.707	0.036
9	0.977	-0.919
11	0.692	-0.482
13	0.933	0.592
15	0.993	-0.598
16	0.917	-2.501
17	0.605	-1.495
20	0.625	0.789
21	1.025	-0.098
22	1.067	0.186
25	0.549	0.000

Based on the data in table 4, it appears that the difficulty level is in the range - 2.501 to 1.595 and the discrimination is in the range 0.289 to 1.109. For the value of the item difference index, Alagumalai has grouped the index into: very good> 0.40, good 0.30–0.39, just 0, 20 - 0, 29 unable to distinguish 0.00 - 0.19, requires examination of items <0.00 (Alagumalai et.al., 2005). Based on this analysis, it can be concluded that all items in this test meet the criteria for good items.

CONCLUSION

The analysis results showed that the student's responses to the fit energy and effort tests with the 1 PL model were 9 items, the 2 PL model was 17 items, and the 3 PL model was 16 items. If viewed from the percentage, the suitability with the 2PL model is greatest than the 1PL and 3 PL. So it can be concluded based on this analysis that the analysis of the

Work and Energy test instrument fits the 2PL parameter model. Further analysis is determining the item parameter value by referring to the 2PL model, namely the item parameter difficulty level (b) and discrimination (a). From the analysis conducted, it was found that the level of difficulty of the problem was in the range - 2.501 to 1.595 and the power of difference was in the range 0.289 to 1.109. Based on this analysis, it can be concluded that all items in this test meet the criteria for good items.

REFERENCES

Alagumalai, S., Curtis, D. D., & Hungi, N. (2005). Applied Rasch measurement: A book of exemplars (pp. 1-15). Dordrecht,

Anggreyani, A. (2009). *Penerapan Teori Uji Klasik dan Teori Respon Butir dalam Mengevaluasi Butir Soal*. Skripsi pada Departemen Statistika IPB: tidak diterbitkan.

Aristiawan, A., Retnawati, H., & Istiyono, E. (2019). *Analysis of Model fit and Item Parameter of Mathematics National Examination Using Item Response Theory*. JPP (Jurnal Pendidikan dan Pembelajaran), 25(2), 40-46.

Baskoro, E. P., & Wihaskoro, A. M. (2013). Modul Perkuliahan Evaluasi Pembelajaran.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Djemari, M. (2008). *Teknik penyusunan instrumen tes dan non tes*. Yogyakarta: Mitra Cendekia.

Ebel, R. L., & Frisbie, D. A. (1986) Essentials of. Educational Measurement

Griffin, P., & Nix, P. (1991). Assessment and reporting: A new approach.

Gronlund, N. E. (1998). Assessment of student achievement. Allyn & Bacon Publishing, Longwood Division, 160 Gould Street, Needham Heights, MA 02194-2310; tele.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise multivariada de dados. Bookman Editora*.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. Boston, MA : Kluwer.Inc

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. Multivariate behavioral research, 19(1), 49-78.

Horn, R., Wolf, L., & Velez, E. (1992). *Sistemas de medición de evaluación educacional América Latina*. Reseña temática y experiencias recientes.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Dorsey Press.

Oriondo, L. L., & Antonio, E. M. D. (1998). Evaluating Educational. Outcomes, Manila: Rex Book Store.

Pollard, A., & Collins, J. (2005). *Reflective teaching*. A&C Black.

Pollard, Andrew (Edited). *Reading Reflective teaching*. London : Continuous International Publishing Group

Retnawati, H. (2008). *Estimasi efisiensi relatif tes berdasarkan teori respons butir dan teori tes klasik*. Disertasi. Yogyakarta: Program Pascasarjana Universitas Negeri Yogyakarta.

Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Nuha Medika.

Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian*. Yogyakarta: Parama Publishing.

Retnawati, H. (2016). *Validitas reliabilitas dan karakteristik butir*. Yogyakarta: Parama Publishing.

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. Educational and psychological measurement, 51(3), 541-565.The Netherlands:: Springer.

Stone, C. A., & Zhang, B. (2003). *Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures*. Journal of Educational Measurement, 40(4), 331-

352.

Susongko, P. (2016). Validation of science achievement test with the rasch model. Jurnal Pendidikan IPA Indonesia, 5(2), 268-277.

Widya, Hamdi dan Ahmad. (2017). Kualitas Perangkat Pembelajaran Fisika Berbasis Model Creative Problem Solving Dengan Pendekatan Open-Ended Pada Materi Usaha Dan Energi Terintegrasi Energi Biomassa. Gravity : Jurnal Ilmiah Penelitian dan Pembelajaran Fisika, 3(2), 158-171.