22 (1), 2021, 1-9

Jurnal Pendidikan MIPA

e-ISSN: 2550-1313 | p-ISSN: 2087-9849

http://jurnal.fkip.unila.ac.id/index.php/jpmipa/

Analysis of Items Parameters on Work and Energy Subtest Using Item Response Theory

Duden Saepuzaman[1,2,*], Haryanto[2], Edi Istiyono[2], Heri Retnawati[2], Yustiandi[3]

[1]Departemen of Physics Education, Universitas Pendidikan Indonesia, Indonesia
[2]Department of Educational Research and Evaluation, Yogyakarta State University, Indonesia
[3]SMA Negeri CMBBS, Indonesia

Abstract: This study aims to describe the Physics test item parameters in Work and Energy and describe students' abilities using the item response theory approach (IRT) dichotomous scoring. This research is quantitative descriptive. The research subjects were 1175 high school class XI students in West Java and Banten provinces consisting of 450 male students and 725 female students. The instrument used was Physics of Work and Energy as many as 25 items in multiple choices with dichotomous scoring. Student response data with dichotomous scoring were analyzed using the item response theory approach using the BILOG-MG program. The results showed that most of the items fit the 2PL model. Subsequent analysis of the items' characteristics indicates that all items have different power and a level of difficulty in the good criteria.

Keywords: item parameters, item response theory, physics test.

Abstrak: *Penelitian ini bertujuan mendeskripsikan parameter butir soal Fisika pada materi Usaha dan Energi dan mendeskripsikan kemampuan peserta didik dengan menggunakan pendekatan teori respon butir (Item Response Theory, IRT) penskoran dikotomus. Penelitian ini merupakan penelitian deskriptif kuantitatif. Subyek penelitian sebanyak 1175 siswa SMA kelas XI yang tersebar di provinsi Jawa Barat dan Banten yang terdiri 450 siswa laki-laki dan 725 siswa perempuan. Instrument yang digunakan berupa soal Fisika materi Usaha dan Energi sebanyak 25 butir berbentuk multiple choice dengan penskoran dikotomus. Data respon siswa dengan penskoran dikotomus dianalisis menggunakan pendekatan teori respon butir menggunakan program BILOG-MG. Hasil penelitian menunjukkan bahwa butir soal paling banyak fit dengan model 2PL. Analisis berikutnya mengenai karakteristik butir soal menunjukkan bahwa keseluruhan butir memiliki daya beda dan tingkat kesulitan dalam kriteria good.*

Kata kunci: *parameter butir, teori respon butir, tes fisika.*

Duden Saepuzaman et al.
Email: dudensaepuzaman.2019@student uny.ac.id

DOI: http://dx.doi.org/10.23960/jpmipa/v22i1.pp1-9
Received: 12 February 2021
Accepted: 25 March 2021

▪ INTRODUCTION

One of the elements that must be considered in implementing learning outcomes is to seek and ensure that the evaluation or assessment of learning outcomes accurately describes students' abilities. Assessment is an important component of education because it provides information about the learning process, measures student achievement, and evaluates learning activities' effectiveness (Wang & Bao, 2010; Pellegrino et al., 2001; National Academy of Sciences, 1996; Black et al. 2004). An assessment is called accurate if the results of the assessment contain the smallest possible error or error. To get accurate information, tell students' abilities, the instruments' quality must be valid, reliable, and have good item parameters. For this purpose, two types of approaches can estimate item parameters, namely classical test theory and item response theory (Neşe Gŭler, et al., 2014). Classical test theory is an approach that is very simple and easy to understand in an empirical analysis of questions. Traditionally, the ability of each examinee is reported in terms of the number of items answered correctly. It is a limitation or weakness of problem analysis with the classical test theory approach. Where students with the same number of items answered correctly may have different response patterns (i.e., correct answers on different items) and, as such, may not have a proficiency level. The same was measured by the test (Cappelleri, Lundy & Hays, 2014). The most prominent disadvantage of classical test theory is that the examinee's characteristics and the test characteristics are inseparable, each of which can only be interpreted in another context (Hambleton, Swaminathan, & Rogers, 1991). That is, the test only determines the ability of the examinees. When the test is difficult, the examinee will appear to have a low ability. And when the test is easy, the examinee will appear to have higher ability. In other words, the item parameters depend on the subject/test taker and vice versa. The item's characteristics will change when the examinee changes. And the characteristics of the examinees will change when the items change. In this case, classical test theory cannot be used as a standard because it depends on the test taker subject.

Item response theory is a solution to overcoming weaknesses in classical test theory because item response theory has the concept of releasing the relationship between items and samples or test taker subjects. Item response theory is a type of measurement model based on a statistical framework (Junker, 1999). Item response theory (IRT) is a popular statistical framework used to model the interaction between unobserved psychological constructs (or traits) and item-level stimuli (Chalmers et al., 2017). In the item response theory (IRT) approach, complexity has been measured almost exclusively by estimating parameters (Bonifay & Cai, 2017). The examinees' characteristics/abilities will remain the same even though they are working on different characteristics.

Conversely, the items' characteristics will remain the same even though test takers perform them with different abilities. Also, item response theory is based on items, not on test kits. According to Hambleton et al. (1991), item response theory rests on two postulates. *First*, a test taker's performance on test items can be predicted (or explained) by a set of factors called traits, latent traits, or abilities. *Second*, the relationship between the test taker's performance and the item can be explained by a monotonically increasing function called the item characteristic curve (ICC) function. This function explains that as ability increases, the respondent's probability of answering correctly for an item also increases. Figure 1 shows that the test-takers with a higher ability will have a greater probability of answering correctly than the group of test-takers with low ability.

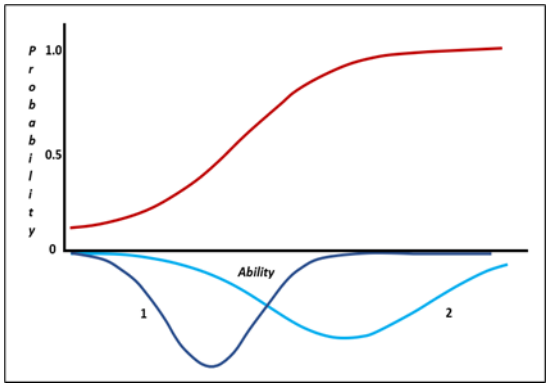

Figure 1. ICC curve and ability distribution in the two test-taker groups (adapted from Hambleton et al., 1991)

The function of item response theory can be applied when the model used is compatible with the test data (Hambleton et al., 1991). Stone & Zhang (2003) stated that grain parameter estimation could be disturbed when the model used does not match the data. Hambleton et al. (1991) describe several logistic models in the goods 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 different number of item parameters. This item parameter functions as a form of the grain response function.

The statistical method is performed by calculating the chi-square value (χ^2), comparing it with the chi-square value from the table, or reviewing the probability value (significance). Items are said to fit the model if the calculated chi-square value is smaller than the chi-square table or the value of sig> α. The correct parameter model's final determination is determined from the suitability of items with the most logistical parameters (1PL, 2PL, and 3 PL). As for the graph method, it can be seen from the item characteristic curve (ICC). Through this curve, it can be seen how precisely the data distribution is compared to the model. This model is suitable if the match line's point distance is very close (Retnawati, 2014). The appropriate parameter model or fit's final determination is the same as determining statistics determined from the most suitable items with the logistic parameter type (1PL, 2PL, and 3 PL).

This study is focused on the analysis of the physics item parameter analysis with the dichotomous item response theory approach. It becomes important to reduce the weaknesses caused by the classical test theory approach.

▪ METHOD

This research is quantitative descriptive. The research subjects were 1177 high school class XI students in West Java and Banten provinces consisting of 451 male students and 726 female students. Response data with dichotomous scoring were analyzed using the item response theory approach with the BILOG-MG program. The initial step taken was the IRT assumption test, namely unidimensional, local independence, and invariant parameters. Further analysis is to identify the characteristics of each item's parameters based on the BILOG MG output. An item is said to be good if the difficulty level (b) is in the range -2 to +2 (Hambleton & Swaminathan, 1985) and the difference (a) is in the range 0 to 2 (Hambleton & Swaminathan, 1985).

▪ RESULT AND DISCUSSION

Before the fit model test, the first analysis is the assumption test is a unidimensional test. Unidimensional means that each item measures only one ability (Retnawati, 2014). At the same time, multidimensional implies that some or all items measure more than one dimension. The dimensional test in this study was proven through factor analysis using SPSS. Factor analysis 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 sample's adequacy, while the Barlett test serves to prove the data's homogeneity. Factor analysis 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, KMO-SMA and Barlett scores, 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. Factors in factor analysis have greater eigenvalue than one expressed as significant factors (Retnawati, 2014; Putri dkk., 2015). It means that the factors or components contained in the instrument are known from the number of more than one eigenvalues.

Table 2. Eigenvalue of every components

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 total eigenvalues with more than one indicate one factor (Retnawati ,2014; Putri dkk. 2015). Based on these eigenvalues, the Work and Energy test instruments have four factors. These four factors can explain the 36.851% variance. Although four factors have eigenvalues of more than one, it appears that if analyzed, the first factor or component has an eigenvalue that is much greater than the eigenvalues of other factors. It shows that there is only one dominant factor in the test set to fulfill the unidimensional assumptions.

This eigenvalue can then be presented in the scree plot in Figure 2. A scree plot that shows the eigenvalues sorted from greatest to smallest is often used to represent and analyze dimensions or factors (Reckase, 1979).

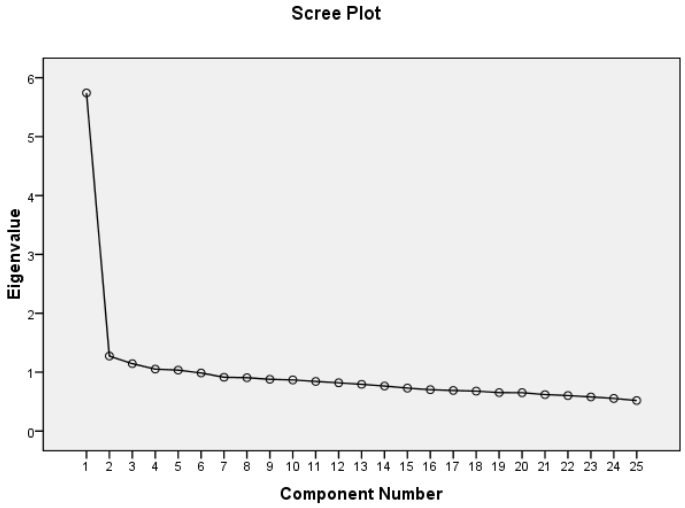

Figure 2. Scree plot factor analysis

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 one dominant factor in the Work and Energy test. It shows that there is only one dominant factor in the Work and Energy instrument test set to fulfill the unidimensional assumptions. This study's results are consistent with the assumption in the item response theory approach where a set of questions or tests only has one latent trait. This result implies that each examinees' performance is assumed governed by a single factor, referred to as ability (Eleje & Onah, 2018).

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, the model's suitability was determined using statistical methods by determining each item's chi-square on each logistic parameter. This method's technique compares the calculated chi-square value with the chi-square table value at certain degrees of freedom. An item is deemed suitable to the logistic parameter model if the calculated chi-square value (χ^2) does not exceed the table chi-square value or critical (χ^2_{crit}) value. The suitability of each item in the 1PL, 2 PL, and 3 PL models is presented in Table 3.

Table 3. The fitness of each item in the 1 PL, 2 PL, and 3PL models

Item	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	Not fit	13.1	6	16.81	Fit	28.8	7	18.48	Not fit
2	90.8	6	16.81	Not fit	19.3	7	18.48	Not fit	34.1	7	18.48	Not fit
3	116.9	8	20.09	Not fit	21.3	9	21.67	Fit	12.1	9	21.67	Fit
4	43.5	6	16.81	Not fit	4.3	6	16.81	Fit	26.3	7	18.48	Not 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	Not 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	Not fit	26.4	8	20.09	Not fit	30.1	8	20.09	Not 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	Not fit	38.1	9	21.67	Not fit	33.4	9	21.67	Not fit
13	23.1	8	20.09	Not fit	8.6	9	21.67	Fit	6.9	9	21.67	Fit
14	77	7	18.48	Not fit	38.4	9	21.67	Not fit	10.8	8	20.09	Fit
15	21.9	6	16.81	Not fit	5	7	18.48	Fit	6.7	7	18.48	Fit
16	13.4	4	13.28	Not fit	7.1	4	13.28	Fit	42.8	5	15.09	Not fit
17	6.6	7	18.48	Fit	6.3	8	20.09	Fit	27.5	8	20.09	Not fit
18	29.6	8	20.09	Not fit	23.7	9	21.67	Not fit	3.6	9	21.67	Fit
19	30.6	8	20.09	Not fit	22	9	21.67	Not 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	Not fit	4.8	8	20.09	Fit	3.1	8	20.09	Fit
22	68.4	8	20.09	Not 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	Not fit	29	8	20.09	Not fit
24	34.4	8	20.09	Not fit	24.3	9	21.67	Not fit	34.1	9	21.67	Not 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**

Based on table 3, it can be seen that 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 greatest than the 1PL and 3 PL. So it can be concluded based on this analysis that the Business and Energy test instrument's

analysis fits the 2PL parameter model. It is possible because the 2 PL model is a parameter model that has fewer assumptions than the 1PL. Mardapi (2008) states that the 1 PL model is the item response theory model with the most assumptions compared to the 2 PL and 3 PL models. Model 1 PL will only estimate the item difficulty parameter.

In contrast, the discrimination parameter or item difference power must be considered the same, and the false guess parameter must be assumed to be zero. Model 2 PL has item parameters of difficulty level and difference power, while the pseudo guess parameter is supposed to be zero. Because the 1 PL model has the most assumptions, this 1 PL model will produce several items that are appropriate. It is in great agreement with the data in table 3, which shows that only about 36% (9 of 25) items fit with the PL model 1. The number of items that fit with 1 PL is smallest than the 2PL and 3 PL models.

The fit and failure of the data with the 1PL, 2PL, or 3 PL models are due to several things related to the test takers' behavior. Meijer (1996) states that at least seven test takers' behaviors 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 difficult 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 was carried out, namely estimating the item parameter value by referring to the 2PL model, namely the item parameter of difficulty level and different power. In general. the results of parameter estimation using the 2PL model are presented in Table 4.

Table 4. Parameters for the test items for work and energi test instrument

Indicator	Item	a	b	Criteria
Defines work concept	1	1.154	-1.511	good
	2	1.355	-0.981	good
	3	0.227	1.838	good
Determine the magnitude of the work using the equation	4	1.163	-0.870	good
	5	0.776	-0.260	good
	6	1.075	-0.334	good
	7	0.405	-1.119	good
Determines work based the F-s graph	8	0.690	-0.047	good
	9	0.774	-1.170	good
Determines potential energy	10	1.061	-0.650	good
	11	0.696	-0.504	good
	12	0.170	-1.228	good
	13	0.864	0.667	good
Defining energy kinetic concept	14	0.328	-0.656	good
	15	1.002	-0.609	good
	16	1.123	-1.266	good
Analyze relationship	17	0.599	-1.425	good
	18	0.828	0.092	good

Indicator	Item	a	b	Criteria
work with the change of energy	19	0.864	0.476	good
	20	0.668	0.670	good
Understand power concept	21	1.012	-0.134	good
	22	1.096	0.179	good
Analyze The Conservation energy Law	23	0.592	-1.846	good
	24	0.394	-1.514	good
	25	0.598	0.129	good

Table 4 indicates that the overall parameters of the test items of effort and energy using dichotomous scoring have good criteria. An item said to be a good criterion if the difficulty level (b) is in the range -2 to +2 (Hambleton and Swaminathan. 1985) and the difference (a) is in the range of 0 to 2 (Hambleton and Swaminathan. 1985).

Based on these findings, the instrument developed has good construct validity and reliability. Further analysis shows that the items' characteristics, including the power of difference and the level of difficulty of the whole items, are in good criteria. This test is empirical evidence that the effort and energy test instrument is suitable for measuring student ability as a measuring tool for assessing learning outcomes.

▪ CONCLUSION

Based on the study results, the most suitable scoring model (fit) in estimating item parameters and the ability of Work and Energy Physics questions is the 2-PL model. The dichotomous scoring item parameters' analysis resulted in the Work and Energy questions having good value for the difficulty level parameters and the different power. This research is expected to provide an overview in testing the validity and reliability of a test instrument's construct. This research is still limited to testing the validity and reliability of constructs that previously passed the expert's content validity test. The next study conducted tested the criteria's validity so that the resulting instruments were more reliable in measuring student learning outcomes and abilities.

▪ REFERENCES

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappa, 86*(1), 8-21.

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research, 52*(4), 465-484.

Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical therapeutics, 36*(5), 648-662.

Chalmers, R. P., Pek, J., & Liu, Y. (2017). Profile-likelihood confidence intervals in item response theory models. Multivariate Behavioral Research, 52(5), 533-550.

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

Eleje, L., I & Onah, F., E. (2018). Comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test item analysis results. *European Journal of Educational & Social Sciences*. 3(1), 71-89

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of educational measurement*, *44*(4), 325-340.

Güler, N., Kaya Uyanik, G., & Taşdelen Teker, G. (2014). Comparison of classical test theory and item response theory in terms of item parameters. *European Journal of Research on Education*, *2*(1), 1-6.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). Análise multivariate 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

Junker, B. W. (1999). Some statistical models and computational methods that may be useful for cognitively-relevant assessment. Prepared for the National Research Council Committee on the Foundations of Assessment. Retrieved April, 2, 2001.

Mardapi, D. (2008). Teknik penyusunan instrumen tes dan non tes. Yogyakarta: Mitra Cendekia.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9*(1), 3-8.

National Academy of Sciences-National Research Council, Washington, DC., National Research Council (US)., National Research Council Staff, National Research Council, Board on Science Education Staff, Division of Behavioral, ... & Assessment Staff. (1996). National science education standards. Joseph Henry Press.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). Knowing what students know: The science and design of educational assessment. National Academy Press, 2102 Constitutions Avenue, NW, Lockbox 285, Washington, DC 20055.

Putri, N. K. L., Asih, N. M., & Nilakusmawati, D. P. E. (2015). Faktor-faktor yang Menentukan Kepuasan Pelanggan Sepeda Motor Matic Honda di Kota Denpasar. *E-Jurnal Matematika*, *4*(1), 1-7.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of educational statistics*, *4*(3), 207-230.

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

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.

Wang, J., & Bao, L. (2010). Analyzing force concept inventory with item response theory. *American Journal of Physics*, *78*(10), 1064-1070.