



Application Rasch model using R program to analyze the characteristics of chemical items

Untung Desy Purnamasari *, Badrun Kartowagiran

Universitas Negeri Yogyakarta. Jalan Colombo No. 1, Yogyakarta 55281, Indonesia

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

Received: 30 March 2019; Revised: 11 June 2019; Accepted: 7 August 2019


Abstract

Chemistry is one of the subjects taught in high school. To find out and assess students' understanding regarding chemistry subjects in one semester can be proven by a test. The tests used must have good quality. This study aims to provide information about the characteristics of chemical items test using the Rasch model. Descriptive explorative was used in this study. The subject of the study was tenth-grade students in Xaverius Senior High School taken the final semester examination on chemistry subject. The object of this research was the form of item tests and student answer sheets. Data collection techniques used documentation. Student answer sheets were analyzed using the R program. The results showed that the reliability of the tests was 0.819 or high category. Subsequently acquired a good level of difficulty about which amounted to 28 items. Also, the average student ability is 0.008, with a minimum ability of -2.309 and a maximum of 2.233. ICC and IIC obtained are very accurate in predicting students' abilities. Teachers can use chemicals items used in the final semester examination as an item bank for use in the evaluation of students' abilities. However, two items need to be revised level of difficulty to produce a good question.

Keywords: characteristics of chemical items, Rasch model, R program

How to Cite: Purnamasari, U., & Kartowagiran, B. (2019). Application rasch model using R program in analyze the characteristics of chemical items. *Jurnal Inovasi Pendidikan IPA*, 5(2), 147-156.

doi:<https://doi.org/10.21831/jipi.v5i2.24235>

 <https://doi.org/10.21831/jipi.v5i2.24235>

INTRODUCTION

Education is a program that involves several components and works together in a process to achieve programmed goals. As a program, education is a conscious and deliberate activity that is directed towards achieving a goal (Tshabalala, Mapolisa, Gazimbe, & Ncube, 2015). To find out whether the implementation of the program can achieve its objectives effectively and efficiently, it is necessary to do an evaluation. Therefore, evaluation is carried out on the components and work processes so that if there is a failure to achieve the objectives, the components and processes that are the source of failure can be traced (Dada & Ohia, 2014). Evaluation is decision making based on measurement results and standard criteria. Measurement and evaluation are two continuous activities. Decision making is done by comparing the measurement results with the specified criteria. Therefore, there are two activities in evaluating, namely making measurements and making decisions (Miller, Linn, & Gronlund, 2009).

The evaluation aims to measure and control the quality of education as stated in the Republic of Indonesia Law Number 20 of 2003 concerning the National Education System and Republic of Indonesia Government Regulation Number 19 of 2005 concerning National Education Standards (Presiden Republik Indonesia, 2005). The realization of the things above could be seen at the end of each semester in each education unit evaluating student learning outcomes in the form of a final semester examination. The final semester exam is held twice in one school year, namely in odd and even semesters. The final semester exam is carried out in the form of a written test. This test is conducted to measure students' abilities on certain subjects tested.

The test is one of the tools to make measurements, namely a tool to collect information on the characteristics of an object (de Gruijter & Van der Kamp, 2008). Tests can also be interpreted as several questions that must be given a response to measure the level of a person's ability or reveal certain aspects of the person subject to the test (Tshabalala et al., 2015).



A test is made by compiling items based on the existing grid. The test is used as a technique or measurement tool which is used as an "objective" and "standardized" measure of behavior samples (Cohen et al., 2002).

Test participants' responses to several questions and statements describe abilities in a particular field (Gregory, 2004). The purpose of conducting the test is to find out the learning achievements or competencies that have been achieved by students for a particular field (Lord & Novick, 2008). Test results are information about the characteristics of a person or group of people. This characteristic can be a person's cognitive abilities (Allen & Yen, 2001; Bahar, 2013; Baumgartner, Jackson, Mahar, & Rowe, 2007; Mardapi, 2017; Taub, Floyd, Keith, & McGrew, 2008).

To obtain accurate and precise measurement results, the final semester examination items used must be really good, and by what goals are to be measured. Therefore, to meet the criteria of a good, reliable, valid instrument and be able to produce accurate data by the objectives of measurement, it is necessary to validate the instrument items and measure their reliability. Validity is defined as the accuracy and accuracy of the instrument in carrying out its measurement functions (Anderson, 2003, p. 10; Kubiszyn & Borich, 2013, p. 3; Van de Walle, 2010).

Validity shows the extent to which the scale can accurately and accurately reveal data about the attributes that have been designed. Validity as the main characteristic that must be possessed by each measuring instrument must be completely compiled and designed according to the theoretical concept. Validity comes from the word validity, which means the accuracy of a test or scale in carrying out its measurement function. Judging from the validity of the test or the measurement validity that has been classic, validity is defined as the extent to which the test measures what is intended to be measured (Retnawati, 2016). Cronbach (1984) emphasized that the validation process does not aim to validate the test tool but does validate the interpretation of data obtained by certain procedures.

Reliability is a criterion that is no less important as validity. The reliability concept explains how far the results of a measurement process can be trusted. The measurement results can be trusted if in several times the implementation of the same subject group will get relatively the same results. The quality or failure of the questions can be known from the degree of

difficulty or the level of difficulty possessed by each question (Retnawati, 2016). A reliable measuring instrument consists of valid items. So, every reliable must be valid, but every valid one is not necessarily reliable.

Analysis of the quality of the final semester examination items is very important to do to improve the quality of the questions and improve the quality of the questions that will be tested in the next period. The questions are analyzed to find out the good questions and the bad ones. Good questions can be used as a measurement tool and reference in making questions in the next period. The bad questions that can still be revised are corrected so that they can be stored in the question bank so they can be reused. Whereas the question is not good, which requires significant revisions should be discarded (Miller et al., 2009).

Analyzing items becomes an activity that must be done for educators. The poor quality of the questions tested is the cause of the event so that educators are required to improve the quality of the questions that have been written. Item analysis is a structured statistical group, used to evaluate the quality of tests during the process of development and construction of tests. With the item analysis activity, educators can make decisions in making judgments on the process of collecting, summarizing, and using information obtained from students' responses (Tshabalala et al., 2015).

The purpose of analyzing the items is to obtain the quality of questions by reviewing the items before the questions that are made will be used in the test. Also, the analysis of items can help identify deficiencies in the test and find out whether students have or have not understood the material that has been taught (Allen & Yen, 2001). Item analysis can help improve understanding of why test scores can predict multiple criteria, show why a test is reliable or not reliable, and improve test characteristics.

Analyzing the items cannot be separated from the use of the classical test theory (CTT) and item response theory (IRT) approach. This is evidenced by the number of researchers who use these two approaches (Crocker & Algina, 2008). In CTT, scores are obtained based on the number of individual responses to various items. However, there could be a gap in the participants who took the exam. Either because the test place is not conducive, excessive anxiety or nervousness makes some participants not focus on answering questions or choosing the wrong answer option.

The questions tested are sometimes very difficult to do. This became a problem for the examinees, which resulted in a reduction in the score obtained. To overcome this problem, the researcher used the IRT approach (Harrison, Collins, & Müllensiefen, 2017).

One of the techniques for data analysis is to use item response theory 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. Also, 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 to see the ability of individual examinees (de Gruijter & Van der Kamp, 2008; Embretson & Reise, 2000; Finch & French, 2015; Hambleton & Swaminathan, 1985; Linden & Hambleton, 1996; Ostini & Nering, 2006; Reckase, 2009). Therefore, to overcome this problem, experts make new theories to complete and correct the limitations that exist in classical test theory. This theory is what we later know as the item response theory (Embretson & Reise, 2000).

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, there are assumptions underlying the item response theory, and the most commonly used are *unidimensionality* and *local independence*. *Unidimensionality* means measuring only one ability (θ) in a test for each examinee, while *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 with different items (Finch & French, 2015; Mardapi, 2017)(Hambleton, 2018).

The most well-known item response theory models are the logistic parameter model (PL) i.e. 1-PL model, or Rasch model, the 2-PL model and 3-PL model (Hambleton, 2018). These models contain an estimate of the latent nature of reading or depression, the ability to distinguish between individuals with different levels of the construct, 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 (Lord & Novick, 2008). Item response theory 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, but what distinguishes it is that the Rasch model has a discriminant value set equal to 1. The 1-PL model is one of the most widely used models. If using the 1-PL model, the item used is only tested the level of difficulty. In the 2-model that is focused only on the level of difficulty test and the discriminant of the item. The last is 3-PL model where this model tests the parameters of difficulty, discriminant, and guessing items (Downing, 2003).

The level of the item difficulty parameter is an opportunity to answer correctly on a problem at a certain level of ability. The difficulty index item (b) is measured by the item score produced by the answers of several test participants. 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, 2018). 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. Usefulness for educators is as re-learning and giving suggestions for students about the learning outcomes and preventing biased items. The usefulness for testing and teaching is to make a test with the accuracy of the data on the problem and to know the weaknesses and advantages of the school curriculum and the presence of biased items (Lord & Novick, 2008).

Another thing about item analysis is model fit-data. Model fit-data can be investigated at the item or person level. Especially Item-fit model, items are said to be fit with the model if the probability value (significance) < 0.05 . Fit models that can be used are 1-PL or Rasch, 2-PL, and 3-PL models. The last thing that is not less important in item analysis is *Item Characteristic Curve (ICC)* and *Item Information Characteristic (IIC)*. ICC describes the opportunity relationship to answer correctly with the level of ability of examinee. Also, it can be seen which items are the easiest and most difficult on a test. Each item has an information function, and the number is an information function of the test so that the function of the test package information will be high if the constituent items have a high

information function. To obtain the information function, it can be seen the IIC graph. The information function obtained can be a function of test and item information (Ackerman, Gierl, & Walker, 2003; Hambleton & Swaminathan, 1985; Salirawati, 2011; Sutrisno, 2016).

There are several studies that have been carried out using item response theory. Research conducted by (Iskandar & Rizal, 2017) to determine the quality of questions in universities using the TAP application. Also, the research conducted by (Aziz & Prasetyo, 2015) was conducted to determine the characteristics of odd semester final exam questions in high school physics subjects using Parscale 4.0. Both of these studies both use applications in analyzing questions. They analyze validity, reliability, level of difficulty, differentiation, and others. But they did not analyze the ICC, IIC, the item information function, the average student ability, and other complex information. This research uses dichotomous questions, but one of them uses polytomous.

To obtain good quality items, it is very important to analyze the characteristics of the items using the Rasch model with the R program. Questions that are of good quality can accurately measure the achievement of learning objectives.

METHOD

This study was ex-post facto design indicates that research is carried out after something has happened (Ary, Jacobs, Irvine, & Walker, 2018). The subject was all students of tenth-grade students in Xaverius Senior High School taken the final semester examination on chemistry subject in the academic year of 2018/2019 were 49 students. The object of this research was the form of item tests and student answer sheets. Data was collected through multiple-choice tests as many as 30 items. The data analyzed use R program.

RESULT AND DISCUSSION

The item response theory is the relationship between the ability of the test participant and the probability of answering an item correctly. Rasch model is a very simple model in item response theory, which only measures the level of difficulty of the item. The minimum number of samples used in the Rasch model is 30 people

(Linacre, 2015). Before analyzing using the Rasch model, it is very important to test assumptions so that the analysis results with the Rasch model are not biased (Retnawati, 2016).

There are two basic assumptions of the Rasch model, namely unidimensional and local independence. The unidimensional scale can be evaluated by performing a factor analysis on items designed to evaluate the structure of factors (Van Alphen, Halfens, Hasman, & Imbos, 1994). In other words, each item only measures one dimension of students' abilities. Unidimensional assumptions can be proven by doing factor analysis using the help of SPSS software. The results of the general assumption test are shown in Table 1.

Table 1. Result of KMO and Barlett Test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.541
Bartlett's Test of Sphericity	Approx. Chi-Square	641.262
	df	435
	Sig.	.000

Table 1 shows the KMO value of 0.541 (> 0.05) means that the sample used in this study is sufficient so that it can be used in the subsequent analysis. Then unidimensional test can be seen in the scree plot as in Figure 1.

Figure 1 shows that there is 1 dominant factor in the Final Semester Examination of Chemistry subject at Xaverius High School in Ambon city. This can be seen from the change in eigenvalue from the first factor to the second factor, which is so large. On the second factor onwards, the change in the eigenvalue is not that big. This is also evident in the steepness of the first factor to the second factor, which is so large. For this reason, the unidimensional assumption test in the Final Semester Examination of Chemistry subject at Xaverius High School in Ambon city is fulfilled.

Test of unidimensional assumption has been fulfilled, then automatically the assumption of local independence is also fulfilled (Retnawati, 2016). This means that among the factors in the Final Semester Examination of Chemistry subject in Xaverius High School in Ambon city correlate with each other. Because both assumptions have been fulfilled, the item analysis can be carried out using the R Program.

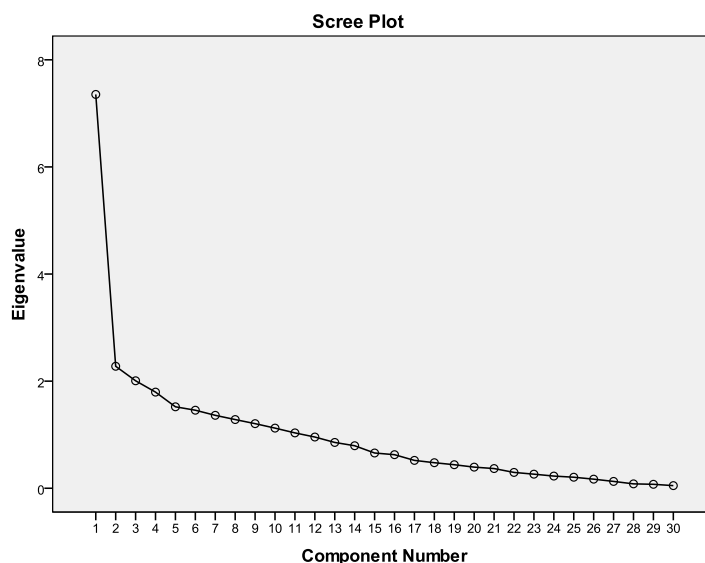


Figure 1. Scree Plot Results of Factor Analysis

The LTM package is one package to analyze IRT in program R. This package is not only able to analyze IRT with models 1, 2, and 3-PL, it can also analyze items with polytomous responses (Rizopoulos, 2006). As with most R packages, the use of this package must use a scripted approach (CLI, Command Line Interface) with the syntax:

```
library(ltm)
```

```
SMAxaverius_rasch<-rasch(dataSMAxaverius,  
constraint =  
cbind(ncol(dataSMAxaverius)+1,1))
```

After that, to analyze the parameters of the item, namely the parameter level of difficulty of the item using the command with the syntax:

```
coef.rasch(SMAxaverius_rasch)
```

Based on the results of the analysis, it is known that for each item has discriminant, which is assumed to be equal to the value of 1. Therefore, the opportunity to answer correctly in each item is assumed to be the same for all test participants. The difficulty level of the item if the difficulty level is close to -2, the item difficulty index is low, whereas if the difficulty value decides +2, the item difficulty index is very high for a test group (Allen & Yen, 2001; Finch & French, 2015; Hambleton, 2018). The results obtained are shown in Table 2.

Based on Table 2, the results show that the items in the Final Semester Examination of Chemistry subject in Xaverius High School in Ambon City have one item that is classified as difficult, namely item 9 with difficulty level (b) > +2, and one item that is classified as easy, namely

item 4 with difficulty level (b) <-2. For other items classified as moderate with difficulty level (b) ranging from -2 to +2.

Table 2. Parameter of the Difficulty Level of Items

No. of items	Difficulty	Category
Item_1	-0.921	Good
Item_2	-0.049	Good
Item_3	0.373	Good
Item_4	-2.036	Not good
Item_5	-1.423	Good
Item_6	-0.259	Good
Item_7	-0.366	Good
Item_8	-1.867	Good
Item_9	2.424	Not good
Item_10	1.045	Good
Item_11	1.046	Good
Item_12	0.161	Good
Item_13	0.927	Good
Item_14	-0.366	Good
Item_15	1.294	Good
Item_16	-0.154	Good
Item_17	1.046	Good
Item_18	0.056	Good
Item_19	-1.562	Good
Item_20	0.927	Good
Item_21	1.294	Good
Item_22	1.294	Good
Item_23	0.589	Good
Item_24	0.589	Good
Item_25	1.294	Good
Item_26	1.564	Good
Item_27	1.294	Good
Item_28	-1.866	Good
Item_29	1.426	Good
Item_30	1.426	Good

Furthermore, the model fit test to determine fit Rasch models with observed data. The model match test uses the Bootstrap test because it can determine the value of the actual latent variable; in this case, the ability of the test participant (Maydeu-Olivares, 2013). Test the suitability of the model using the syntax command (Finch & French, 2015):

GoF.rasch(SMAXaverius_rasch, B=1000)

The results of the analysis model fit of test obtained p-value = 0.907 (> 0.05) shows that fit the Rasch model of the data being tested. Model fit of test results is shown in Table 3.

Table 3. Results of the analysis model fit

```
Bootstrap Goodness-of-Fit using Pearson chi-
squared
Call:
rasch(data = dataSMAXaverius, constraint =
cbind(ncol(dataSMAXaverius) + 1, 1))
Tobs: 83389623
# data-sets: 1001
p-value: 0.907
```

Furthermore, the model fit can also be done by comparing the number of items that are fit on the model. Items that are fitted with the model are shown in Table 4 and can be known to use the syntax command (Finch & French, 2015):

*item.fit(SMAXaverius_rasch,
simulate.p.value=TRUE)*

Table 4. Result of item fit with Rasch Model

No. of items	Item fit Value	Category
Item_1	0.099	Fit
Item_2	0.624	Fit
Item_3	0.416	Fit
Item_4	0.465	Fit
Item_5	0.129	Fit
Item_6	0.772	Fit
Item_7	0.515	Fit
Item_8	0.366	Fit
Item_9	0.822	Fit
Item_10	0.386	Fit
Item_11	0.990	Fit
Item_12	0.198	Fit
Item_13	0.327	Fit
Item_14	0.337	Fit
Item_15	0.010	Not fit
Item_16	0.852	Fit
Item_17	0.455	Fit
Item_18	0.347	Fit
Item_19	0.455	Fit
Item_20	0.535	Fit
Item_21	0.683	Fit
Item_22	0.356	Fit

No. of items	Item fit Value	Category
Item_23	0.693	Fit
Item_24	0.495	Fit
Item_25	0.644	Fit
Item_26	0.733	Fit
Item_27	0.723	Fit
Item_28	0.901	Fit
Item_29	0.168	Fit
Item_30	0.792	Fit

Table 4 shows that there is only one item that is not fit with the Rasch model, namely item 15. In item response theory, we can know the reliability of the test. The reliability of the test used in the Final Semester Examination of Chemistry subject at Xaverius High School in Ambon city is shown in Table 5.

Table 5. Reliability of the test

```
$reliability
Coefficient Alpha
0.819
```

Table 5 shows that the level of precision and consistency of the test scores have good accuracy. With the known value of reliability ranges from 0-1. The higher the reliability coefficient of a test (close to 1), the higher the accuracy. It is proven by the score of coefficient Alpha of 0.819. Thus this test has reliability in the high category. A reliable measuring instrument consists of valid items.

The item information function is a method to explain the strength of an item on the test device, selection of test items, and a comparison of several test devices. The item information function states the strength or contribution of test items in revealing the latent trait measured by the test. The information function for each item with ability parameters (θ) -4 to 4 is shown in Table 6.

Table 6. Item Information Function

No. of items	Item Information Function
Item_1	0.95 (94.85%)
Item_2	0.96 (96.41%)
Item_3	0.96 (96.17%)
Item_4	0.87 (87.49%)
Item_5	0.92 (92.52%)
Item_6	0.96 (96.3%)
Item_7	0.96 (96.18%)
Item_8	0.89 (89.16%)
Item_9	0.83 (82.75%)
Item_10	0.94 (94.42%)
Item_11	0.94 (94.42%)
Item_12	0.96 (96.37%)
Item_13	0.95 (94.87%)
Item_14	0.96 (96.18%)
Item_15	0.93 (93.26%)

No. of items	Item Information Function
Item_16	0.96 (96.37%)
Item_17	0.94 (94.42%)
Item_18	0.96 (96.41%)
Item_19	0.92 (91.61%)
Item_20	0.95 (94.87%)
Item_21	0.93 (93.26%)
Item_22	0.93 (93.25%)
Item_23	0.96 (95.81%)
Item_24	0.96 (95.81%)
Item_25	0.93 (93.26%)
Item_26	0.92 (91.59%)
Item_27	0.93 (93.26%)
Item_28	0.89 (89.16%)
Item_29	0.92 (92.5%)
Item_30	0.92 (92.5%)

The Item Characteristics Curve (ICC) shows the characteristics of items that indicate the ability of test participants with the probability of answering questions correctly. The curve maps the ability of the test participant to Y and the probability of answering the question correctly on X. The ICC of the Final Semester Examination of Chemistry subject at Xaverius High School in Ambon city is shown in Figure 2 and can be known to use the syntax command (Finch & French, 2015):

```
plot(SMAXaverius_rasch,type=c("ICC"))
```

Figure 2 shows that the most difficult item is item 9, so only test participants have the ability

between 0-4 who can answer the item correctly. This is evidenced in the ICC where the distribution of point 5 on the curve is at the far right while the easiest item is item 4 so that test-takers who have the ability of -4 to 4 can answer the item easily.

The last analysis is an estimation of the function of information; this is intended to inform the test compiler of how well the ability at each level is. The information function is not dependent on test distribution because this technique is an application of item response theory. Item information function is shown in the form of an item information curves (IIC). The IIC of the Final Semester Examination of Chemistry subject at Xaverius High School in Ambon city is shown in Figure 3 and can be known to use the syntax command (Finch & French, 2015):

```
plot(SMAXaverius_rasch,type=c("IIC"))
```

Based on Figure 3, it can be seen that the information for each item is different where the graph of each item is marked with a different color. The peak point of the curve which is below zero capability provides information that suitable items are given to low-ability test participants, and the peak of the curve that is above zero ability provides information that the item is suitable to be given to test takers with high ability.

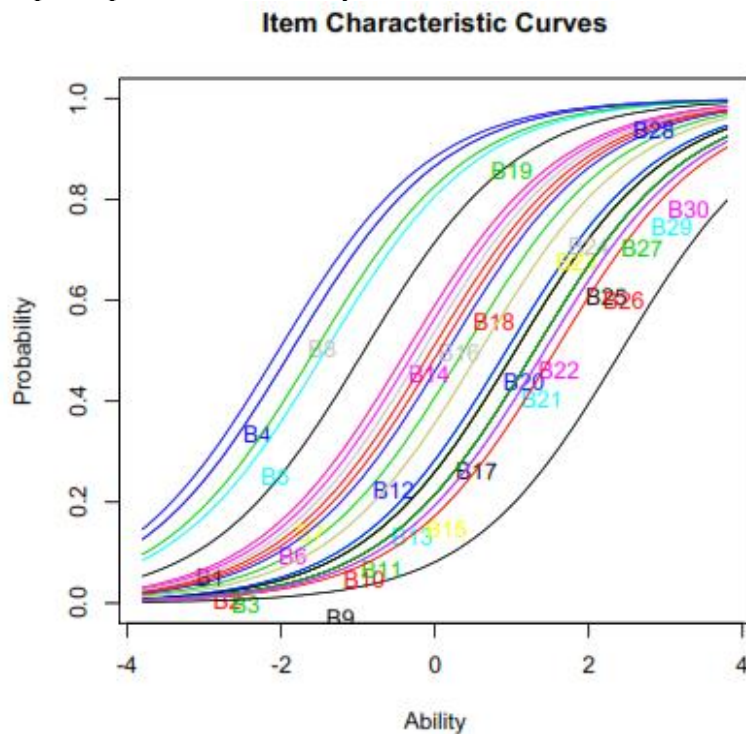


Figure 2. Item Characteristics Curve

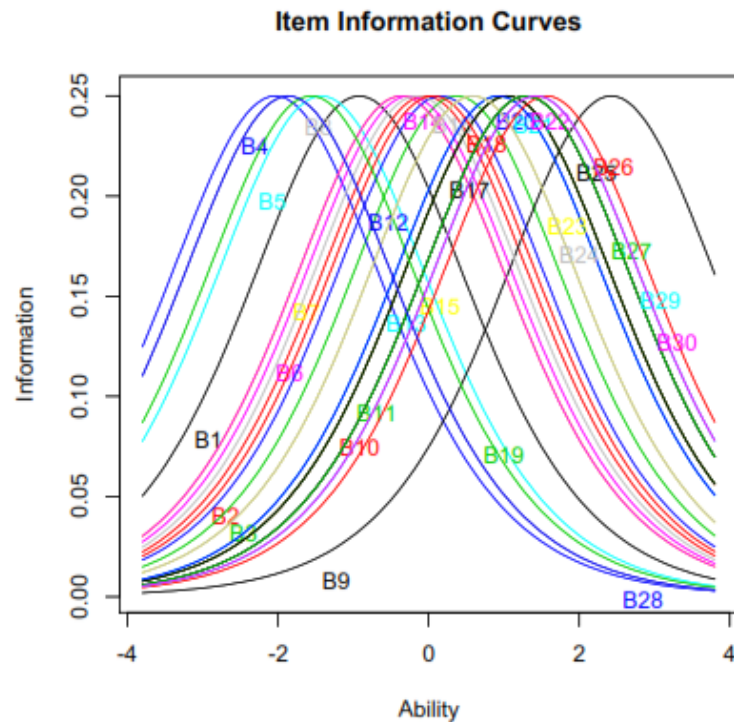


Figure 3. Item Informations Curves

Based on the above curve shows the curve that has the highest information (peak curve, which is at the ability level approaching -2 to 2. Because the average student ability is 0.008 with the minimum ability of the test participants is -2.308 and the maximum ability of the test participants the is 2.233. It shows that the sample is normally distributed.

CONCLUSION

Based on the results of data analysis and discussion of the research that has been stated there are several things related to the characteristics of chemical items, there are: (a) from 30 items there are 28 items in good category based on the level of difficulty parameters, while 2 items still need to be revised to produce good items. It can be seen in the form of characteristic items curve where item 4 is very easy while item 9 is very difficult. (b) The reliability of the items is 0.819 with the high category. Because the test is reliable, then automatically the tests are valid also. (c) The average ability of students is 0.008 with a minimum ability of -2.309 and a minimum ability of 2.233. Experts make shown in detail in item information curves.

REFERENCES

- Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, 22(3), 37–51. <https://doi.org/10.1111/j.1745-3992.2003.tb00136.x>
- Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. California: Waveland Press, Inc.
- Anderson, L. W. (2003). *Classroom assessment: Enhancing the quality of teacher decision making*. New Jersey: Lawrence Erlbaum Associates, Inc.
- Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. (2018). *Introduction to research in education*. Cengage Learning.
- Aziz, A., & Prasetyo, Z. (2015). Karakteristik soal ujian akhir semester gasal mata pelajaran fisika SMA kelas X di Kabupaten Lombok Tengah Nusa Tenggara Barat. *Jurnal Evaluasi Pendidikan*, 3(2), 99–111. Retrieved from <http://journal.student.uny.ac.id/ojs/index.php/jep/article/view/1266>
- Bahar, A. (2013). *The influence of cognitive abilities on mathematical problem solving performance*. The University of Arizona.
- Baumgartner, T. A., Jackson, A. (Tony), Mahar, M., & Rowe, D. (2007). *Measurement for evaluation in physical education and exercise science*. New York: McGraw-Hill.

- Cohen, L., Manion, L., & Morrison, K. R. B. (2002). *Research methods in education*. New York, N.Y.: Routledge.
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Ohio, USA: Cengage Learning.
- Cronbach, L. J. (1984). *Essentials of psychological testing*. New York: Harper & Row Publisher.
- Dada, E. M., & Ohia, I. (2014). Teacher-made language test planning, construction, administration and scoring in secondary schools in ekiti state. *Journal of Education and Practice*, 5(18), 71–76. Retrieved from <https://www.iiste.org/Journals/index.php/JEP/article/view/13928>
- de Gruijter, D. N. M., & Van der Kamp, L. J. T. (2008). *Statistical test theory for the behavioral sciences*. Chapman and Hall.
- 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, Inc.
- Finch, W. H., & French, B. F. (2015). *Latent variable modeling with R*. New York, N.Y.: Taylor & Francis.
- Gregory, R. J. (2004). *Psychological testing: History, principles, and applications*. Allyn & Bacon.
- Hambleton, R. K. (2018). Emergence of item response modeling in instrument development and data analysis. *Medical Care*, 38(9), II60–II65.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. New York, N.Y.: Springer Science+Business Media. <https://doi.org/10.1007/978-94-017-1988-9>
- 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. (2017). Analisis kualitas soal di perguruan tinggi berbasis aplikasi tap. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 21(2), 12–23. <https://doi.org/10.21831/pep.v22i1.15609>
- Kubiszyn, T., & Borich, G. D. (2013). *Educational testing and measurement: Classroom application and practice* (10th ed.). London: Wiley.
- Linacre, J. M. (2015). Sample size and item calibration stability. *Journal of Applied Measurement*, 3(1).
- Linden, W. J. van der, & Hambleton, R. K. (1996). *Handbook of modern item response theory*. New York, N.Y.: Springer Science+Business Media. <https://doi.org/10.1007/978-1-4757-2691-6>
- Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. (F. Mosteller, Ed.). Addison-Wesley.
- Mardapi, D. (2017). *Pengukuran penilaian dan evaluasi pendidikan*. Yogyakarta: Nuha Medika.
- Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, 11(3), 71–101. <https://doi.org/10.1080/15366367.2013.831680>
- Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching*. (L. Reinkober, Ed.) (10th ed.). Kevin M. Davis.
- Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, California: SAGE Publications, Inc.
- Presiden Republik Indonesia. Peraturan pemerintah Republik Indonesia no 19 th 2005 tentang standar nasional pendidikan, Pub. L. No. 19, Peraturan pemerintah Republik Indonesia 1 (2005).
- Reckase, M. D. (2009). *Statistics for social and behavioral sciences: Multidimensional item response theory*. New York, N.Y.: Springer Science+Business Media.
- Retnawati, H. (2016). *Validitas reliabilitas dan karakteristik butir*. Yogyakarta: Parama Publishing.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modeling. *Journal of*

- Statistical Software*, 17(5).
<https://doi.org/10.18637/jss.v017.i05>
- Salirawati, D. (2011). Pengembangan instrumen pendeteksi miskonsepsi kesetimbangan kimia pada peserta didik SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 15(2), 232–249.
<https://doi.org/10.21831/pep.v15i2.1095>
- Sutrisno, H. (2016). An analysis of the mathematics school examination test quality. *Jurnal Riset Pendidikan Matematika*, 3(2), 162–177.
<https://doi.org/10.21831/jrpm.v3i2.11984>
- Taub, G. E., Floyd, R. G., Keith, T. Z., & McGrew, K. S. (2008). Effects of general and broad cognitive abilities on mathematics achievement. *School Psychology Quarterly*, 23(2), 187–198.
<https://doi.org/10.1037/1045-3830.23.2.187>
- Tshabalala, T., Mapolisa, T., Gazimbe, P., & Ncube, A. C. (2015). Establishing the effectiveness of teacher-made tests in Nkayi District Primary Schools. *Nova Journal of Humanities and Social Sciences*, 4(1), 1–6.
<https://doi.org/10.20286/jhss.v4i1.29>
- Van Alphen, A., Halfens, R., Hasman, A., & Imbos, T. (1994). Likert or rasch ? Nothing is more applicable than good theory. *Journal of Advanced Nursing*, 20, 196–201.
<https://doi.org/10.1046/j.1365-2648.1994.20010196.x%0A>
- Van de Walle, J. A. (2010). *Elementary and middle school mathematics : teaching developmentally*. Boston: Pearson /Allyn and Bacon.