PEDAGOGIK | JURNAL PENDIDIKAN
P-ISSN : 2528-5793
E-ISSN : 2354-7960

ITEM RESPONSE THEORY ANALYSIS ON STUDENT STATISTICAL LITERACY TESTS

Mila Yulia Herosian[1]; Yeni Rafita Sihombing[2]; Delyanti Azzumarito Pulungan[3]
[1,2,3] Universitas Prime Indonesia, North Sumatera, Indonesia
[2] Contributor: yenirafitasihombing@unprimdn.ac.id

Received: June 2022	**Accepted:** January 2023	**Published:** January 2023
DOI: https://doi.org/10.33650.pjp.v9i2.3800		

Abstract : *This study aims to evaluate the quality of statistical literacy instruments. With the ex post facto approach, the results of the statistical literacy tests of 150 students were used. Data were analyzed using PARSCALE, based on the Item Response Theory with a two-parameter GRM (2-PL) logistic model, namely item discriminant and item difficulty. The results showed that of the 12 items analyzed, all were good, and the GRM-IRT 2-PL model was the right model for statistical literacy instruments with polytomous answer types. The elements in the statistical literacy instrument have been constructed with accurate information about statistical literacy abilities ranging from -2.5 to +1.2. The least SEM occurs when the maximum test information function is assigned to the test set; therefore, it can be recommended for item analysis on other tests. The statistical literacy test items examined in this study can be used to assess student literacy statistics in various schools and regions.*

Keywords : *Item Response Theory; Model 2-paramater Logistics; Statistical Literacy Test.*

Abstrak : *Penelitian ini bertujuan untuk mengevaluasi kualitas instrumen tes literasi statistic. Dengan pendekatan ex post facto, digunakan hasil tes literasi statistik dari 150 siswa. Data dianalisis menggunakan PARSCALE, berdasarkan Item Response Theory dengan model logistik dua parameter GRM (2-PL), yaitu daya beda dan tingkat kesukaran. Hasil penelitian menunjukkan bahwa dari 12 soal yang dianalisis dengan IRT GRM 2-PL, semua butir soal baik dan model GRM IRT 2PL merupakan model yang tepat untuk instrumen literasi statistik dengan tipe jawaban polytomous. Unsur-unsur dalam instrumen literasi statistik telah dibangun dengan informasi yang akurat tentang kemampuan literasi statistik mulai dari -2,5 hingga +1,2. SEM paling kecil terjadi ketika fungsi informasi tes maksimum diberikan ke set tes, oleh karena itu dapat direkomendasikan untuk analisis item dalam tes lainnya. Butir tes literasi statistik yang diteliti dalam penelitian ini dapat digunakan untuk menilai statistik literasi siswa di berbagai sekolah dan daerah.*

Kata Kunci : *Item Response Theory; Model 2-Parameter Logistic; Instrumen Literasi Statistik.*

INTRODUCTION

The quality of learning will be known based on a good assessment. On the contrary, good learning will produce good assessments. Thus, the quality of learning and assessment have a strong relationship. A good assessment requires good measuring tools, in the form of tests and non-tests. This is because the test must be able to provide accurate information about the student's ability as well as evaluate the learning process. A good test must have items that meet the requirements based on test features and must provide information with the least number of errors (Mulianah & Hidayat, 2013; Suwarto, 2016). Therefore, the nature of the test must be checked so that it can be trusted to measure the abilities of students(Isnani et al., 2019; Leal Filho et al., 2018; Purnamasari et al., 2020).

One of the abilities that can be measured is statistical literacy. In order to measure students' statistical literacy well, a high-quality statistical literacy test is needed. The purpose is to differentiate between students with high, medium, and low statistical literacy. In order to get a test with that ability, it is necessary to analyze the statistical literacy test items with a point analysis approach. Question-item analysis can be done qualitatively or quantitatively, but the goal is to get information about the characteristics of each question item, both through the examination of question items and empirical analysis. The characteristics of a good test will be focused on the analysis of quantitative questions. Richard and Sheila (1999) explained that quantitative grain analysis is a grain study based on empirical data from tested tests (Suwarto, 2022). There are two types of quantitative analysis, namely analysis based on classical test theory and modern test theory.

Two approaches that are always used to analyze items in education measurement are classical test theory (CTT) and item response theory (IRT) (Awopeju & Afolabi, E. R. I., 2016). Measurement theory, especially in the field of education, initially used Classical Test Theory (CTT) as a detailed analysis approach, which later became the basis for the development of measurement theory (Erguven, 2013). Unfortunately, CTT has shortcomings in describing students' abilities. This is because in the CTT, students' ability is explained by the number of scores obtained without considering the interaction of each student with the test (Amelia & Kriswantoro, 2017). This is contrary to the fact that every student must have different abilities (Akinde et al., 2017; Ida, 2021).

Another weakness is the level of difficulty of the question and the power of the question depending on the sample (Hambleton & Swaminathan, 1985). This dependency means that the characteristics of the items analyzed with CTT will vary according to the student's abilities. Thus, a question item will have a low level of difficulty (be an "easy item") if it is completed by a student with high ability. If completed by students with limited ability, it will be extremely difficult. Another disadvantage of CTT is that it is more test-oriented than question-oriented (Hambleton et al., 1992). The limitations of CTT are then solved with the IRT (item response theory) approach, given the nature of group dependence and item dependence.

There are many reference studies that discuss IRT, but very few explain the steps and guidelines of how teachers apply IRT so that they can improve the quality of exams and learning. There are also many references that explain the characteristics of the item and discuss the response function of the item, but have not yet explained how to measure it or what it is used for. As a result, IRT was developed to address the shortcomings of classical test theory and work best in overcoming the limitations of classical theory.

IRT developed from classical measurement theory to overcome various limitations (Hambleton, 1985). IRT is a general form of a mathematical function that explains the interaction between subjects and test items (Sumintono & Widhiarso, 2014). The estimation of item parameters or students' ability in IRT does not depend on a specific item sample or students selected in the test. Thus, the characteristics of the question items will remain the same for different groups of test students. The level of student ability will also increase in some test items when IRT is applied for item analysis (Isnani et al., 2019). More specifically, with the items whose characteristics have been analyzed with the IRT approach, a group of students with any level of ability has no problem as long as they have enough ability to do the questions. Whatever group test the student takes will not be a problem as long as the group test can be done well by the student.

IRT is a statistical theory that consists of various mathematical models that have the following characteristics: a) they predict student scores based on student ability or latent traits; b) they build relationships between student item performance and the set of traits underlying the item. performance through a function called "item characteristic curve" (Hambleton et al., 1992). IRT models provide invariance of item parameters and abilities for test items and people when the IRT model fits the available test data. In other words, the same item used in different samples will maintain its statistical properties, and the students' scores that represent the ability or latent nature of a particular construct will not depend on the particular test item given to them.

IRT assumption is a prerequisite test that must be completed before performing IRT analysis. Before doing the IRT analysis, it is necessary to make assumptions that must be fulfilled by the question and the test participants. These assumptions include unidimensional parameters, local independence, and invariance. In addition to the opinion about the three general assumptions, there is also an opinion that states that at least two assumptions must be fulfilled in IRT: unidimensionality and local independence (Embretson & Reise, 2013). The assumption that each test set measures only one ability is called unidimensional. Unidimensional means that the test only measures one characteristic of the participant (Crocker & Algina, 1986; Jumailiyah, 2017). Unidimensional means that the test only measures one character or ability of the person taking it. Common ways to check dimensional assumptions are: plotting eigenvalues (variability of 20% or more in the first factor); parallel analysis; or confirmatory factor analysis (testing the hypothesis of one

factor); a technique to evaluate unidimensionality is tetrad analysis and the Hull method (Hattie, 1985; Horn, 1965; Lorenzo-Seva et al., 2011).

Furthermore, the influence of the participant's ability and the question is considered constant when the student's response to the question does not have a statistically related relationship. This condition is assumed with local independence. The assumption of local independence is categorized as satisfactory when the student's answer to one item does not affect the answer to another item. There is no correlation between student responses to multiple items on a series of tests (Hambleton et al., 1992). As a result, local independence is assumed item by item, and participants are evaluated individually. An invariant assumption is defined as a function of the characteristics of an item in a constant or fixed condition. The function of the item did not change, even though the participant group responded to the item change. Thus, in the same group, the characteristics of the questions will remain even if the questions answered change. From the characteristics of the items and the characteristics of the participants, the difficulty of the items and the ability to differentiate the items are still prominent even if the items of the question are answered by high-ability groups or low-ability groups. The student's ability will be constant or remain unchanged even if the question he answers changes. One of the most common assumptions is that, in each test, only one ability is measured by the item instrument. This assumption is called a "unidimensional assumption".

The type of IRT model will depend on the research question, the field of study, and how many item parameters are estimated and held constant. There are 3 models in IRT: 1-PL, 2-PL, and 3-PL. The IRT 2-PL model, used in our example (below), estimates item discrimination (slope) and item difficulty. The IRT 3-PL model estimates item discrimination, item difficulty level, and prediction parameters. Items that contain item difficulty parameters (parameter b) are called Rasch models or 1-PL models. If it contains 2 parameters, then the item difficulty level parameter (b) and power difference (a) are called logistic model parameters 2 (2-PL). If it contains three parameters in the form of difficulty item parameters (b), power difference (a), and pseudo-guessing (c), then it is called 3-PL.

The slope of the item allows one to determine how well the item identifies the patient at different levels of the latent trait. Discrimination power is an index that shows the ability to distinguish between students who perform well (top group) and those who perform poorly (bottom group). Levels offer better discrimination than less steep slopes, as depicted in the ICC. The item discrimination parameter estimate for item i is denoted by the symbol a. The theoretical value range for AI is - to +; however, items with negative AI values are problematic because they indicate that respondents with higher levels of latent traits are less likely to support more severe response options. This has the potential to happen if the item cannot differentiate well between those who have high and low ability levels. Tests that do not have discriminating power will not produce an accurate picture of student ability. Therefore, it is important to measure the discriminating power of a test item to produce a good test device.

Item difficulty (location) means how difficult it is to achieve a probability of 0.5 of a correct response for a specific item based on the respondent's latent variable (theta) level.

The more difficult it is for a student to have a 50% chance of answering an item correctly, the higher the level of ability required to achieve this goal. Question difficulty is defined as the proportion of test takers who answer the question correctly. The item index (p-values) can be calculated based on the test participant's response to the item, and a good item should have a certain level of difficulty because it may not be too difficult (Hartati & Yogi, 2019; Vincent & Shanmugam, 2020). Question items that are too easy or too difficult will result in a distribution of scores, so it is difficult to identify the reliability of achievement between students who perform well and students who perform poorly. The level of difficulty shows how easy or difficult the test questions are for that group. So the level of difficulty is influenced by the student's competence.

One of the most popular models used for polytomous items, common with many psychological tests, is the GRM (Graded Response Model). GRM is suitable for use when dealing with categories ordered on a rating scale (for example, a Likert scale that reflects the degree of agreement or disagreement) and is considered a generalization of the two-parameter logistic (2PL) model. This 2PL model is used to assign the probability that a person should receive a certain score (or higher), given the level of the underlying latent trait. The more attributes (positive affect, for example) the respondent has, the greater the probability they will respond with a higher-scoring answer or the greater the probability they will choose one of the more positive ratings on the scale item (Hambleton et al., 1992; Zanon et al., 2016). The GRM model is used for categorical responses and is a development of the 2-PL IRT (Logistic Parameter) model that can express two parameters of an item, namely the level of difficulty and the discriminating power of the item.

METHOD

This research is ex post facto quantitative research. The data used is the result of a statistical literacy test administered to 150 students. Student test response in the form of polytomous data (0, 1, 2, 3, and 4), which shows the measurement level of student test results The data was analyzed with the IRT technique of the GRM 2PL model with the help of PARSCALE. Before the data are analyzed, the IRT assumptions must be fulfilled. The IRT assumptions are the Unidimensional assumption test, local independence assumption test, and invariance assumption test. After the assumptions are met, then the data will be analyzed in 2 parameters model, namely parameter a (item discriminant) and parameter b (item difficulty). In this study, an analysis will also be conducted to find out how the ability of statistical literacy test items can measure the level of statistical literacy of students.

RESULT AND DISCUSSION

The unidimensional test is carried out through factor analysis using the SPSS program. Factor analysis requires that the data matrix mustt have sufficient correlation so that factor analysis can be carried out. To test whether there are correlations between dimensions, the Bartlett test of sphericity is used. If the result is significant it means that the correlation matrix has a significant correlation with several dimensions. Another test used

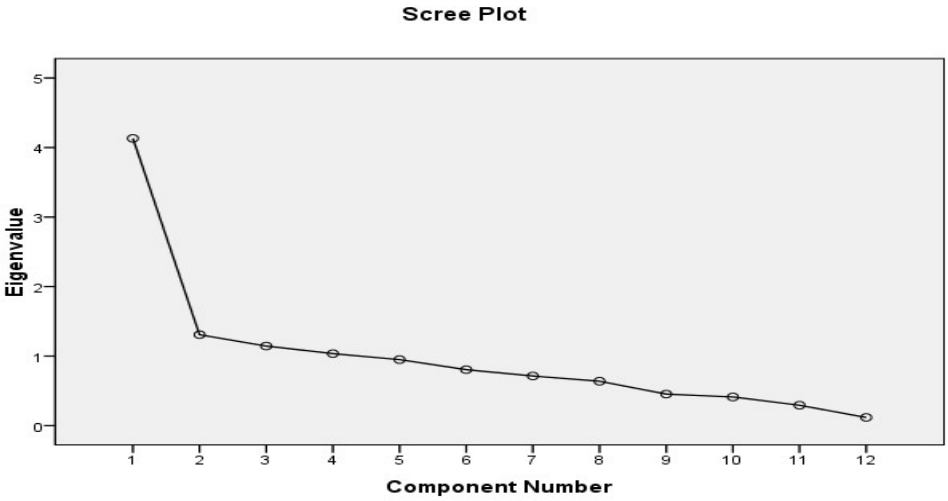
to see the intercorrelation between variables and whether or not factor analysis is performed is the measure of sampling adequacy (MSA). If the MSA value is 5 0.05, factor analysis can be carried laut. The results of the MSA test and Bartlett's test are shown in Table 1.

Table 1: Bartlett Test Results of Sphericity and KMO-MSA

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.770
Bartlett's Test of Sphericity	Approx. Chi-Square	706.801
	df	66
	Sig.	.000

Source: Result of Research Data Process, 2022

The Bartlett test of sphericity and KMO-MSA test table shows that the results are significant. It shows that the significant value is 0,000 which means that there is a correlation between dimensions. Besides, the MSA value of the analyzed data ≥ 0.05 so that factor analysis can be continued. The results of the factor analysis show that there is only 1 dominant factor measured by the statistical literacy test.



Source: Result of Research Data Process, 2022

Figure 1: The plot of Eigenvalues For Unidimensional Assumption

From Figure 1 it appears that factor 1 has a value of 4.231 followed by factor 2 with a value of 1.308 and other factors below it. This means that factor 1 is very dominant over other factors, while other factors have adjacent or similar eigenvalues. When the first-factor eigenvalue is several times the value of the second-factor eigenvalue, while the second-factor eigenvalue and so on are almost the same, it can be said that the unidimensional condition has been fulfilled (Aristiawan et al., 2019; Jumailiyah, 2017; Rahim & Haryanto, 2021)

Local independence requirements are met if the value of covariance between the participant's ability intervals located in the diagonals is small and close to zero (Hambleton et al., 1992; Hambleton & Swaminathan, 1985). The results of the local independence test of the statistical literacy test items, showed all elements outside the main diagonal matrix have values close to zero. This condition explains that the assumption of local independence has been fulfilled. Local independence can also be detected through unidimensional assumption

Test (Jumailiyah, 2017; Sudaryono, 2011). So that the assumption of local independence will automatically be fulfilled after the unidimensional assumption is fully fulfilled.

The test on the assumptions of the invariance of the ability parameters is done by separating between groups of students who have high abilities and low abilities. Each group as much as 27% of the total test-takers. Then the answers of the two groups of students were estimated. The results of which were in the form of test item parameters (slope, item difficulty, and guessing) from the two groups of students. Furthermore, the item parameter pairs from both the high and low ability groups are correlated. If it turns out the correlation is high, then the assumption of test item invariance is met. Results of invariance analysis indicate that all data are invariant. Invariance testing shows that the correlations between the calibration results in both groups were all high (0.753 - 0.916).

Item response theory contains two parameters, item parameters and participant parameters. The participant characteristic parameter θ states the participant characteristic with ability θ, while the item parameter is expressed through a suitable logistic model. In this study, 2 parameters were used with the Graded Partial Model (GRM) calibration. Therefore, item parameter estimates are expressed in terms of item discriminant (a), item difficulty index (b). The estimation results of these parameters can be seen in the phase 2 PARSCALE program output.

The results of the estimation parameters of the statistical literacy test items in the form of different grain power can be seen in the value of the slope contained in the PARSCALE PH2 output. Based on the results of data analysis obtained information on the parameter *a* (slope) all items are in the range of values from 0.191 to 1.598. The index of slope classifications can be seen in Table 2.

Table 2: The Results of Slope (parameter *a*) of Statistic Literacy Item

Item	Slope	Criteria
1	0,915	Sufficient
2	1,541	High
3	0,828	Sufficient
4	0,811	Sufficient
5	1,598	High
6	1,003	Sufficient
7	1,269	Sufficient
8	1,012	Sufficient
9	1,580	High
10	0,796	Sufficient
11	0,203	Low
12	0,191	Low

Source: Result of Research Data Process, 2022

From Table 2, information is obtained that from 12 items of students' statistical literacy tests, there are 7 items in the "enough" category to determine students' abilities, 3 items in the "high" category to determine students' abilities, and 2 items in the "low" category

to determine student ability. Item 5 is the item with the highest difference power index. So it can be concluded, item 5 is the best item to determine student ability in statistical literacy.

Items that are considered good for determining student ability are items that have a minimum category of "sufficient". Then there are 84% items from 12 items of statistical literacy test students can determine students on the ability of statistical literacy. So it can be concluded that the statistical literacy test used can distinguish students who have high literacy and students who have low statistical literacy.

The item difficulty of items is a function of students' abilities (Djemari Mardapi, 1991: 11). High-ability students will easily work on test questions, while low-ability students will find it difficult to work on test questions. The level of difficulty of the item moves from the scale $-\infty \leq b \leq \infty$ on the item response theory. But in practice, the items announced are items that have an agreed level of conformity (b) $-2 \leq b \leq +2$. Items with item difficulty near or below the scale -2 Select a test item in the easy category. Whereas items that have a level of conformity (b) close to or located above the +2.00 scale indicate that the items are in the difficult category (Hambleton, Swaminathan, & Rogers, 1991: 13). The provision for item response theory with polytomous response is a value of b> 2, then items are categorized as "very difficult", at interval -1 <b ≤ 2, items are categorized as "difficult", -1 <b ≤ 1 item is categorized as "Medium" and for b <-2 items are categorized as "very easy" (Hidayatullah, 2013). The results of the estimation of the statistical literacy test item parameters in the form of parameter b (Item difficulty) can also be seen in the value of the slope contained in the PARSCALE PH2 output. Based on the results of data analysis obtained information index parameter b (Item difficulty) all items are in the range of values -1,701 to 1,725. Classification of different power index items can be seen in Table 3.

Table 3: The Result of Item Difficulty (parameter *b*) of Statistics Literacy Items

Item	Item Difficulty	Criteria
1	1,725	Difficult
2	0,828	Difficult
3	-0,968	Easy
4	-0,930	Easy
5	0,794	Difficult
6	-1,701	Easy
7	1,078	Difficult
8	-0,453	Easy
9	0,904	Difficult
10	-0,862	Easy
11	0,293	Difficult
12	0,516	Difficult

Source: Result of Research Data Process, 2022

Based on Table 3. Information is obtained that 7 items have an item difficulty with the category of "difficult" and 5 items have an "easy" category. Based on the analysis of the estimated parameters a (Slope) and parameter b (item difficulty), it can be determined how many good items meet the IRT criteria. The quality of items empirically uses criteria: good, quite good, and not good. The criteria for each parameter can follow the criteria; a) An item

is classified as good if the item has a slope index of 0.0 to 2.0 and Item difficulty index of -2.0 to 2.0; b) The items are classified as quiet good if The slope index is more than 2.0 and the item difficulty index is less than -2.0 or more than 2.0; and 3) An item is not good if it does not fall into the two previous categories.

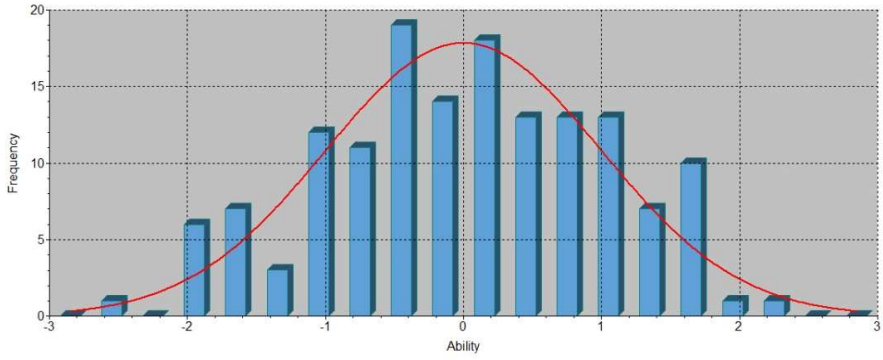
Besides slope, and item difficulty. Good items will also be selected based on the parameter model used is fit or not. Based on the analysis using GRM 2PL, all items meet the criteria, fit model. That is fit test items using IRT 2-PL analysis. The results of the recapitulation of items based on slope, item difficulty and fit model can be seen in Table 4.

Tabel 4: Results of Item Parameter Estimation of 2-PL

Item	Slope	Item Difficulty	Category
1	0,915	1,725	Good
2	1,541	0,828	Good
3	0,828	-0,968	Good
4	0,811	-0,930	Good
5	1,598	0,794	Good
6	1,003	-1,701	Good
7	1,269	1,078	Good
8	1,012	-0,453	Good
9	1,580	0,904	Good
10	0,796	-0,862	Good
11	0,203	0,293	Good
12	0,191	0,516	Good

Source: Result of Research Data Process, 2022

The participant characteristic parameter θ states the participant trait with ability. The estimated ability of participants can be seen in the PARSCALE PH3 output. Based on the PARSCALE PH3 output obtained information that the average ability of students is 1,0042. The average positive ability indicates that most students tend to have good statistical literacy skills. As seen in the following ability graph.

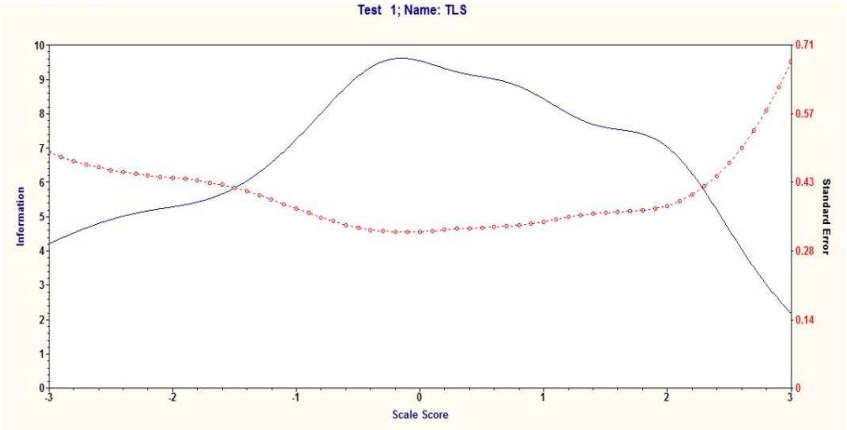


Source: Result of Research Data Process, 2022

Figure 2: Estimated of Student Statistical Literacy Ability Graph

Based on the graph we can seen that the proportion of students who have high abilities is greater than those who have low abilities. The Total Information Curve (TIC) for GRM produces tests that are accurate enough to assess participants with abilities of -1.3 to

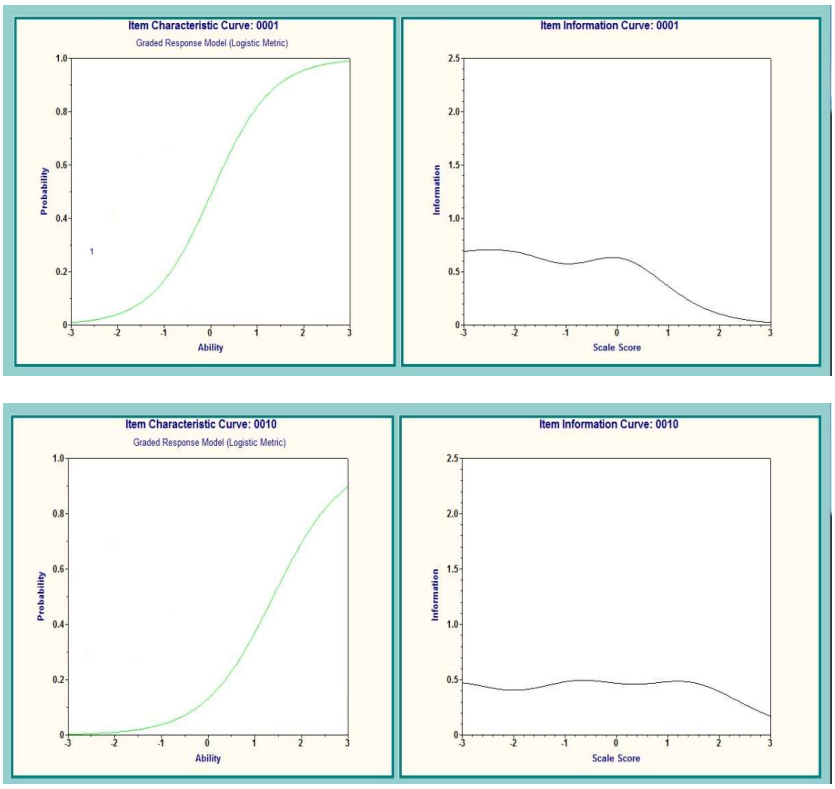
2.2. The test gives the highest information on a theta value of around -0.2. More can be seen in Figure 3.



Source: Result of Research Data Process, 2022

Figure 3: TIC of Student Statistical Literacy Tests with GRM

Unbroken lines indicate the value of information and broken lines show the magnitude of standard errors. Explanation about ICC graph or grain characteristic curve, item information function, test information function Ana standard error measurement (SEM) in the two-parameter logistics model (2PL). *ICC and Item Function 2-PL Model.* The following is an example of an ICC chart and item information function in the analysis Item of Student Statistical Literacy Test using the 2PL model.

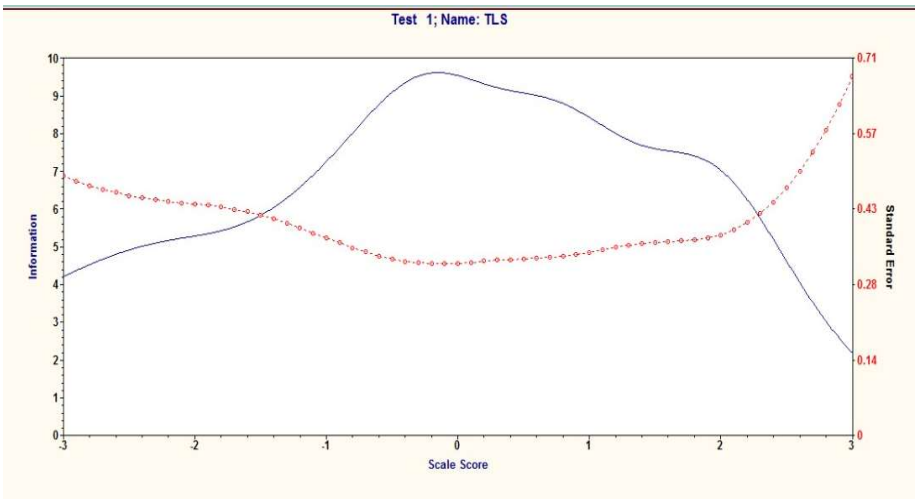


Source: Result of Research Data Process, 2022

Figure 4: ICC and 2PL Item Function Information Model

The 2PL model only involves two-item parameters namely slope (parameter *a*) and item difficulty (parameter *b*). Based on Figure 4 above, For example, the value in item 1 is known to be 0.916 for *the parameter a* and 1,725 for *parameter b* and item 10 that is 0.796 fo *parameter a* and -0,862 for *parameter b*. Based on this it is known that the slope values (*parameter a*) for item 2 are more lower than item 10 so that the ICC in item 2 will appear to be slower than that point 10. Value can be interpreted that a minimum capability of 1,725is required on the scale to answer correctly item 2 with a 50% chance. Information function item 2 will be maximized when on a scale of 1.725. Similar to item 2, item 10 in value it can be interpreted that a minimum capability of -0,862 is required on the sale to answer correctly point 10 with a 50% chance.

Graph of the relationship between the test information function with SEM on the 2PL model Approved by Figure 3 below.



Source: Result of Research Data Process, 2022

Figure 5: Test Information Function with SEM Model 2PL

Based on figure 5 above, it is known that a statistical literacy test item is being developer able to provide good information on the ability of statistical literacy capable students -1,5 to +2,4. The smallest SEM occurs when the test device delivers maximum test information function.

CONCLUSION

This study produced 12 items to measure students' statistical literacy. The characteristics of each item have a good different power values, 0.191 - 1.598 and grain difficulty values at intervals of -2 to 2. Based on the information generated, the model GRM scoring is suitable to model the Scoring Test Statistic Literacy Students are administrated. IRT is very useful in item analysis a test. Through IRT, the item parameter index can be known with ease. The index is the basis for selecting items. Other than that, the information

function can provide consideration of how the test should be used. quality literacy items that have been analyzed for quality in this study can be used to measure different samples. The 2PL GRM IRT analysis can be a reference for further research to analyze the quality of items with a political score.

BIBLIOGRAPHY

Akinde, O. A., Harr, D., & Burger, P. (2017). Field Experience: Experiential Learning as Complementary to the Conceptual Learning for International Students in a Graduate Teacher Education Program. *International Journal of Higher Education*, *6*(4), 137-143. https://doi.org/10.5430/ijhe.v6n4p137

Amelia, R. N., & Kriswantoro. (2017). Implementasi item response theory sebagai basis analisis kualitas butir soal dan kemampuan kimia siswa kota yogyakarta. *Jurnal Kimia Dan Pendidikan Kimia*, *2*(1), 1–12.

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. https://doi.org/10.17977/um047v25i12018p040

Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination. *European Scientific Journal, ESJ*, *12*(28), 263. https://doi.org/10.19044/esj.2016.v12n28p263

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory, Holt, Rinehart & Winston.1986*. ERIC.

Embretson, S. E., & Reise, S. P. (2013). Item Response Theory For Psychologists. In *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410605269

Erguven, M. (2013). Two approaches in psychometric process: Classical test theory & item response theory. *Journal of Education*, *2*(2), 23–30.

Hambleton, R. K., & Swaminathan, H. (1985). Assumptions ofItem Response Theory. In *Item Response Theory* (pp. 15–16). Springer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1992). Fundamentals of Item Response Theory. In *Contemporary Sociology*, *21*(2), 101-117. https://doi.org/10.2307/2075521

Hartati, N., & Yogi, H. P. S. (2019). Item Analysis for a Better Quality Test. *English Language in Focus (ELIF)*, *2*(1), 59-70. https://doi.org/10.24853

Hattie, J. (1985). Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement*, *9*(2), 139–164. https://doi.org/10.1177/014662168500900204

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Ida, I. (2021). Evaluation Of The Jumat Taqwa Program (Jumtaq) To Constructing The Religious Character Of Students. *Pedagogik: Jurnal Pendidikan*, *8*(2), 368–386. https://doi.org/10.33650/pjp.v8i2.2967

Isnani, I., Utami, W. B., Susongko, P., & Lestiani, H. T. (2019). Estimation of college students' ability on real analysis course using Rasch model. *Research and Evaluation in Education*, *5*(2), 95–102. https://doi.org/10.21831/reid.v5i2.20924

Jumailiyah, M. (2017). Item response theory: A basic concept. *Educational Research and Reviews*, *12*(5), 258–266. https://doi.org/10.5897/err2017.3147

Leal Filho, W., Raath, S., Lazzarini, B., Vargas, V. R., de Souza, L., Anholon, R., Quelhas, O. L. G., Haddad, R., Klavins, M., & Orlovic, V. L. (2018). The role of transformation in learning and education for sustainability. *Journal of Cleaner Production*, *199*(2), 286–295. https://doi.org/10.1016/j.jclepro.2018.07.017

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). The hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*(2), 340–364. https://doi.org/10.1080/00273171.2011.564527

Rahim, A., & Haryanto, H. (2021). Implementation of Item Response Theory (IRT) Rasch Model in Quality Analysis of Final Exam Tests in Mathematics. *Journal of Educational Research and Evaluation*, *10*(2), 57–65. https://doi.org/10.15294/jere.v10i2.51802

Sudaryono. (2011). Implementasi Teori Responsi Butir (Item Response Theory) Pada Penilaian Hasil Belajar Akhir di Sekolah. *Jurnal Pendidikan Dan Kebudayaan*, *17*(6), 719–732. https://doi.org/10.24832/jpnk.v17i6.62

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

Suwarto, S. (2022). Karakteristik Tes Ilmu Pengetahuan Alam. *Jurnal Pendidikan*, *31*(1), 109-121. https://doi.org/10.32585/jp.v31i1.2269

Vincent, W., & Shanmugam, S. K. S. (2020). The Role of Classical Test Theory to Determine the Quality of Classroom Teaching Test Items. *Pedagogia : Jurnal Pendidikan*, *9*(1), 5–34. https://doi.org/10.21070/pedagogia.v9i1.123

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexao e Critica*, *29*(1), 73-86. https://doi.org/10.1186/s41155-016-0040-x