

An analysis of Javanese language test characteristic using the Rasch model in R program

*¹Muchlisin; ²Djemari Mardapi; ³Farida Agus Setiawati

^{1,2,3}Department of Educational Research and Evaluation,
Graduate School of Universitas Negeri Yogyakarta

Jl. Colombo No. 1, Karangmalang, Depok, Sleman, Yogyakarta 55281, Indonesia

*Corresponding Author. E-mail: muchlisinjanuary@gmail.com

Submitted: 28 February 2019 | Revised: 02 May 2019 | Accepted: 07 May 2019

Abstract

One skill required to solve a problem in the 21st century is communication. Two international languages that are important in communication and thought at school are English and German language. However, beside international language, the local language, such as the Javanese language, is also essential and need to be maintained. The purpose of this study is to analyze the Javanese language test characteristics. This study was explorative research with secondary data collected by documentation of 220 students responses to the 50 multiple choice item of Javanese language test in the 11th grade of vocational high school. Data were analyzed using the Rasch model assisted by R program. Rasch model fits the data with 42 items after three times calibration. Based on difficulty level, ICC, and items reliability, there were 28 of 42 items (66.67%) that were good. This study finds out that generally, the Javanese language test is in the moderate category of difficulty. Hence, the need of evaluating the Javanese language test to make a better test that gives more accurate information about examinees' ability is crucial. The evaluation of the Javanese language test can be used to plan the next learning to get better Javanese language learning.

Keywords: *Javanese language test, Rasch model, R program*

Permalink/DOI: <https://doi.org/10.21831/reid.v5i1.23773>

Introduction

In the 21st century, there are some skills that are required. One of these skills is communication (Dede, 2010, pp. 7–8; Trilling & Fadel, 2009, p. 54; Zubaidah, 2017, p. 1). We need language to carry out communication. Some international languages are important, taught in the school, and widely used in the world, such as English, German language, Chinese language, etc. Beside international language, the local language, such as the Javanese language, is important and need to be maintained.

Central Java and Yogyakarta Special Region, two provinces in Indonesia, are very rich in terms of tradition and culture of Java. One of these traditions is the Javanese language that is used to speak to each other in daily life.

This is why the Javanese language lesson at school, especially in Java, still be held nowadays. At every end of the semester, a test is conducted to assess students ability in the Javanese language.

The assessment of the Javanese language test can be carried out by analyzing test characteristics, which was begun by collecting the information about the previous results of the test score (Sumintono & Widhiarso, 2015, p. 12). Besides to give a score to the students, the students' response can also be used to predict or explain the students' ability and item characteristic by analyzing test characteristic based on the Item Response Theory (IRT).

Test is very important both for teacher and students. A test can be used to classify the weakness in terms of verbal skills, me-

chanical skills, etc. (Allen & Yen, 1979, p. 1). Besides, a test is a powerful method of data collection with an impressive array for gathering numerical data rather than verbal kind (Cohen, Manion, & Morrison, 2007, p. 414). A test is defined as the standardized procedure for sampling behavior and describing it with categories or scores (Gruijter & van der Kamp, 2008, p. 2). The essential features of a test are a standardized procedure, a focused behavioral sample, and description in term of scores or categories mapping (Gruijter & van der Kamp, 2008, p. 2). The result of the test (scores) can be used to predict or explain the item and test performances (Lord & Novick, 2008, p. 358). Thus, the Javanese language test has to be analyzed in terms of its characteristics to get a better test in the next chance that can reach the test goal and give more accurate information about the examinee's ability.

The test has some uses. Five uses of a test include classification, diagnosis and treatment planning, self-knowledge, program evaluation, and research (Gregory, 2015, p. 29). A test can be a useful tool, but it can also be dangerous if misused (Allen & Yen, 1979, p. 5), depending on our professionalism in ensuring the use of the test accurately and as fairly as possible. Many extraneous factors can influence the test (Gregory, 2015, p. 31). Several sources that may influence the test are the manner of administration, the test characteristic, the testing context, examinee's motivation and experience, and the scoring method (Gregory, 2015, p. 31).

In a test, some plannings need to be prepared, including identifying the purposes, the test specifications, and selection of the contents, considering the form, the writing test, the layout, the timing, and planning the scoring of the test (Cohen et al., 2007, p. 418). We can make a good Javanese language test by paying attention to the planning and some influencing factors. Besides, a good result of the test, which is accurate, rich, and beneficial for evaluation will be obtained by analyzing the characteristics of the items or test of Javanese language using Item Response Theory (IRT).

There are some alternative ways to analyze test characteristics, including classical test theory (CTT) and item response theory (IRT). In CTT, it is difficult to analyze a test with a large amount of calculation to get useful information (Baker, 2001, p. 1). Besides, CTT has some weakness, such as the result of the measurement depends on the test characteristic used, item parameter depends on the examinee's ability, and the error measurement provided is limited for group measurement instead of individual information (Mardapi, 2017, p. 187). In CTT, if test is 'hard', the examinee ability will below; it is 'easy', the examinee ability will be higher (Ronald K. Hambleton, Swaminathan, & Rogers, 1991, p. 2). Therefore, CTT is considered to be not effective to analyze the Javanese language test.

The weakness of CTT is that it can be covered by IRT. IRT is one of the modern psychometric theories that provide useful tools for ability testing (Harrison, Collins, & Müllensiefen, 2017, p. 1). IRT is a powerful tool used to solve a major problem of CTT (Downing, 2003, p. 739). Item response theory (IRT) models, including Rasch, show the relationship between the ability of test participants from latent trait (e.g., Javanese language skills) and the opportunity to master the given items (answer the items correctly) in the form of logistic models (Finch & French, 2015, p. 181). IRT has 3 assumptions (Finch & French, 2015, p. 181; Mardapi, 2017, p. 187). These are monotonicity, unidimensionality, and local independence.

CTT has served development well in a test over several decades, but IRT has become mainstream rapidly as the theoretical measurement basis (Embretson & Reise, 2000, p. 3). The feature of IRT is specification of a mathematical function relating probability of an examinee's response on a test item to an underlying ability (Embretson & Reise, 2000, p. 8; Finch & French, 2015, p. 177; Gruijter & van der Kamp, 2008, p. 133; R K Hambleton & Swaminathan, 1985, p. 9; Ostini & Nering, 2006, p. 2; Reckase, 2009, p. 68; van der Linden & Hambleton, 1996, p. iii). In other words, the function describes in probabilistic terms, a person with low and high ability give

a different response (Ostini & Nering, 2006, p. 2). IRT is an important thing that can solve the problem of dealing the relationship between ability (examinee's mental traits) and response (performance) to the item (Lord & Novick, 2008, p. 397). IRT is used in so many education fields, not only in social science, even in medical education, it has some potential benefits (Downing, 2003, p. 739). In the IRT, some information about the test characteristic can be gained accurately, so that analyzing the Javanese language test using IRT needs to be conducted.

One of the models in IRT is the Rasch model. The Rasch model was developed by Georg Rasch, a Danish mathematician, in 1960 (Hailaya, Alagumalai, & Ben, 2014, p. 301; Jambulingam, Schellhorn, & Sharma, 2016, p. 50; Mallinson, 2007, p. 1; Young, Levy, Martin, & Hay, 2009, p. 545). There are some points of view about the Rasch model. Rasch model is a special case of one-parameter logistic (1 PL) model with item discrimination value is set equal to 1 (Finch & French, 2015, p. 181). Discrimination shows the ability of an item to differentiate among examinees ability (Finch & French, 2015, p. 181). The Rasch model can be expressed as:

$$P(x_j = 1 | \theta, b_j) = \frac{e^{(\theta - b_j)}}{1 + e^{(\theta - b_j)}} \quad (1)$$

In equation (1), x_j is the response to the item j with 1 being correct in the context of an achievement test. θ represents an individual ability, and b_j is the difficulty level of item j .

Analysis of the Javanese language test using Rasch model has practical benefits. We can check the model fits the data. Rasch model can define the probability of a specified response in relation to examinee's ability and item difficulty of a Javanese language test (Hailaya et al., 2014, p. 301; Jambulingam et al., 2016, p. 50). Using Rasch model, there is no need to differentially weight items to produce a total score that gives the maximum possible amount of information about latent trait; the number-right score is the best possible total score to use (Allen & Yen, 1979, p. 260). Rasch model produces the latent-trait (Javanese ability) and the item difficulty scale

that have desirable. Analyzing the Javanese language test using the Rasch model can be done by the R program.

The Javanese language test in the school has to be analyzed the characteristic using the Rasch model in IRT by R program to get some information. This information can be gained from the Item Characteristic Curves (ICC). ICC can provide the probability of the examinees at a given ability level of answering each item correctly (Hambleton & Swaminathan, 1985, p. 13). Beside ICC, there are the other important information about the items or the test that we can get by using the Rasch model in IRT. The Javanese language test in the school has to be analyzed the characteristic using Rasch model in IRT by R program to get some information. This information can be collected from the Item Characteristic Curves (ICC). ICC can provide probability of the examinees at a given ability level of answering each item correctly (Hambleton & Swaminathan, 1985, p. 13). Beside ICC, there are the others important information about the items or the test that we can get by using the Rasch model in IRT.

There are many studies of IRT application. They compared the use of IRT and CTT or studied the application of IRT to analyze the test characteristic. A study conducted by Downing (2003) contrasts the IRT with CTT and explores the benefit of IRT application in typical medical education settings. Downing just compares these models and explore the benefit of IRT theoretically; he did not go further discussing the application of IRT in the analysis. In this study, IRT was used to analyze the test by the Rasch model in the R program. Essen, Idaka, and Metibemu (2017) analyze the model-data fit in IRT using Bilog and IRTPRO program. They used two programs to analyze the model-data fit, but in this study, one model in one program was used to analyze the model's fit data, item fit model, the difficulty level of the items, items characteristics curve (ICC), item information curve (IIC), test information curve (TIC), the information given by each item, and the Javanese ability distribution. More complex information would be revealed in this study.

The study of Purnama (2017) was conducted to understand the characteristics of Accounting Vocational Theory test items by IRT using BILOG Program. In this study will analyze the characteristics of the Javanese language test using the Rasch model in the R program. Purnama's study analyzes the test using 2 PL, employing the Rasch model, which is the special case of 1 PL. Purnama's study did not use the ICC to analyze the item characteristics, while in this study, ICC will be used. Another study conducted by Setiawati, Izzaty, and Hidayat (2018b, 2018a) using IRT to analyze the test employs Bilog program, while this study employs the R program. A study by Iskandar and Rizal (2018) has some relevancy with this study. These studies use a program to conduct analysis. In their study, they analyze the validity, reliability, difficulty level, and the other cases, but not the items and test characteristic curve, the information functions, the ability average of examinees, etc. Those aforementioned studies used CTT, while this study uses IRT. It is hoped that this study would present findings which can contribute to analyzing the characteristic of the Javanese language test, so that there would be an evaluation for the Javanese language test to get a better one.

The Javanese language test will be analyzed by IRT. Analyzing the Javanese language test will be more accurate and can be used to estimate the relationship between the examinee ability and the examinee response to the items of the Javanese language test. Analyzing the Javanese language test using IRT will produce the analysis not just for the overall test, but also for individual items characteristic. The characteristics of item and test (IIC and TCC) estimate how accurate the Javanese language test will give us the information (IIC and TIC) and the other characteristics. Based on the explanations, the researchers decided to analyze the Javanese language test characteristics based on item response theory using the Rasch model in the R program.

Method

This study is explorative research, that is research which aims at finding the fact and

characteristics systematically and accurately about the Javanese language test (Arikunto, 2010, p. 14). The characteristics of the Javanese language test were analyzed using the Rasch model in the R program. This research was conducted in Yogyakarta from May to June 2018.

The data analyzed in this study are secondary data. The data were collected by the documentation method, which is collecting the answer sheet of 220 students' responses to the Javanese language test in Depok 1 Vocational High School, Yogyakarta. The Javanese language test consists of 50 multiple choice items.

The instrument unit, the Javanese language test, was made by the Javanese language teacher. Then, the researchers summarize the responses in the dichotomy data table. The wrong responses are denoted by 0, and the true responses are denoted by 1. The item number 1 was symbolized with B1, item number 2 was B2, item number 3 was B3, and so on. The data of the Javanese language test were analyzed based on IRT using Rasch model in the R program.

After the data were collected and analyzed using the Rasch model in the R program, some findings are gained. It described how the characteristics of the Javanese language test told us the probability of an examinee's response on the test item to an underlying ability (Javanese language ability). The researchers analyzed the model fits of the overall data, the difficulty level, and item fits of the model, ICC, TCC, IIC, TIC, item information, the Javanese language ability distribution, and the descriptive statistics for the Javanese language ability.

The model fits the overall data. The goodness of fit model was conducted to test whether the Rasch model fits with the overall data, whereas item fits model was done to test whether the model fits for individual items as well. Both will be fit if the p-value more than 0.05. If the Goodness of Fit Model has not met the fit criteria, then the item fits model would be conducted, and the items that did not fit would be removed. Then, the goodness of fit of the remained items would be re-

analyzed until the criteria were met, and we can continue to the next analysis.

In practice, the researchers set the category, e.g., a difficult level is said to be good if it has a difficulty value ranging from - 2.0 to 2.0 (Hambleton & Swaminathan, 1985, p. 107). In this study, an item can be said a good item if have difficulty level from - 3.0 to 3.0. The ICC will show about how the relationship between examinee ability with the true response probability, whereas TCC shows the relationship between examinee ability and the true score (sum of the true response probability). The IIC and TIC show the information that we can get based on the item or test for certain examinee ability. The item information is useful for item selecting. The criteria of the reliable item are if the item information value more than 0.5. The Javanese language ability distribution and descriptive statistics are all about examinee ability in this test. All of the information would explore the Javanese language test characteristics in this study.

Findings and Discussion

After the data were collected and analyzed, some results are gained. It describes how the characteristics of the Javanese language test told us the probability of an examinee's response to the test item to an underlying ability (Javanese language ability). It can be seen from model fits data, the difficulty level, and item fits model, ICC, TCC, IIC, TIC, the distribution of Javanese language ability, etc.

The first step of the analysis of the characteristic of the Javanese language test is the assessment of the model fit for the Rasch model. We have to make sure that overall model fit for Rasch model. It can be said that the model fits the data if the frequency of the observed and the model-predicted individuals for each response pattern are close to one another (Finch & French, 2015, p. 189). To analyze the model fit, we used the bootstrap chi-square procedure in R program (whether the model fits for the overall data). The bootstrap chi-square test of overall model fit for a Rasch model was conducted by command `GoF.rasch(model.rasch, B=1000)`. First, the re-

searchers analyzed the model fits for all items (50 items). The result shows that p-value is 0.006. If the p-value is less than 0.05, it means that the model does not fit the data. Thus, it is said that the model did not fit the data (for all items). Then the items fit model was analyzed (whether the model fits for the individual items as well) by command `item.fit(model.rasch, simulate.p.value = TRUE)`. There were three items that did not fit the model. These items are item number 27, 32, and 35. The data for these three items were removed, and the researchers analyzed the model which fits the data again.

The second analysis of the model fit of the data was done, and we got the p-value 0.017. It was still less than 0.05. It means that the Rasch model did not fit the data. Then the researchers analyzed the items fit the model for these 47 items. They got that the items number 3, 11, 13, 36, and 48 did not fit the model. The data for these items were then removed. Then, the researchers reanalyzed the model fit of the data with 43 items remained. The third analyzing of the model fit of the data showed that the model fits the data. It could be seen from the p-value were 0.053 (more than 0.05). Finally, after three times calibration of the fit-model, the researchers got the Rasch model fits the data without the items number 3, 11, 13, 27, 32, 35, 36, and 48 (there are 42 items that would be analyzed). In other words, the researchers had gotten the overall model-fit for the Rasch model, then, they could continue the other analysis.

The researchers analyzed the difficulty level of the items, and the items fit the model. The summary of the analysis is clearly presented in Table 1.

The center of item difficulty level is 0; negative value represents relatively easy, and positive value indicates relatively more difficult items (Finch & French, 2015, p. 184). Based on that statement, it indicates that when the value of difficulty is increasingly negative, then the difficulty level of the problem is easier and when the value of the difficulty becomes more positive then the level of difficulty becomes increasingly difficult.

From the Rasch's analysis of the difficulty level of the items, it is found that the easiest question is item number 20 (with difficulty level -15.7892) and the hardest problem is item number 23 (with difficulty level 0.9702).

In theory, the difficulty levels are in the range of minus infinity to infinity. There are some items that have a good category based on their difficulty level. There are 28 good

items, and the rest, 14 items, are not good based on the difficulty level. The not good items based on difficulty level are item number 5, 6, 7, 12, 14, 16, 17, 18, 19, 20, 25,29, 38,and 46. There are 69.77% of 43 items that are good in the difficulty level. Hence, the test in the moderate category based on the difficulty level.

Table 1. Difficulty level of items and the items fit of the model

Item No.	Difficulty level of the items	The items fit of the model
1	-0.8355	0.0792
2	-1.0570	0.6634
4	-0.3796	0.4554
5	-4.6802*	0.7165
6	-4.6802*	0.5149
7	-3.5262*	0.3861
8	-1.0317	0.1683
9	-2.8874	0.2574
10	-1.5902	0.3366
12	-5.0950*	0.6832
14	-5.7976*	0.6436
15	-2.6885	0.9208
16	-4.1508*	0.3465
17	-3.7959*	0.0891
18	-3.9593*	0.9208
19	-5.7976*	0.1584
20	-16.0705*	0.1881
21	-0.2267	0.0396#
22	-0.5127	0.9802
23	0.9695	0.3960
24	-1.8959	0.8713
25	-4.3832*	0.8614
26	-1.3516	0.7426
28	-1.7202	0.9604
29	-3.1221*	0.0693
30	-1.5902	0.2970
31	-0.4016	0.4356
33	-2.6282	0.4059
34	-1.7202	0.4653
37	-1.9713	0.9406
38	-3.5263*	0.3168
39	-2.0908	0.1287
40	-2.0102	0.1386
41	-1.1084	0.1386
42	-1.5589	0.2277
43	-1.4678	0.2277
44	-2.0908	0.8119
45	-2.9610	0.3762
46	-3.3073*	0.6436
47	-2.1756	0.4158
49	-1.3235	0.9505
50	-1.6541	0.3366

Notes:

*item is not good based on the difficulty level

#item misfit with the Rasch model

The teacher should pay attention to the not good category items. All of the items that are not good based on the difficulty level are categorized at too easy items. These items are not good because they are too easy for every examinee. It was indicated by all of their indexes of difficulty level which are smaller than -3.0.

Rasch model had fit with the data, but there is one item that did not fit with the Rasch model. This item is item number 21. We could not decide on these items. It was because these items did not fit with the model. It means that the characteristics of this item (item no. 21) based on the Rasch model were not adequately accurate.

The analysis of item characteristics is displayed in the form of curves for all items can be seen in Figure 1. The item characteristic curve (ICC) places the test participant's location on the latent trait measured on the x-axis and the ability to master an item on the y-axis (Finch & French, 2015, p. 184). The latent trait refers to the Javanese language ability, and the ability to master an item (probability answer correctly) refers to the probability of the examinee to respond correctly to the item. From ICC, it can be known about the probability of correctly answer from someone with a certain ability on an item. The command to get ICC for all items (42 items) together is `plot(model.rasch,type=c('ICC'))`. It gives us all the ICC of the item in the test.

Figure 1 shows the ICC of 43 items. It was difficult to interpret the curve if we used all ICC together. The ICC of the items number 23 was located at the most right position of the x-axis (Finch & French, 2015, p. 185). It means that the item number 23 is the most difficult item. The easiest item was not able to find, because it was so complex. However, it is clear that the item number 20 is the easiest item based on the difficulty level of the item. If the curve from these items is separated, we can see it more clearly. Thus, the ICC for item number 20, 23, and two other numbers can be compared. The ICC for item number 20 and 23, and two other items are presented in Figure 2.

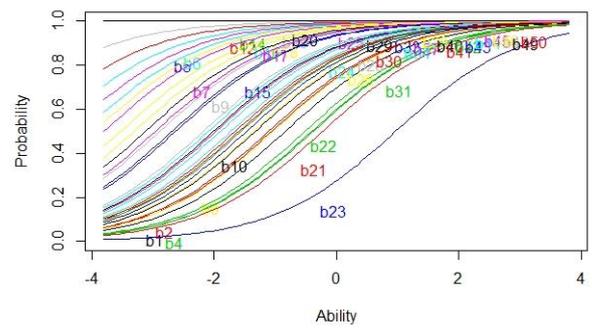


Figure 1. The ICC of Javanese language test

From Figure 1, some of ICCs are not good because the correct response probability for the examinee with low ability is high. These items are item number 5, 6, 7, 12, 14, 16, 17, 18, 19, 20, 25, 29, 38, and also 46 (total of 14 items). All of these items have fitted the model.

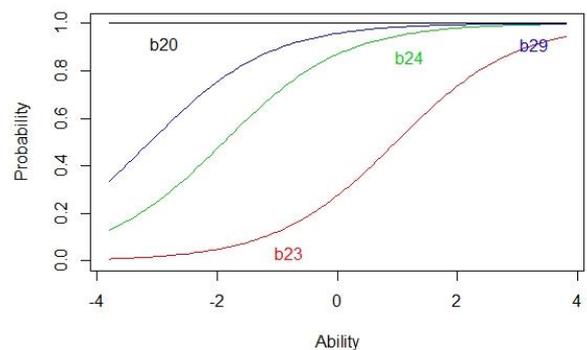


Figure 2. The ICC for items number 20, 23, 24, and 29

However, the difficulty levels of these items are not good. Thus, these items (see Figure 3) are not good based on the ICC and difficulty level.

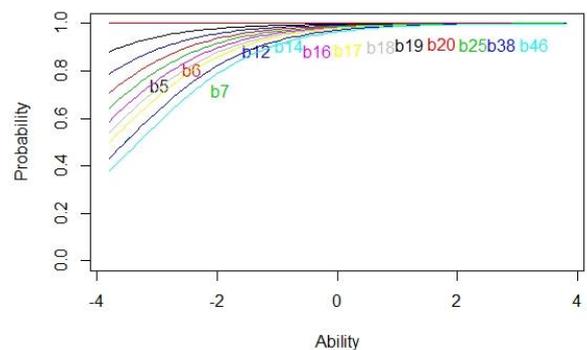


Figure 3. Items with not good ICC

To look the ICC of a specific item, let us say that items number 20, 23, 28, and 29,

used the command `plot(model.rasch,type=c("ICC"),items=c(17,20,21,25))`. It is a little different from the command for all ICC, in which, sort number from specific items was mentioned. It would make every ICC of some items in one graph to be able to compare easily.

Figure 2 presents some item characteristic information. For item number 20, regardless of the student's ability, the probability to answer correctly is the same for all examinee, which is 1.0 (always true). It indicates that the item number 20 is too easy for every examinee. It means that examinee with any Javanese language ability will be able to respond the item correctly (the examinee with ability value -4 through 4 could respond to this item correctly). For the hardest item (item number 23), the examinee with ability 1 will have probability approximately 0.5 to answer this item correctly. To get high probability about 0.9 or more, the examinee should have Javanese language ability almost 4. The Javanese language ability would be needed to increase the opportunity to answer this item correctly.

The test characteristic in correlating the ability with true score can be found by TCC (Test Characteristic Curve). True score is the sum of correct answer probability. The Javanese language test TCC is shown in Figure 4.

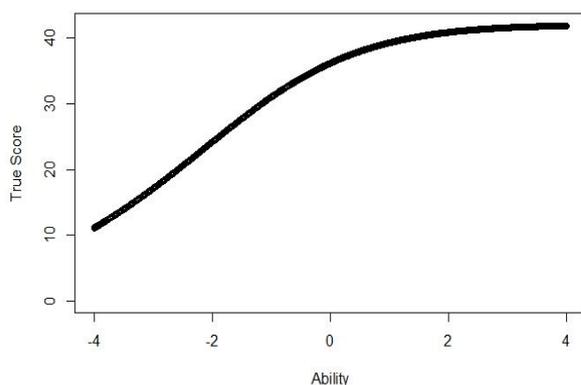


Figure 4. The TCC of the test

From Figure 4, it is known that the test is an easy category. The examinee with a low ability (-3) will have true scores approximately 19, and the examinee with an average ability (0) will have true scores approximately 35 (near to the maximum true score, that is 42).

The examinee with ability value 0 (average ability) will have a different probability for each item. He/she will have probability 0.2

for item number 23, probability approximately 0.8 or more for item number 24 and 29, and probability 1.0 (true response) for item number 20. Figure 2 explains that the difficulty level of item number 20 is easier than item number 24 and 29, and item number 24 and 29 are easier than item number 23. Figure 1 shows that some ICCs are not good since the correct response probability for examinee with low ability is high. These items are item number 5, 6, 7, 12, 14, 16, 17, 18, 19, 20, 25, 29, 38, and 46 (14 items). Those items have fitted the model. The item characteristic for every item can be described the same way as we had done to the item number 20, 23, 24, and 29, by separating it from the other ICC so that it will be seen clearly.

In addition to the ICC, we used the R program to plot the item information curve (IIC). The IIC describe the information function of an item. It refers to the degree to which item reduces the uncertainty in the estimation of Javanese language ability (the latent trait) value for an individual (Finch & French, 2015, p. 185). A high value of information for a specific range of ability distribution indicates that the item provides relatively more information regarding the latent trait (Javanese language ability) in that region than another region in the distribution (Finch & French, 2015, p. 186). Based on the IIC, we can see how reliable the item in giving information. All the IIC are shown in Figure 5. There are 50 IIC with each degree in estimating the information given by each item. The command to get IIC for all item in the test is `plot(model.rasch,type=c("IIC"))`. The command for specific IIC is `plot(model.rasch,type=c("IIC'),items=c(18,21,25,40))`, that will produce IIC for item number 20, 23, 28, and 47. The IIC for 43 items is shown in Figure 3, and the IIC for item number 20, 23, 28, and 47 is shown in Figure 7.

There are 43 IIC that can describe how reliable each item in the giving information about the Javanese language ability value for an individual. There are just 43 IIC of the 43 items that the Rasch model fits for the data. From Figure 4, we can get the most accurate and inaccurate items in giving information about the examinee's ability in the Javanese language. These are shown by item number

20 and 23. The IIC for these numbers is shown separately from the others in Figure 5 with item number 28 and 47.

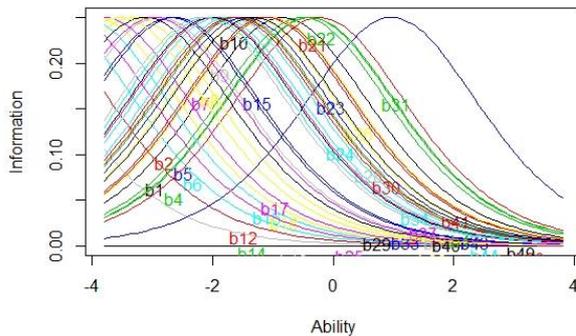


Figure 5. The IIC of 42 items

Some of IICs give maximum information for examinee with a low ability (Figure 6). These items are item number 5, 6, 7, 12, 14, 16, 17, 18, 19, 20, 25, 29, 38, and 46 (14 items). These items did not give maximum or give low information for the examinee with the medium or high ability. These items are not good, because they give maximum or high information just for low ability examinee and these items based on the ICC and the difficulty levels are not good. Therefore, we can conclude that these items are not good based on the ICC, IIC, and difficulty level.

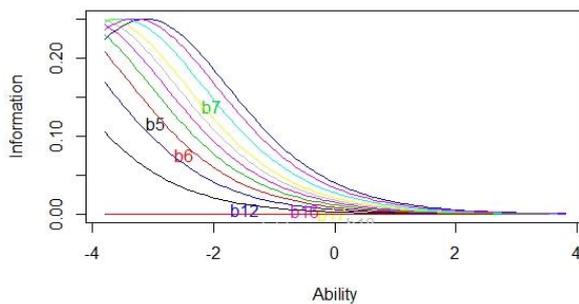


Figure 6. Item with not good IIC

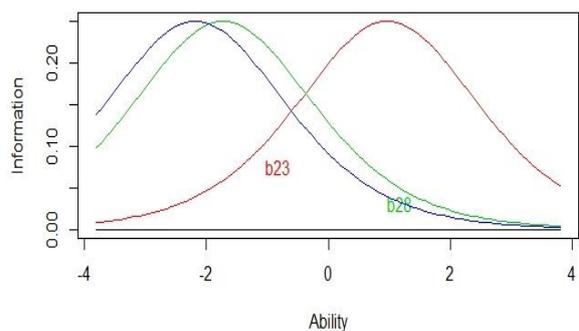


Figure 7. The IIC for item number 20, 23, 28, and 47

Figure 7 shows the IIC for item number 20 is the most inaccurate in giving information about the examinee's Javanese language ability. This item cannot give the information accurately because any examinee with any ability shows 0 information value that can be provided by this item. We cannot differentiate the examinee's ability. There is no information about the examinee ability (in the Javanese language) that we can get if we use this item to measure them. The IIC for item number 23 shows that it is needed ability approximately 1 to get information about 0.25, in other words that item 23 provides maximum information for estimating θ (Javanese language ability) around values of 1. The item number 28 and 47 will give maximum information about the examinee if he/she has ability about -2. The IIC for every item is different, but this study shows more specific item information curve for item number 20, 23, 28, and 47. If we want to look at the IIC from the other item, we can separate it from the others.

Item information curves show the information function for every item in the test. For the total information, the function can get from Test Information Function. There are some features of the test information function. These are defined for a set of the test items at each point on the ability scale, the amount of the information is influenced by quality and number of test items, etc. One of the most important features of the test information function is that the contribution of each item to complete information is additive (Hambleton & Swaminathan, 1985, p. 104). The test information curve that shows the total information function is like Figure 8. The command to get test information curve is `plot(model.rasch,type=c("IIC"), items=c(0))`.

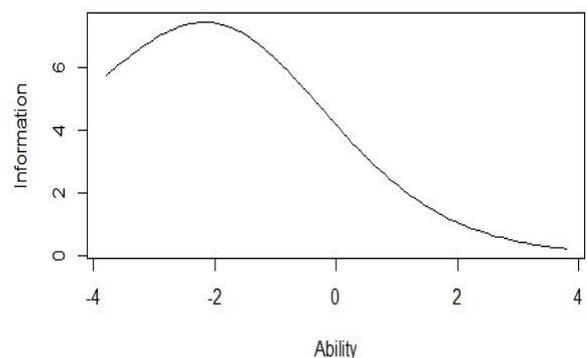


Figure 8. Test information curve

Figure 8 shows the estimate of the test information function on the curve. TIC presents how reliable the Javanese language test is. The TIC interpretation is similar to the IIC interpretation. The test provides us maximum information for estimating θ around values of -2 . Thus, the test will be good to be used for examinee with low Javanese language ability. The test was less accurate in giving information on examinee with Javanese language ability 0 (average ability) or more than 0 ability.

The information function (IIC or TIC) has some application in the test construction,

item selection, measurement precision assessment, test comparison, scoring weight determination, and scoring methods comparison (Hambleton & Swaminathan, 1985, p. 101). In item selection, we can select the item that can provide accurate information on examinee's ability. The item's IIC, which does not provide information, means the item should not be used in the test (like item number 20). The item does not provide information in any theta (ability), so it should not be used in the test.

Table 2. The information of each item in theta -3.0 until 3.0

Item No.	Information	Percentage
1	0.88	87.60%
2	0.86	85.78%
4	0.90	89.93%
5	0.16	15.74%
6	0.16	15.66%
7	0.37	36.94%
8	0.86	86.01%
9	0.53	52.58%
10	0.79	79.35%
12	0.11	11.01%
14	0.06	5.82%
15	0.57	57.43%
16	0.24	24.03%
17	0.31	31.04%
18	0.28	27.67%
19	0.06	5.82%
20	0	0.09%
21	0.90	90.31%
22	0.89	89.44%
23	0.87	86.55%
24	0.74	74.38%
25	0.2	20.06%
26	0.83	82.61%
28	0.77	77.38%
29	0.47	46.78%
30	0.79	79.39%
31	0.9	89.86%
33	0.59	58.87%
34	0.77	77.38%
37	0.73	73.00%
38	0.37	37.05%
39	0.71	70.70%
40	0.72	72.27%
41	0.85	85.29%
42	0.8	79.81%
43	0.81	81.09%
44	0.71	70.70%
45	0.51	50.76%
46	0.42	42.14%
47	0.69	68.98%
49	0.83	82.95%
50	0.78	78.42%

The complete information of the test across all values of the Javanese language ability (latent trait) can be obtained by using the command `information(model.rasch, c(-10,10))`. The subcommand `c(-10, 10)` identifies the range of the theta (ability) for which information is requested. The total information that is provided by the test at the examinee's ability ranges from -10 to 10 equal to 41.93 or 100%. It means that the test will give maximum information if the test were used in the examinees with ability -10 until 10. If we request for the ability values in range 0 to 10, with the command `information(model.rasch, c(0,10))`, is 5.9 or 14.08% of the total information provided by the Javanese language test. In the normal distribution raw, the area of range -3 to 3 equals to 95% of the total area. The total information that could be given by the test if we measure in the ability range of -3 to 3 is 24.98 or 59.58% of the total information. There is still moderate information which we could obtain by using this instrument in measuring the examinee with the ability in this range.

Beside the ICC, TIC, and the total information, we can get the information given by each item in the range of a certain ability (theta). In this study, the information, that is given by each item in the ability range of -3 until 3, are listed in Table 2. We can know the percentage that we get from the total information of each item.

Based on Table 2, we can see the information given by each item in the theta -3.0 until 3.0. The information can be used for item selection. How reliable the item depends on the percentage of information gotten from each item in this range of theta. We can set the criteria for reliable item like we need. For example, if we will compose a test, we cannot use item number 20, because it gives us very small information. If we set the criteria for reliable information of each item by more than 50%, we get 28 reliable items of 42 items that can be used (there are 66.67%). The remaining unreliable items (14 items) are not good. Incidentally, these unreliable items are also categorized as not good based on the ICC, IIC, and difficulty level.

Obtaining latent trait (Javanese language ability) estimates for the Rasch model in R program, we used the command `theta.rasch<-factor.scores.rasch(model.rasch)` to save the θ estimates from the Rasch model. Then, we used the command `summary(theta.rasch$score.dat$z1)` to get a basic descriptive statistic of ability (θ). The output of this command is shown in Table 3.

Table 3. The latent trait estimates

Min.	Median	Mean	Max
-2.0780	-0.1534	-0.1138	1.6538

We can see that the mean of Javanese language ability for the sample is -0.1138, with the minimum being -2.0780 and the maximal being 1.6538. The standard deviation of Javanese language ability gotten by the command `sqrt(var(theta.rasch$score.dat$z1))`. The result of the standard deviation of Javanese language ability is 0.750783. The plot of the latent trait (Javanese language ability) was gotten by the command `plot(theta.rasch)`. The plot of the latent trait (Javanese language ability) based on the Rasch model is shown in Figure 9.

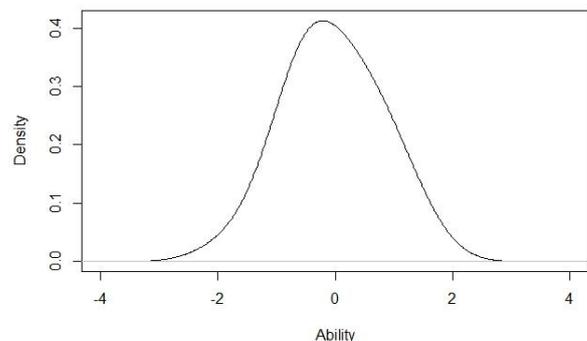


Figure 9. Plot of theta

Figure 9 shows that the distribution of Javanese language ability almost centered at 0. The center of the plot ability shows the mean of ability, that is -0.1367. Thus, that is the reason why it is almost centered for those with Javanese language ability value of 0. The highest density of Javanese language ability is located in the mean ability value. The distribution of the theta (Javanese language ability) based on the analysis using the Rasch model in R program shows the normal distribution

curve. The right side and the left side of the distribution curve are almost balanced.

Figure 8 shows that maximum information will be obtained when the Javanese language ability value is -2 . However, the mean ability from the examinees is -0.1367 , meaning that generally, the Javanese language test did not give maximum information on the examinee's Javanese language ability. It can be said that the test is less accurate. Thus, evaluation of the Javanese language test is needed.

The evaluation of the Javanese language test will make the test better, so that it can give more accurate information for a teacher in the assessment of precision measurement. The teacher will have further steps or ideas to be applied in the next Javanese language lesson if they know the examinee's ability generally to make the examinee's Javanese language ability increase. It is hoped that, with the increasing of the Javanese language ability, the student will practice it in their daily life. They retain the culture and character of Javanese language in their lives, which there are so much positive learning, culture, character, interaction in Java, and so much more.

This study analyzed the Javanese language test based on the Rasch model in the R program. For the next study, we hope they can use the other model to analyze the Javanese language test based on the procedure for each model. It is hoped there will be more test analysis, maybe about mathematics test, a certain language test, or the other test, especially the Javanese language test. Therefore, it will give the teacher a view to making a better test in the next chance that gives accurate information about the examinee ability and measures the examinee ability more accurate. It is better to use item response theory to analyze the test because there are some benefits that we can get. We can know about each item characteristic, the information function of each item, and the other benefits.

Conclusion

Based on the result of the analysis of Javanese language test using the Rasch model in R program, the interpretation, and the discussion, the researchers can conclude some points of the characteristic of the Javanese

language test. The calibration of the fit-model was done in three times. It was done to get model fits the data with 42 items in the fit-model. Analysis of the difficulty level shows that there are 28 items of 42 items (66.67% of 43 items) that are a good category. Therefore, the Javanese language test is in the moderate category based on the difficulty level.

We can see the characteristic of the item in predicting the true probability for examinee with a certain ability in the ICC and the test characteristic from the TCC. Based on ICC and IIC, there are 28 good items (66.67%). Based on the information that we can get from each item (item information) in the theta -3.0 to 3.0 , there are 28 items (66.67% give information more than 50%) of 42 items can be used (moderate category based on the information in this range of theta). From descriptive statistic, it can be said that the ability of examinees are in the moderate category because the mean of ability is -0.1138 (near from 0.00 /average ability). Generally, the Javanese language test is in the moderate category. It will be better if we evaluate the Javanese language test to make a better test that gives more accurate information on the examinees' ability. The evaluation of the Javanese language test can be used by the Javanese language teachers to plan the next learning in their class to get better Javanese language learning.

Acknowledgment

The researchers thank Depok 1 Vocational High School, which had permitted the researchers to collect the data. Gratitude is also sent to contributors, Ali, Desy, and Laras for their help during the data collection and dealing with the dichotomy data tabulation.

References

- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Cole Publishing.
- Arikunto, S. (2010). *Prosedur penelitian: Suatu pendekatan praktik* (Revised ed). Jakarta: Rineka Cipta. Retrieved from <https://doi.org/10.1017/CBO9781107415324.004>

- Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.
- Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). London and New York, NY: Routledge Falmer.
- Dede, C. (2010). Comparing frameworks for 21st century skills. In J. Bellance & R. Brandt (Eds.), *21st century skills: Rethinking how students learn* (pp. 51–76). Bloomington, IN: Solution Tree Press.
- Downing, S. M. (2003). Item response theory: Applications of modern test theory in medical education. *Medical Education*, 37(8), 739–745. <https://doi.org/10.1046/j.1365-2923.2003.01587.x>
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists: Multivariate applications book series*. London: Lawrence Erlbaum Associates.
- 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. <https://doi.org/10.4314/gjedr.v16i2.2>
- Finch, W. H., & French, B. F. (2015). *Latent variable modeling with R*. New York, NY: Taylor & Francis.
- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications* (7th ed.). New York, NY: Pearson Education.
- Gruijter, D. N. M., & van der Kamp, L. J. T. (2008). *Statistical test theory for the behavioral sciences*. New York, NY: Taylor & Francis Group.
- Hailaya, W., Alagumalai, S., & Ben, F. (2014). Examining the utility of Assessment Literacy Inventory and its portability to education systems in the Asia Pacific region. *Australian Journal of Education*, 58(3), 297–317. <https://doi.org/10.1177/0004944114542984>
- Hambleton, R K, & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.
- Hambleton, Ronald K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.
- Hambleton, Ronald K, & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Nijhoff.
- Harrison, P. M. C., Collins, T., & Müllensiefen, D. (2017). Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation. *Scientific Reports*, 7(1), 1–19. <https://doi.org/10.1038/s41598-017-03586-z>
- Iskandar, A., & Rizal, M. (2018). Analisis kualitas soal di perguruan tinggi berbasis aplikasi TAP. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 22(1), 12–23. <https://doi.org/10.21831/pep.v22i1.15609>
- Jambulingam, T., Schellhorn, C., & Sharma, R. (2016). Using a Rasch model to rank big pharmaceutical firms by financial performance. *Journal of Commercial Biotechnology*, 22(1), 49–60. <https://doi.org/10.5912/jcb734>
- Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. (F. Mosteller, Ed.). Reading, MA: Addison-Wesley.
- Mallinson. (2007). Rehabilitation institute of Chicago in rehabilitation research provides new insights. *Atlanta*, pp. 1–3.
- Mardapi, D. (2017). *Pengukuran, penilaian, dan evaluasi pendidikan* (2nd ed.). Yogyakarta: Parama Publishing.
- Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: SAGE Publications.

- Purnama, D. N. (2017). Characteristics and equation of accounting vocational theory trial test items for vocational high schools by subject-matter teachers' forum. *REiD (Research and Evaluation in Education)*, 3(2), 152–162. <https://doi.org/10.21831/reid.v3i2.18121>
- Reckase, M. D. (2009). *Multidimensional item response theory (Statistics for social and behavioral sciences)*. New York, NY: Springer.
- Setiawati, F. A., Izzaty, R. E., & Hidayat, V. (2018a). Analisis respons butir pada tes bakat skolastik. *Jurnal Psikologi*, 17(1), 1–17. <https://doi.org/10.14710/jp.17.1.1-17>
- Setiawati, F. A., Izzaty, R. E., & Hidayat, V. (2018b). Items parameters of the space-relations subtest using item response theory. *Data in Brief*, 19, 1785–1793. <https://doi.org/10.1016/j.dib.2018.06.061>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada asesmen pendidikan*. Bandung: Trim Komunikata. Retrieved from https://umexpert.um.edu.my/file/publication/00013268_127390.pdf
- Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. San Francisco, CA: Jossey-Bass.
- van der Linden, W. J., & Hambleton, R. K. (1996). *Handbook of modern item response theory*. New York, NY: Springer Science+Business Media. <https://doi.org/10.1007/978-1-4757-2691-6>
- Young, D. J., Levy, F., Martin, N. C., & Hay, D. A. (2009). Attention deficit hyperactivity disorder: A Rasch analysis of the SWAN rating scale. *Child Psychiatry and Human Development*, 40(4), 543–559. <https://doi.org/10.1007/s10578-009-0143-z>
- Zubaidah, S. (2017). Keterampilan abad ke-21: Keterampilan yang diajarkan melalui pembelajaran. In *Isu-Isu Strategis Pembelajaran MIPA Abad 21*. Sintang, West Kalimantan: Program Studi Pendidikan Biologi STKIP Persada Khatulistiwa Sintang. Retrieved from https://www.researchgate.net/publication/318013627_keterampilan_abad_ke-21_keterampilan_yang_diajarkan_melalui_pembelajaran