Al-Jabar: Jurnal Pendidikan Matematika
Volume 13, Number 1, 2022, Pages 109 – 122
http://ejournal.radenintan.ac.id/index.php/al-jabar/index

An analysis of item response theory using program R

Ali[1]*, Edi Istiyono[1]
[1] Study Program of Educational Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta, Indonesia
✉ stevenalix959.2017@student.uny.ac.id

Artikel Information
Submitted Jan 21, 2022
Revised May 28, 2022
Accepted June 01, 2022

Keywords
Item Analysis;
Item Response Theory;
R Package Program.

Abstract
The test is one of the instruments used to assess the extent of student understanding in learning. Multiple choice is a type of test commonly used in testing students. In addition to testing students' understanding, the quality of the tests used also needs to be tested. This study aims to determine the characteristics of the national mathematics test items in Baubau in the 2015/2016 academic year and the test information function with the item response theory approach. This research is an ex-post-facto study with a sample size of 574 students using a random sampling technique. Data was collected through documentation and analyzed using the LTM R package program. Findings indicated that there were four items (I1, I2, I4, and I8) for the 1-PL model, six items (I1, I2, I4, I7, I8, and I10) for the 2-PL model, and seven items (I1, I2, I3, I4, I7, I9, and I10) (3-PL) that fit the model (FM). The percentage of good (G) item parameters using R was 90% for (b) (1-PL), 90% (b) and 100% (a) (2-PL), and 90% (b), 10% (a), and 70% (c) (3-PL). The percentage of good quality items in each model for the 1-PL model was 40% or four items, the 2-PL model was 60% or six items, and the 3-PL model was 0%, or none was included in the good quality item category.

INTRODUCTION

Educators could measure the extent of students' competencies and comprehension from the instructional practices employed by using a test. Giving tests is a learning process where learning is an essential factor in achieving learning objectives (Anggoro et al., 2019). The test is used as a technique or measurement tool (Kaplan & Saccuzzo, 2009; Cohen & Swerdlik, 2009), which is used as an "objective" and "standardized" measure of behavior samples (Anastasi, 1988). A good test needs to have good items in the questions provided. There are many kinds of items in a test: multiple-choice, true-false, open-ended, short answers, and descriptive & persuasive. To measure good quality items listed in the test, an analysis of items must be administered to ensure the quality of items is relevant to the standard and quality of constructive alignment in designing the test.

Murphy & Davidshofer (2005) claimed that item analysis is a structured statistical group. Urbina (2004) and Kaplan & Saccuzzo (2009) stated that item analysis is used to evaluate the quality of tests during the process of development and construction of a test. In this context, item analysis is shown as analyzing items in questions is a vital component among educators before the development of testing. Poor quality of items demands judgment in the decision to produce a good test. This includes collecting, summarizing, and using information obtained from students' responses (Nitko, 1996).

The purpose of analyzing the items is to obtain the quality of questions by reviewing the items before the questions, to assist identification of deficiencies in the test (Anastasi & Urbina,

How to cite Ali, A., & Istiyono, E. (2022). An analysis of item response theory using program R. *Al-Jabar: Pendidikan Matematika*, *13*(1), 109-122.
E-ISSN 2540-7562
Published by Mathematics Education Department, UIN Raden Intan Lampung

1997), and to confirm students' comprehension of the materials utilized in the instructional practices (Aiken, 1994). In addition, Murphy & Davidshofer (2005) stated that item analysis could help to improve the comprehension and the reasons for test scores that could predict multiple criteria, indicate the reliability of a test, and specify the improvement of test characteristics.

Items analysis is compulsory using the Classical Test Theory (CTT) and Item Response Theory (IRT) approach. It is evidenced by the number of researchers who employed these two approaches (Champlain, 2010; Holland & Hoskens, 2003; Hays et al., 2006; Linden & Hambleton, 1997). In CTT, scores are obtained based on the number of individual responses to various items (Kaplan & Saccuzzo, 2009). However, there could be a gap among the examinees who sit for the examination due to some non-conducive examination environment factors and excessive anxiety that intrude the participants to provide the correct answer. The levels of questions provided are high, as it leads to a reduction in the score obtained. To overcome this matter, the researcher employed the IRT approach.

One of the techniques for data analysis is to use item response and theoretical models. This technique is an update of classical test theory. The use of classical test theory is relatively easy but has some limitations for psychometric experts, such as estimating the ability of students to depend on items. Besides, the estimated measurement errors do not include each individual but together or in groups. Of course, this will be a problem in the learning process, especially in seeing the ability of individual examinees. Therefore, to overcome this matter, experts design new theories to complete and correct the limitations that exist in classical test theory. This theory is what we later recognized as the IRT.

IRT is a statistical model that uses responses to test items to estimate the level of examinees in the measured construct. In item response theory, some assumptions underlying the item response theory and the most commonly used are unidimensionality and local independence. Unidimensionality means measuring ability (θ) in a test for each examinee. Local independence means that when the abilities that affect test performance are maintained, the examinee's response to each item pair is statistically independent, which means that there is no relationship between the test participants' responses to different items (Hambleton & Swaminathan, 1985; Hambleton et al., 1991; Demars, 2010).

The most well-known IRT model is the logistic parameter model (PL), i.e., the 1-PL model or Rasch model, 2-PL model and 3-PL model (Hambleton et al., 1991; Demars, 2010; Fox, 2010; Reckase, 2009). These models contain an estimate of the latent nature of reading or depression, the ability to distinguish between individuals with different construct levels, and the possibility of chance or guessing. The construct in question is the latent variable measured on items formed based on indicators as variables observed in the factor analysis model (Muthen & Lehman, 1985; Thissen et al., 1993). According to Hays et al. (2006), IRT theoretically provides several advantages invariant items and latent traits that estimate standard errors and information underlying constructive anchoring estimates of item content and explicit evaluation of assumptions model.

The Rasch model is known as the 1-PL model (Linden & Hambleton, 1997; Baker, 2001; Demars, 2010; Fox, 2010; Reckase, 2009; Hambleton et al., 1991; Embretson & Reise, 1998), but what distinguishes it is that the Rasch model has a *discriminant* value set equal to 1 (Finch & French, 2015; Demars, 2010; de Gruijter & van der Kamp, 2008; Embretson & Reise, 1998).

The Rasch or 1-PL model is used with a sample of as small as 100 or 200 test participants (Demars, 2010). The 1-PL model is one of the most widely used models. If using the 1-PL model, the item used only tested the *level of difficulty*, in the 2-PL model that is focused only on the *level of difficulty* test and the *discriminant* of the item. But it requires a sample size generally estimated at 500 or less if the item has a moderate difficulty level and normal distribution ability (Drasgow, 1989; Harwell & Janosky, 1991; Stone, 1992). The last is the 3-PL model, where this model tests the parameters of *difficulty level*, *discriminant*, and *guessing* items. To estimate the *guessing* parameter requires a larger sample than the previous model.

Furthermore, a test that is made to have the quality of the items using the IRT model can be seen from the parameters previously mentioned, namely the discriminant (a), the difficulty level of item (b) and guessing (c). Examinees have different abilities of high and low abilities. With the discriminant parameters, the ability of the examinee is distinguished. It means that the discriminant parameter is the ability of an item distinguished by the examinees who have mastered the material and those who have not mastered it. Suppose the item is not available to distinguish the examinee's ability. In that case, the answer key is incorrect, has more than one answer key, the measured competency is unclear, and the deceiver does not work. The usual range for item discriminant parameter (a) is (0, 2) (Hambleton & Swaminathan, 1985; Hambleton et al., 1991). The higher the discriminant of a question, the better the item will be.

The next parameter is the level of item difficulty, which is an opportunity to answer correctly on a problem at a certain level of ability. The item score produced by the answers of some test participants measures the difficulty index item (b). The more test participants were able to answer the test questions given, the lower the level of difficulty of the test and vice versa. A good question item lies in the interval $-2 \leq \theta \leq 2$ (Hambleton et al., 1991). The value of b approaches -2 indicates that the item is getting easier, and the value of b approaches +2 indicates that the item is getting harder. The level of difficulty of the item has usefulness for the educator and testing and teaching (Nitko, 1996). Usefulness for educators is re-learning and giving suggestions to students about the learning outcomes and preventing biased items. The usefulness for testing and teaching is to make a test with the data accuracy on the problem and to know the weaknesses and advantages of the school curriculum and the presence of biased items. However, it is different from guessing parameters.

According to Baker (2001), guessing parameters is an opportunity to answer an item that is correct by guessing it yourself. Hambleton & Swaminathan (1985) state that guessing (parameter c) shows the opportunity for low-ability test participants to be able to answer the item correctly. In a multiple-choice test consisting of a choice of alternative answers, the parameter *c* is located around 1/*k*, where *k* is the number of alternative answers.

Another thing about item analysis is model fit data. Model fit data can be investigated at the item or person level. In especially the Item-fit model, items are said not to be fit with the model if the probability value (significance) $<\alpha$ with $\alpha = 0.05$ (Retnawati, 2014). Several studies of fit data for items can be seen in research conducted by Thissen & Steinberg (1988), Meijer et al., (1990) and Reise & Waller (1993). Fit models that can be used are 1-PL or Rasch, 2-PL, and 3-PL models.

The final element important in item analysis is Item Characteristic Curve (ICC) and Item Information Characteristic (IIC). ICC describes the opportunity relationship to answer correctly with the examinee's level of ability. In addition, it can be seen which items are the easiest and

most difficult on a test. Each item has an information function. The number is an information function of the test (Hambleton et al., 1991). The function of the test package information will be high if the constituent items have a high information function. The information function can be seen from the IIC graph. The information function obtained can be a test and item information function.

Computer programs are available to teach the complex IRT theory (Penfield, 2003). With computers, participants who have difficulties can be identified (Bolt, 2003; Schmidt & Embretson, 2003). Exist many effective computer programs for applying response theory item models nowadays, such as BILOG-MG, TESTFACT, MULTILOG, and PARSCALE (Baker, 2001). However, one computer program also has the same purpose, namely the R program. R program is open-source software that everyone can write functions and add to the software. You can access the information at http://cran.r-project.org/ for more detail. R programs are available on various computing platforms, most notably Windows, Macintosh OS, and Unix/Linux. The R program offers researchers to perform data analysis from the most basic to the most complex. Thus, the R program is highly recommended for data analysis for statisticians and researchers in other fields who use statistics to inform their work. R Program has a variety of functions that are applied for the models of IRT and the application of formulas used in mathematics. Of course, this program includes a complex program because it requires a command to run statistically with what a researcher's plan analyzes. Therefore, many researchers use this program such as (Chalmers, 2012; Dahlke & Wiernik, 2019; Chan, 2018; Ferraro & Giordani, 2015; Lemenkova, 2019; Chen et al., 2020; Lemenkova, 2018; Kruschke, 2014; and Ostrouchov et al., 2012). It becomes an interest for researchers to apply it in this study.

Several studies have been conducted related to item response theory analysis, such as item response theory analysis, especially the Rasch model using the QUEST program (Rizbudiani et al., 2021), item response theory analysis using IRTPROV3.0 and BILOG-MG V3.0 programs to investigate item level diagnostic statistics and models - data fit (Essen et al., 2017), and item response theory analysis by comparing the fit of the 2-PL and 3-PL models (Reise & Waller, 2003). Research on the R program, such as using the R program, in particular, developing the R PLmixed package into the existing R package lme4 (Jeon & Rockwood, 2017), and using the R program to see the level of difficulty and suitability of the item model as well as the item characteristic curve (ICC) and item information curve (IIC) on the Rasch model (Muchlisin et al., 2019). Based on several existing studies, no research has been found regarding the analysis of item response theory in the 1-PL, 2PL, and 3-PL models using the R program to see the item difficulty level, discriminant, guessing, model fit, item characteristic curve (ICC) and item information curve (IIC). Based on the description above, this study aims to determine the characteristics of the items of the national exam mathematics in Baubau city in the 2015/2016 academic year and the information function test questions with a question response theory approach. In this study, the 1-PL, 2-PL, and 3-PL models were used to analyze the data with the help of the R program.

METHODS

The researchers employed an ex-post facto design. The designation ex-post-facto, derived from Latin for "after the fact" (Ary et al., 2010), indicates that research is carried out after something

has happened. The population in this study was 3,079 students who took the National Examination Mathematics in the city of Baubau in the 2015/2016 academic year with five code packages. Samples were taken using random sampling techniques so that the package code was P0C5520, and 574 students were sufficient to represent the population.

This study focuses on analyzing item characteristics and test information functions using an item response theory approach with the help of the R ltm package program (latent trait model), analyzing 40 items on the multiple-choice test of the Junior High School Mathematics National Examination in the 2015/2016 academic year in Baubau city and taken ten random items to be carried out as the focus of this study.

This research was conducted through three stages: the preparation, data collection, and data analysis stages. The analysis was carried out on each logistic parameter model using the R program at the data analysis stage. The results of the analysis of the characteristics of the items are seen from the item fit model, the percentage of good item parameters, item quality, ICC, IIC and TIF with the following details, namely: the number of item fit models (number of items fit in model 1-PL, 2-PL, and 3-PL), percentage of good item parameters (percentage of values *a*, *b* and *c* are included in the good category in each logistic parameter model), percentage of Item quality (can be seen from the "good" or "not good" category of an item), ICC (the curve that can show which items are the most difficult and easiest, and describe the opportunity relationship to answer correctly with the level of ability of examinee), IIC (the curve that can provide information on an item), and TIF (the curve that provides information on a test).

In the analysis stage, the first researcher enters the data in the form of responses to student answers. Answer responses are divided into 2, namely response answers that contain alternative answers, namely A, B, C, and D and response answers that contain the numbers 0 and 1 (dichotomous data). The R program's utilization with the ltm package (latent trait model) uses dichotomous data in the form of numbers 0 and 1. Dichotomous data are used because they often involve item response theory analysis (Finch & French, 2015; Hambleton & Swaminathan, 2013; Van der Linden, 2017; Steinberg & Thissen, 2013; Paek & Cole, 2019; Mair, 2018; Ayala, 2018; DeMars, 2018; Primi et al., 2016; Reise, 2014). The program is used to analyze each model of the logistics parameter. Then the results of the analysis are identified in the form of item parameter values *a*, *b* and *c*. A category of item quality will be created from the value of the item parameters. The quality of items is considered good if the item parameters include good categories and fit items on each model. Otherwise, the quality of the item is considered not good if the item parameters are not good and items are not fit for each model. It can be seen in Table 1 below.

Table 1. Item Quality Categories in Logistic Parameter Model

Category	Parameter Logistic Model		
	1-PL	2-PL	3-PL
Good (G)	$-2 \le \theta \le 2$ fit on the model (FM)	$-2 \le \theta \le 2$ $0 \le a \le 2$ fit on the model (FM)	$-2 \le \theta \le 2$ $0 \le a \le 2$ $c \le 0.25$ fit on the model (FM)
Not Good (NG)	$\theta > 2$ or $\theta < 2$ not fit on the model (NFM)	$\theta > 2$ or $\theta < 2$ $a > 2$ or $a < 0$ not fit on the model (NFM)	$\theta > 2$ or $\theta < 2$ $a > 2$ or $a < 0$ $c > 0.25$ not fit on the model (NFM)

The analysis of items (*I*) in the 1-PL, 2-PL, and 3-PL model using the R Program will obtain the quality item (*I*) based on the categories in table 1. Further, in the ICC analysis, an item curve is displayed. From the curve, the easiest and most difficult items will be identified. The last thing to do is analyze the IIC and TIF. IIC aims to provide information on an item to test-takers ability, and TIF aims to provide information on the overall test.

RESULTS AND DISCUSSION

Table 2 shows the analysis of the results of 10 items in the 1-PL model using the R Program. The analysis results indicated the value of the difficulty levels of each item, the probability of each item and the category of difficulty level and probability.

Table 2. Results Analysis of 10 Items (*I*) in the 1-PL model using R

I	Difficulty Level		Chi-Square	
	b	Category	Probability	Category
*I*1	0.124	G	.352	FM
*I*2	1.577	G	.861	FM
*I*3	1.870	G	.000	NFM
*I*4	1.124	G	.057	FM
*I*5	0.833	G	.013	NFM
*I*6	0.617	G	.001	NFM
*I*7	2.261	NG	.003	NFM
*I*8	0.270	G	.375	FM
*I*9	2.051	G	.003	NFM
*I*10	1.416	G	.003	NFM

Based on table 2, the estimated parameter b_i from all ten items has almost a good *level of difficulty* (*b*) except *I*7. It can be seen from the results of the b_i parameter estimation. However, the b_i item parameter value for I7 is not included in the range because the bi parameter value of $b_i > 2$, so only I7 is classified as not good (NG). The number of fit items obtained is four items, namely *I*1, *I*2, *I*4, and *I*8. This is due to the chi-square's significance value (*Probability*) being above the 0.05 significance level. Four quality items are included in the G category, namely *I*1, *I*2, *I*4, and *I*8. It is because all four items have item parameters categorized as G and item fit on the model.

Next, the 2-PL model was analyzed using the R Program. Table 3 presents the analysis of the results of 10 items in the 2-PL model using the R program. The analysis results show the value of the *level difficulty, discriminant,* and *probability* of each item and the category of *difficulty level, discriminant,* and *probability*.

Table 3. Results Analysis of 10 Items (I) in the 2-PL model using R

I	Discriminant		Difficulty Level		Chi-Square	
	A	Category	*B*	Category	Probability	Category
*I*1	0.599	G	0.129	G	.031	NFM
*I*2	0.463	G	2.065	NG	.173	FM
*I*3	1.128	G	1.143	G	.020	NFM
*I*4	0.679	G	1.039	G	.158	FM
*I*5	0.816	G	0.649	G	.122	FM
*I*6	1.059	G	0.371	G	.024	NFM
*I*7	0.905	G	1.652	G	.210	FM
*I*8	0.483	G	0.351	G	.421	FM
*I*9	0.059	G	20.254	NG	.037	NFM
*I*10	0.928	G	1.002	G	.174	FM

Table 3 provides information about the values of *discriminant* (*a*) and *difficulty level* (*b*) and the *probability* of ten items. All items have *discriminant* values in the range of 0 to 2, so all items are in a good category. It can be seen that the parameter value *b* is almost all in the range $-2 \leq \theta \leq 2$. Therefore, the estimation for parameter *b* shows that only *I*2and *I*9 are not a good category, and the others are in a good category. The *probability* of 10 items shows that only *I*1, *I*3, *I*6, and *I*9 are NFM because the probability value is less than 0.05. From table 3, it can be concluded that the quality of items including category G is *I*4, *I*5, *I*7, *I*8, and *I*10. This is because the six items have discriminant parameters and the level of difficulty of the items, including category G and items that fit in the model.

The last model is the 3-PL model presented in table 4 using the R program. The 3-PL model provides information about values of *discriminant* (*a*), *difficulty levels* (*b*), *guessing* (c), and *probability*.

Table 4. Results Analysis of 10 Items (I) in the 3-PL model using the R Program

I	**Discriminant *a* (Category)**	**Difficulty Level *b* (Category)**	**Guessing *c* (Category)**	**Chi-Square Probability (Category)**
*I*1	6.396 (NG)	1.268 (G)	0.428 (NG)	.058 (FM)
*I*2	4.026 (NG)	1.601 (G)	0.244 (G)	.499 (FM)
*I*3	6.795(NG)	1.173 (G)	0.166 (G)	.961 (FM)
*I*4	9.791 (NG)	1.352 (G)	0.289 (NG)	.253 (FM)
*I*5	1.327 (G)	0.724 (G)	0.105 (G)	.000 (NFM)
*I*6	2.432 (NG)	0.788 (G)	0.227 (G)	.001 (NFM)
*I*7	4.346 (NG)	1.399 (G)	0.147 (G)	.165 (FM)
*I*8	2.258 (NG)	1.444 (G)	0.392 (NG)	.022 (NFM)
*I*9	6.942 (NG)	2.392 (NG)	0.228 (G)	.322 (FM)
*I*10	2.274 (NG)	1.101 (G)	0.164 (G)	.161 (FM)

Based on table 4, it was found that the parameter value *a*, only at *I*5, is included in the G category while the others are in the NG category. *I*5 is categorized as G because the parameter value is in the range of 0 to 2. Likewise, the parameters *b* included in the NG category were *I*9, while the others included category G.

For parameter *c*, three items are included in the NG category, namely *I*1, *I*4, and *I*8. These three items have a parameter value of $c > 0.25$, so they are categorized as NG categories. The probability value indicates five items categorized as FM, namely *I*1, *I*3, *I*4, *I*9, and *I*10. These items have a probability value > 0.05, so it belongs to the FM category. The item quality is said to be good if the parameters in the 3-PL model are categorized as good and the items fit the model. Based on these things, it can be concluded that none of the quality items are included in the G category of 10 items. The analysis results using the item response theory approach with the help of the *ltm* R package program above show that none of the logistical parameter models produced ten items, including the FM category of the ten items analyzed. In addition, the quality of items obtained from 10 items for each logistic parameter model is not yet fully good.

Table 5. The Number of Item Fit and Percentage Quality of Item G in Each Logistic Parameter Model

Item	**Analysis Using R Program**		
	1-PL	**2-PL**	**3-PL**
Numbers of items FM	4	6	7
Percentage of quality item G	40%	60%	–

Table 5 shows that the item analysis of 10 items with all three logistical models using the R Program can be concluded that the problem of Junior High School Mathematics National Examination in 2015/2016 in Baubau city is very suitable to be analyzed using a two-parameter model (2-PL). By using the two-parameter model, many items received or, in other words, are classified as good items, as many as six items (60%) for analysis of the R program. In addition, the highest number of items included in the FM category affected the suitability of the analysis on the model.

Considering table 5 above, the highest number of items included in the FM category is obtained from the 3-PL model of 7 items. However, in the absence of a single item with good quality items on the 3-PL model, it cannot be used as a suitable model for analyzing the 2015/2016 academic year of National Mathematics Examination questions in Baubau. Considering these two things, namely a large number of fit items and the highest percentage of item quality belonging to the G category, the 2-PL model is suitable for analysis.

Besides the number of FM items and the percentage of quality of item G, the percentage of item parameters belongs to the G category. The following is the percentage distribution of the parameters of category G items for ten items based on analysis using the R program.

Table 6. The Percentage Distribution of Item Parameter Categories G in the analysis of 10 items using the R Program

Parameter Item	Analysis using R Program		
	Percentage of Parameter Item G		
	1-PL	2-PL	3-PL
B	90%	80%	90%
A	–	100%	10%
C	–	–	70%

Looking at table 6, it appears that the percentage of good item parameters for parameter *b* in the 1-PL model is 90% or as many as nine items, in the 2-PL model for parameter *b* is 90% or as many as nine items, and parameter *a* is 100% or as much as ten items. For the 3-PL model, parameter *b* of 90% or nine items is obtained, parameter *a* of 10% or 1 item and parameter *c* of 70% or seven items. The 3-PL model yields a low percentage, especially in parameter *a*. It shows that all items produced from the analysis of this model are of poor quality, as shown in table 5.

An item characteristic curve (ICC) and item information curve (IIC) curve were shown using the R program for further analysis. It was intended to find out more about the characteristics and information provided by a test or item. For example, the 1-PL model shows the ICC in Figure 1 below.

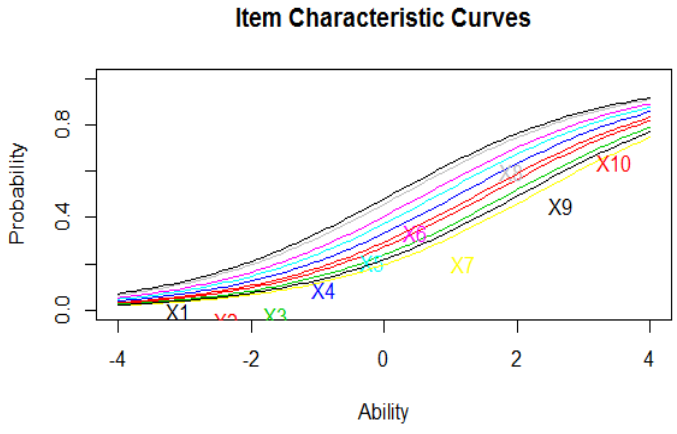

Figure 1. ICC for 1-PL Model Using the R Program

Figure 1 shows the characteristic curves of items in the 1-PL model for ten items. This curve displays the probability of answering correctly for examinees with certain abilities θ. The curves are close together, so it is difficult to determine which items are the most difficult and which ones are the easiest.

The R program was used to compare the items in one ability scale to see which items were the easiest and most difficult of the ten items on the ICC curve. For example, *I*2, *I*8, and *I*9 are taken for the 1-PL model, and ICC curves are made. The curve is presented in figure 2.

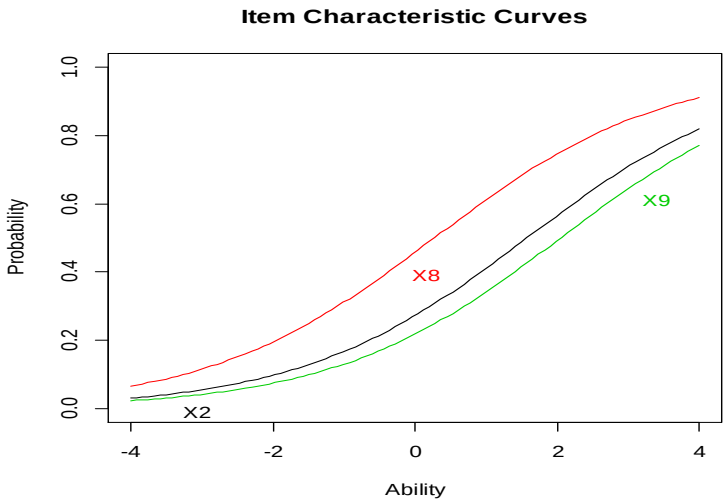

Figure 2. ICC for I2, I8, and I9 in 1-PL Model

Based on Figure 2, it can be seen that item nine (*I*9/X9) is located rightmost, while item eight (*I*8/X8) is located on the far left. The rightmost item shows that the item is the most difficult than the other items. Otherwise, the leftmost item indicates that the item is the easiest. So, *I9* is the most difficult item, and *I8* is the easiest item. It is also evident from *I9*, which has a value of parameter *b* that comes closest to +2 and *I8*, which has a value of parameter *b* closest to -2. From the ICC curve, it can be concluded that the greater the examinees' ability, the greater the probability of answering correctly. The greater the probability of answering correctly, the easier the item is.

The last analysis is IIC, which provides information on a test or item. For example, the IIC curve with the 3-PL model is shown in Figure 3. The image shown is unclear, and it is difficult to identify the information curve for each item.

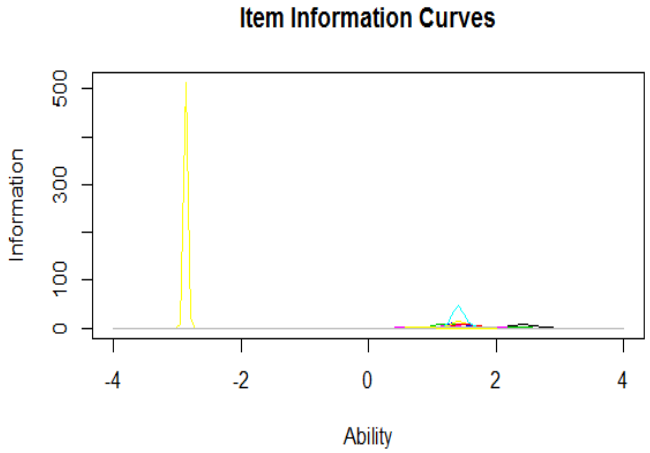

Figure 3. IIC for 3-PL

Compare more items to see which items are better at providing maximum information. For example, *I*2, *I*8, and *I*9 are taken to make the IIC curve. The following IIC curves for I2, I8, and I9 are in Figure 4 below.

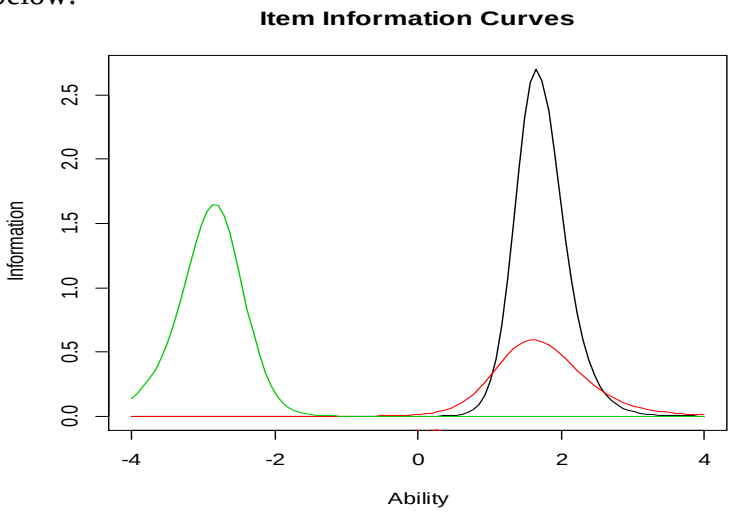

Figure 4. IIC for I2, I8, and I9 in 3-PL Model.

Figure 4 shows three items that provide maximum information. The highest item curve giving maximum information compared to other items is item two (*I*2), around 2.7, with θ ability approaching two. Besides the maximum information given by the item, the maximum information can also be seen on a test. The results of the test information analysis can be Shown using the R program. This is shown in figure 5.

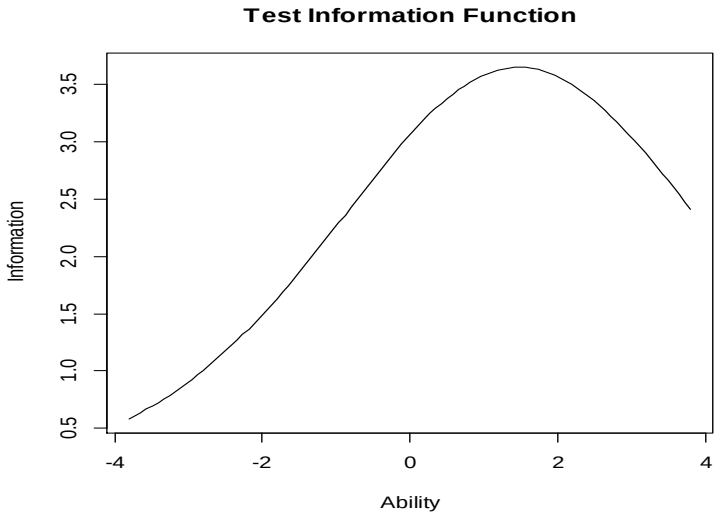

Figure 5. Information functions of Mathematics National Examination Test for Junior High School in 2015/2016 in Baubau City on R Program.

Call:
Rasch (data = data [, 2:41], constraint = NULL, IRT.param = TRUE)
Total Information = 24.76
Information in (-4, 4) = 19.38 (78.26%)
Based on all the items

Previous research conducted by Muchlisin et al. (2019) used the Rasch model with the assistance of the R program to analyze the level of difficulty of the items and the item suitability

model. in their research. They also displayed the item characteristic curve (ICC) and item information curve (IIC). While their research only used the Rasch model in this study using the 1-PL, 2-PL and 3-PL models, the difficulty level of the items was analyzed, but the discriminant and guessing items were also analyzed. In addition, research conducted by Kurniawan (2015) analyzed the quality of questions based on item response theory using the 2-PL model. The results of the analysis of the quality of the questions showed that of the 30 items analyzed with the item response theory of the two-parameter logistic model, 13 items (43.33%) were included in the good category of items. 15 items (50%) were included in the bad category. The BILOG program is used for analysis and only displays the results of the 2-PL model. In this study, the R program shows the results of the 1-PL, 2-PL and 3-PL models.

Based on the study results, it can be stated that the implication that the characteristics of the national mathematics exam items in Baubau City in 2015/2016 can be used as a consideration for the government to make better questions. In addition, the characteristics of the items based on the analysis results can be used as a reference for teachers who still lack knowledge of item analysis to make tests of good quality.

As for the limitations in the study, from the 40 items analyzed, only ten items were taken to be the focus of the research. For further research, the researcher suggests using other applications such as BILOG-MG. Item response theory can be applied in this application. The goal is to compare the results of the analysis between the R program and the BILOG-MG.

CONCLUSIONS

Based on the results of the study, it can be concluded that the analysis of items using the R program on the 1-PL model obtained four items, namely *I*1, *I*2, *I*4, and *I*8, have good item parameters and fit on the model, so that it includes a good item quality category. In the 2-PL model, six items are obtained, namely *I*4, *I*5, *I*7, *I*8, and *I*10, which have good parameters and are fitted on the model to include them in the good quality item category. The percentage of good (G) item parameters using R is 90% for (b) (1-PL), 90% (b) and 100% (a) (2-PL), and 90% (b), 10% (a), and 70% (c) (3-PL). The percentage of good quality items in each model for the 1-PL model is 40% or as many as four items. The 2-PL model was 60% or six items, the 3-PL model was 0%, and none included the good quality item category. Based on this, it can be said that the 2-PL model is a model that can be selected for analysis in the 2015/2016 National Mathematics Examination in the city of Baubau.

ACKNOWLEDGMENT

Researchers would like to thank the participants who contributed to fulfilling research data needs. No additional funding from outside other than researchers in this project.

AUTHOR CONTRIBUTIONS STATEMENT

EI is the coordinator and the author of articles in this research activity. As the author of the article on the revision and processing of instrument data.

REFERENCES

Aiken, L. R. (1994). *Psychological testing and assessment* (eight edit). Allyn and Bacon.

Anastasi, A. (1988). *Psychological testing* (6th Edition). Mcmillan.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (seventh ed). Prentice-Hall, Inc.

Anggoro, B. S., Agustina, S., Komala, R., Komarudin, K., Jermsittiparsert, K., & Widyastuti, W. (2019). An analysis of students' learning style, mathematical disposition, and mathematical anxiety toward metacognitive reconstruction in mathematics learning process abstract. *Al-Jabar: Jurnal Pendidikan Matematika*, *10*(2), 187–200.

Ary, D., Jacobs, L. C., Sorensen, C., & Razavieh, A. (2010). *Introduction to research in education* (8th editio). Nelson Education.

Ayala, R. J. D. (2018). Item Response Theory and Rasch Modeling. In *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (2nd ed.). Routledge.

Baker, F. B. (2001). The basics of item response theory. In *Evaluation* (Second Edi). ERIC.

Bolt, D. (2003). Essays on item response theory. A. Boomsma, MAJ van Duijn, and TAB Snijders (Eds.)[Book Review]. *Psychometrika*, *68*(1), 155-58.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, *44*(1), 109–117.

Chan, B. K. C. (2018). Data analysis using R programming. In *Biostatistics for Human Genetic Epidemiology* (pp. 47–122). Springer.

Chen, C., Razak, T. R., & Garibaldi, J. M. (2020). FuzzyR: An extended fuzzy logic toolbox for the R programming language. *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–8.

Cohen, C., & Swerdlik, S. (2009). *Psychology testing and assessment: An introduction to test and measurement* (seventh ed). McGraw Hill, Inc.

Dahlke, J. A., & Wiernik, B. M. (2019). Psychmeta: An R package for psychometric meta-analysis. *Applied Psychological Measurement*, *43*(5), 415–416.

de Gruijter, D. N. M., & van der Kamp, L. J. T. (2008). *Statistical test theory for the behavioral sciences*. Taylor & Francis Group, LLC.

DeMars, C. (2010). *Item response theory: Understandings statistics measurement*. Oxford University Press.

DeMars, C. E. (2018). Classical test theory and item response theory. *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*, 49–73.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, *13*(1), 77–90.

Embretson, S. E., & Reise, S. P. (1998). *Item response theory for psychologist*. Lawrence Erlbaum Associates, Inc.

Essen, C. B., Idaka, I. E., & Metibemu, M. A. (2017). Item level diagnostics and model-data fit in item response theory (IRT) using BILOG-MG v3. 0 and IRTPRO v3. 0 programmes. *Global Journal of Educational Research*, *16*(2), 87-94.

Ferraro, M. B., & Giordani, P. (2015). A toolbox for fuzzy clustering using the R programming language. *Fuzzy Sets and Systems*, *279*(1), 1–16.

Finch, W. H., & French, B. F. (2015). *Latent variable modeling with R*. Routledge.

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. kluwer.

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.
Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). SAGE Publications, Inc.
Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small item parameter estimation in BILOG. *Applied Psychological Measurement*, *15*(3), 279–291
Hays, R. D., Brown, J., Brown, L. U., Spritzer, K. L., & Crall, J. J. (2006). Classical test theory and item response theory analyses of multi-item scales assessing parents' perceptions of their children's dental care. *Medical Care*, *44*(11), S60–S68.
Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika*, *68*(1), 123–149.
Jeon, M., & Rockwood, N. (2017). PLmixed: An R package for generalized linear mixed models with factor structures. *Applied Psychological Measurement*, *42*(5), 401–402.
Kaplan, R. M., & Saccuzzo, D. P. (2009). *Psychological testing: Principles, applications and issues* (Seventh Ed). Nelson Education.
Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and stan*. Elsevier
Kurniawan, D. D. (2015). Analisis kualitas soal ujian akhir semester matematika berdasarkan teori respon butir. *Prosiding Seminar Nasional Matematika dan Pendidikan Matematika UMS 2015*, 123–132.
Lemenkova, P. (2018). Factor analysis by R programming to assess variability among environmental determinants of the Mariana Trench. *Turkish Journal of Maritime and Marine Sciences*, *4*(2), 146–155.
Lemenkova, P. (2019). Statistical analysis of the Mariana Trench geomorphology using R programming language. *Geodesy and Cartography*, *45*(2), 57–84.
Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In *Handbook of modern item response theory* (pp. 1-28). Springer, New York, NY.
Mair, P. (2018). Item response theory. *Modern Psychometrics with R*, 95–159.
Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the mokken and the rasch approach to IRT. *Applied Psychological Measurement*, *14*(3), 283–298.
Muchlisin, M., Mardapi, D., & Setiawati, F. A. (2019). An analysis of Javanese language test characteristic using the Rasch model in R program. *REiD (Research and Evaluation in Education)*, *5*(1), 61–74.
Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing principles and applications* (sixth edit). Pearson Education.
Muthen, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics*, *10*(2), 133–142.
Nitko, A. J. (1996). *Educational assessment of students*. ERIC.
Ostrouchov, G., Chen, W. C., Schmidt, D., & Patel, P. (2012). *Programming with big data in R*. Oak Ridge National Laboratory and University of Tennessee.
Paek, I., & Cole, K. (2019). *Using R for item response theory model applications*. Routledge.
Penfield, R. D. (2003). IRT-Lab: Software for research and pedagogy in item response theory. *Applied Psychological Measurement*, *27*(4), 301–302.
Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making*, *29*(5), 453–469.
Reckase, M. D. (2009). *Multidimensional item response theory (statistics for social and behavioral sciences)*. Springer Sicence.

Reise, S. P. (2014). Item response theory. *The Encyclopedia of Clinical Psychology*, 1–10.
Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, *65*(1), 143–151.
Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items?. *Psychological Methods*, *8*(2), 164–184.
Retnawati, H. (2014). *Teori respon butir dan penerapannya untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Nuha Medika.
Rizbudiani, A. D., Jaedun, A., Rahim, A., & Nurrahman, A. (2021). Rasch model item response theory (IRT) to analyze the quality of mathematics final semester exam test on system of linear equations in two variables (SLETV). *Al-Jabar: Jurnal Pendidikan Matematika*, *12*(2), 399–412.
Schmidt, K. M., & Embretson, S. E. (2003). Measuring abilities and item response theory. *Comprehensive Handbook of Psychology: Research Methods in Psychology*, 429–445.
Steinberg, L., & Thissen, D. (2013). Item response theory. In *The Oxford handbook of research strategies for clinical psychology* (pp. 336–373). Oxford University Press.
Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*(1), 1-16.
Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin*, *104*(3), 385–395.
Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In *Differential item functioning* (pp. 67–113). Lawrence Erlbaum Associates, Inc.
Urbina, S. (2004). *Essentials of psychological testing*. John Wiley & Sons.Inc.
Van der Linden, W. J. (2017). *Handbook of item response theory: Volume 2: Statistical tools*. CRC Press.