ATLANTIS PRESS

Advances in Social Science, Education and Humanities Research, volume 326

3rd International Conference on Current Issues in Education (ICCIE 2018)

Characteristics of Assessment Instrument for Diagram and Verbal Representation Ability of High School Physics Students in Physics with Partial Credit Model

Johan Pamungkas
Physics Education Department
Yogyakarta State University
Yogyakarta, Indonesia
johan.pamungkas2016@student.uny.ac.id

Supahar
Physics Education Department
Yogyakarta State University
Yogyakarta, Indonesia

Warsono
Physics Education Department
Yogyakarta State University
Yogyakarta, Indonesia

Jumadi
Physics Education Department
Yogyakarta State University
Yogyakarta, Indonesia

***Abstract*—The characteristics of assessment instrument are important to obtain representative information about students' knowledge and abilities in education. The representation ability is fundamental ability to solve problems in learning Physics. This study aims to determine the characteristics of assessment instrument for the representation abilities of diagram and verbal in high school Physics. The assessment instrument consists of two test sets with four category response according to Partial Credit Model (PCM). The subject of the research was a sample of 305 students of XI *IPA* class from high schools in Yogyakarta. The subject was selected through purposive sampling based on the result of national high school examination year 2016/2017. The research data were analyzed with item response theory (IRT) according to PCM. Based on the result estimation of parameter and student ability, that (1) item difficulty index ranged from -1.02 to +1.17, which satisfies the criteria of good test items; (2) all test items fit with the PCM model; (3) the information function is 10 and SEM is 0.21; and (4) the ability estimation ranges from -2.08 to +2.10, indicating that students' abilities are varied. The assessment instrument fits with PCM model and can be categorized as good, meaning that it can be used to measure diagram and verbal representation abilities of high school students in Physics.**

Keywords—assessment instrument, representation ability, Partial Credit Model

I. INTRODUCTION

Assessment becomes an important activity in education. Assessment has been regulated in the Minister of Education and Culture Number 104 of 2004. Assessment includes all the means used to know and interpret the ability of each student to the competence of the material that has been studied on the basis of a specific criterion [1,3]. Teachers always make an assessment to know the students' abilities before, during and after learning. Assessment can be conducted with instrument test. The instrument tests must have validity in order to obtain valuable information. The result was expected to improve the quality of education.

The assessment instrument was used teachers usually in the test form. A test is a measuring instrument which requires correct or incorrect answers to obtain student information about knowledge and ability were learned during the lesson in evaluating achievements [2,4]. The instrument test must have the goal to be achieved based on basic competences (*Kompetensi Dasar/ KD*) in Curriculum 2013. The test plays an important role in the assessment of cognitive aspects. The learning process can be improved based on the results of decision making on the assessment. In the development process, the characteristics of the instrument test must be validated in order to obtain a good test item [5,6]. Based on interviews from teachers in senior high school, the developed assessment instruments have not been through the analysis process of test item both from experts and empirics validation. Assessment instrument developed by teachers is referred to only on a basic competencies and indicators. Assessment instrument developed by teachers was tested directly on students without qualitative and quantitative analysis. This is a weakness of teachers in education,

especially the assessment using the writing test. The lack of information on the characteristics results in the low quality of test items .

Similar things occur on the tests in learning Physics. Students' ability in Physics learning can be known from assessment by using an assessment instrument. The test fits the goals and needs of learning Physics. In senior high school, Physics is the study of the natural laws underlying the phenomenon of everyday life which related in a complex way [7,8], and emphasizes students in problem solving [9]. The Physics was underlying development of study and technology. Therefore, assessment instrument consisting of tests and assessment guides are necessary to measure problem solving skills . So need to develop assessment instruments in the aspects of problem solving. However, because the related study of problem solving ability is still less specific to be measured and still too wide in scope, so it needs to be limited the topic of the study.

Problem solving ability is related to students' representational ability. Problem solving ability is seen as cognitive activities that involve the construction of student representation [10]. The ability of representation is an expression, idea or idea embodied in various ways or forms to solve a problem [11,12]. Representations are classified into verbal representations (text or words), visual (diagrams, pictures, graphics, tables) and symbols (numeric, symbolic). The use of representation is an effective way for students to understand Physics problems so as to assist in solving problems [13,14]. In this study the test used is to measure the ability of students' verbal and diagrams representation.

The ability of representation diagrams becomes one of the fundamental ability of Physics learning. The representation diagram is used to analyze the forces acting on the object [15]. Students who describe the diagram correctly can solve the problem correctly rather than misrepresent the diagram or not at all [16]. Nevertheless, there are still many students who have difficulty when applying the diagram representation in Newton mechanics [17]. This indicates a lack of diagram representation ability. Students need to understand that the ability in diagram representation is an important step in organizing and simplifying the information provided into a more appropriate representation. This can be due to the learning or assessment model. In this case, assessment will be discussed, because inappropriate judgment will make unsuitable results.

The ability of verbal representation is important as the first step in solving the problem [18]. Students must understand the problems faced. Students process information into words and notes (written text). However, students do not use verbal representations effectively, resulting in errors when solving problems [19,20]. Students' habits in studying Physics tend to memorize mathematical equations without emphasizing the concept of understanding or the appearance of other representations. Students do not understand the language of Physics correctly so that understanding the concept of Physics becomes weak. This may be due to an improper assessment model, especially its assessment test.

The representation abilities of verbal and diagram becomes important in solving a Physics problem. Based on the test form, multiple choice tests are often found during observations in some schools. Multiple choice tests are relatively easy in scoring. But in this study using essay form. This is because during the completion of Physics problems related to the representation of diagrams and verbal there is a process which can be assessed by scoring with polytomous models. The weaknesses of the multiple-choice test are (1) the possibility of the student to guess the answers, and (2) the student's thought process that cannot be seen because the scanning of the student's response uses a dichotomous model that has not yet assessed the stage of problem solving. The weakness can be overcome with a form of essay test using scoring with a polytomous models, although the time needed to correct the response of students becomes longer. Therefore, an essay test is valid and has characteristics that can describe the ability of students' verbal and diagrams representation.

In previous research, the test instrument has passed the process of content validation or expert judgment. Therefore, the main problem raised in this study is the test characteristics that measure the representation ability of verbal and diagram. The characteristic of the test must be proven to be an empirically valid and accountable instrument. This will make decisions more meaningful. These results can be used as information about students has mastered competencies that have been studied. But the items developed are still not known characteristics. Therefore, it is necessary to do a test or calibration analysis to find out the characteristics of the test item.

There are two ways of analyzing items, namely the classical test theory (CTT) and the item response theory (IRT). The classical test theory approach still has limitations because it is group dependent and item dependent [21]. IRT was developed to overcome the weaknesses of CTT. IRT uses a probabilistic model to improve the limitations of CTT. IRT is used to identify the problem characteristics by calibrating the items on an IRT model. Item calibration is the process of estimating item parameters (item difficulty index, item discrimination and guessing) and parameters of respondent ability on IRT model [22]. There are several IRT models on known polytomous scans such as the Partial Credit Model (PCM), the Generalized Partial Credit Model (GPCM), and the Graded Response Model (GRM). However, this study used PCM model with scoring model polytomous. The PCM estimates the item difficulty parameter by assuming the same power discrimination for each item. The thing to note in this IRT is the number of test participants because different parameter models will require different numbers of participants to have stable item characteristics [21]. In addition, the item

response theory assumption also needs to be met first. To make it easier in the analysis or the estimation, computer software is used.

The characteristics of assessment instrument can determine a good item, with certain criteria. Hence, the purpose of this research to get the characteristics of assessment instrument for measuring the representation ability of verbal and diagram in learning Physics, especially (1) the goodness of fit on PCM model, (2) the item difficulty index, (3) the information function and (4) the estimation of student ability. It is expected that teachers who want to develop the test calibrate the item to ensure its quality.

II. LITERATURE REVIEW

A. Partial Credit Model

In the execution of the description test, scoring is usually done partially based on the steps to be taken to correctly answer a point. Scoring is done stepping and grading score obtained by the participants by summing the students score each step, and ability is estimated with raw score. This scoring model is not necessarily correct, because the difficulty level of each step is not taken into account. An alternative approach that can be used is the item response theory approach (IRT) for polytomous scoring, one of them with partial credit model (PCM). The PCM is an extension of the Rasch model. PCM is a scoring model of polytomous, while the Rasch model on dichotomous data. The data on this research involve four categories with polytomous model. The scoring on the PCM model is based on a category score showing the number of completion stages. A higher score indicates a greater ability than a low category score. The category scores on PCM indicate the number of steps to correctly complete the item, that the student ability of each test participant can be estimated by calculating the probability of each participant in answering each step in completing a test question.

The results of this type of compound test can be analyzed according to the Rasch model, whereas the essay or description test can be analyzed using the partial credit model (PCM). The early development of IRT in the model polytomous more familiar as an extension of the Rasch model is now called the partial credit model (PCM). There is an assumption used in PCM where each item has the same discrimination power. There are several considerations in the use of PCM which is an extension of Rasch model 1-PL. The first consideration is the use of samples that are not as big as if the process of calibrating data polytomous using 2PL or 3PL model [23]. The second consideration is that the response characteristic of each item following the PCM model is the degree of difficulty (threshold) of a category stage below it to the above categories are not the same among items one and the other. Therefore, the difference between categories is not the same.

It is important to know the difficulty levelof the item, so that the information obtained is more valid. The item is considered as good if it has an item difficulty index (b_i) in the range of -2 to +2 [21]. This indicates that item with index value nearing -2 means that it is categorized as having a low degree of difficulty, while item approaching +2 means that it has a high difficulty level. If the index of difficulty passes the limit it is said that the question can be too easy or too difficult; so it does not describe the ability of the respondent. The equation of PCM with item parameter and item difficulty index (b_i) following equation (1) can be described as follows [24]:

$$\theta \quad P_{jk}(\theta) = \frac{\exp \sum_{v=0}^{k} (\theta - b_{jv})}{\sum_{h=0}^{m} \exp \sum_{v=0}^{k} (\theta - b_{jv})}, \quad k = 1, 2, 3, ..., m \tag{1}$$

with

$P_{jk}(\theta)$: the probability of a participant ability θ obtaining a score on category k in item j

θ : the ability of participant

$m+1$: the number of categories k of items j

b_{jk} : item difficulty index of category k item j

1) Goodness of fit

The goodness of fit was analyzed by statistical test using likelihood comparison test (likelihood ratio test). This test is used to check whether the estimation of item parameters in different scoring groups is equal in the sampling error of the estimate. The test items fit with PCM model, if INFIT MNSQ was close to 1.0 and standard deviation close to 0.0 according to Adams and Khoo [24]. This means that the test instrument which developed for representation skills of diagram and verbal in senior high school fulfil criteria according to PCM (1PL).

2) Information Function

The information function is an important method in Item Response Theory (IRT). The item information function describes the quality of an item on the test instrument, the selection of test items, and the comparison of some test instrument [25]. The item information function states the item contribution of the test in uncovering the latent trait as measured by the test. The item information function can help in the selection model which items fit the model. The test information function is the sum of item information function [21]. The test information function will be high if the test item has a high information function as well. The test information function can be mathematically written as follows:

$$I_i(\theta) = \sum_{i=1}^{n} \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta) Q_i(\theta)} \quad (2)$$

The values of the item parameter index and the ability of the test participants are the results of estimates in which the truth is probability and not independent of measurement error. In IRT, standard error measurement (Standard Error of Measurement, SEM) has a close relationship with the information function. The information function has a relationship of inverse quadratic with SEM [26]. The larger of information function so the value of SEM will be smaller with quadratic. The SEM is defined by the following formula:

$$SEM\left(\hat{\theta}\right) = \frac{1}{\sqrt{I(\theta)}} \quad (3)$$

with

$I(\theta)$: information function

$SEM\left(\hat{\theta}\right)$: standard error measurement estimation value

III. Method

A. Type of research

This research was a quantitative descriptive research. This research was conducted to find out the characteristics of test item in measuring the ability of students' verbal and diagrams representation. The population of research is all students of XI *IPA* class in senior high school in Yogyakarta. The sample used in this research is 305 students of XI *IPA* class in senior high school. The subjects of this study were chosen by purposive sampling, based on the result of national high school examination year 2016/2017. The sample involved all students present in the execution of the test. This is in accordance with the terms of the number of samples for IRT analysis, which ranges from 200 to 1000 people [27].

The research data was obtained from student response from the test. The essay-related test is related to students' representational and verbal ability on the subject of Physics. The scoring model of test was the polytomous model with four categories. The test consists of two test sets of problems, namely A and B test sets. This was aimed to reduce cheating by students who sat side by side during the test.

B. Data analysis techniques

Analysis of a research data is using the response theory of items following the Partial Credit Model (PCM). PCM as an extension of the Rasch (1-PL) model using a sample that is not as large as doing a calibration of polytomous data using a 2-PL or 3-PL model [23]. This model is used because the essay test instrument with the scoring polytomous. Aspects analyzed using the PCM model are (1) the item difficulty index, (2) the goodness of fit, (3) the information function and SEM, and (4) the ability estimation. The estimation of item parameters and the ability of test participants may be assissted by the use of computer software, asmaking iterative calculations can be very difficult if done manually. The analysis program is used to estimate the item and ability parameters [24].

The results of the analysis can be seen to know the characteristics of item and students ability. Items that meet the difficulty criteria of item *b* have an index value of difficulty between -2 to +2 logit scales. The goodness of fit based on INFIT MNSQ value. The test items fit with PCM model, if INFIT MNSQ was close to 1.0 and standard deviation close to 0.0 according established by Adams and Khoo [24]. The interpretation result of the students' ability has a value of -4 to +4 in logit scale. From the plot, the graph of test information function and student ability was obtained.

IV. Results and Discussion

A. The result of parameter estimation with Partial Credit Model

Item parameter estimation was done to get the information of test item which useful in assessments. Student response data obtained from assessment instrument in the form of essay test. The essay test consists of two test set, i.e. A and B test sets. Each test sets has ten items including two anchor items. The assessment instrument of A and B test sets was equivalent in measuring the ability of students' verbal and diagram representation ability. The parameter of item difficulty index in A and B test sets was homogeneous. The scoring model used was the PCM scoring in the polytomous form with four categories. The result of parameter estimation of item difficulty index *b* using the PCM model can be seen in Figure 1.

Based on Figure 1, it is known that the item difficulty index on the assessment instrument of representation ability has a value *b* in the range of -1.02 to +1.17. The item difficulty on the test are various, i.e. easy, average and hard levels. The item difficulty index *b* -1.02 was categorized as test item with easy level which near a value *b* -2 in logit scale. The test item with the lowest difficulty is 6th item. The item difficulty index *b* +1.17 was categorized as test item with difficult level which near a value *b* +2 in logit scale. The test item with the highest difficulty is 18th item. Each test item is in the good category because it has an item difficulty index *b* ranging from -2 to +2 [21]. None of the item test was in the very easy or very hard difficulty level. The item difficulty index has not passed the limit of meeting a good item difficulty criteria.

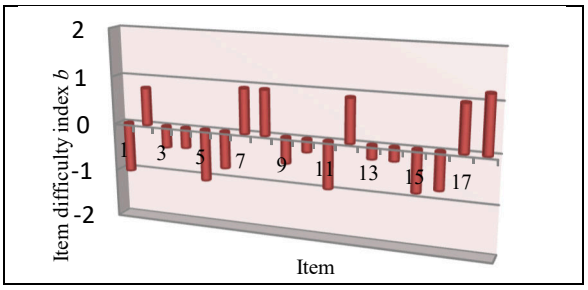

Fig. 1. The parameter estimation of item difficulty index b according to the PCM model.

B. Goodness of fit

Determining the goodness of fit with model is called the item calibration. The field testing providing the characteristics of the test instrument of representations abilities. The test instrument fits with the PCM model based on INFIT MNSQ and standard deviation. All test items fit with PCM model, as the INFIT MNSQ was close to 1.0 and standard deviation close to 0.0, according to the model established by Adams and Khoo [28]. Based on the goodness of fit on PCM model shown in Figure 2, the goodness of fit for each test item can be found. The INFIT MNSQ values of all test items are in the range of 0.80 to 1.17, with standard deviation close to 0.0. According to Adams and Khoo, all test items can be modelled with PCM in the IRT.

```
------------------------------------------------------------
Item Fit
all on all (N = 305 L = 18 Probability Level= .50)
------------------------------------------------------------
INFIT
 MNSQ          .56       .63       .71       .83      1.00
----------------+---------+---------+---------+---------+---
  1 item 1                                    .         *  |
  2 item 2                                    .    *       |
  3 item 3                                    .            |
  4 item 4                                    .            |  *
  5 item 5                                    .    *       |
  6 item 6                                    .       *    |
  7 item 7                                    .  *         |
  8 item 8                                    .     *      |
  9 item 9                                    .            |  *
 10 item 10                                   .            |
 11 item 11                                   .         *  |
 12 item 12                                   .       *    |
 13 item 13                                   .            |*
 14 item 14                                   .            |  *
 15 item 15                                   .  *         |
 16 item 16                                   .   *        |
 17 item 17                                   .       *    |
 18 item 18                                   .         *  |
============================================================
```

Fig. 2. Goodness of fit of Items with PCM model.

C. Test Information Function (TIF)

Based on the analysis result, there is a relationship between the ability with Test Information Function (TIF) and standard error of measurement (SEM). Figure 3 shows the relationship of TIF and SEM on assessment instruments of diagram and verbal representations. The curve of total information function is inversely proportional with standard error of measurement. The test information function (TIF) at the maximum point of 25 with standard error of measurement (SEM) of 0.06. The maximum information function provided from the test is within the ability (θ) of -1.2 in the logit scale. This means that the test instrument provides high information on students' representation ability and error of measurement. The test is appropriate and provides maximum information if done by students with average ability. The analysis result of total information function has an information function of 10 and SEM of 0.21. This test is suitable for students with moderate ability in the range -2.30 $\leqslant$ θ $\leqslant$ +2.18. Students who perform the test are at low to high ability.

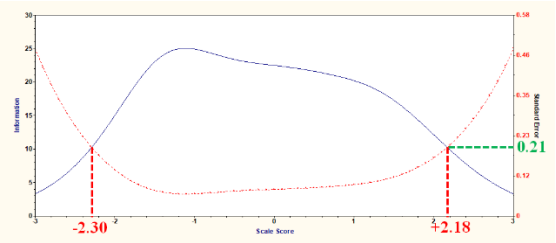

Fig. 3. Test information function (TIF).

D. The result of ability estimation

Based on the result of the students ability estimate (θ) as shown as figure 4, it is known that the students' verbal and diagram representation ability in the learning Physics produces the score distribution between -2.08 to +2.10. The students' representation ability is in the range of -4 to +4 in scale logit. The average result of student representation ability is 0.00. Based on the average result of the students' representation ability in learning Physics, it is known that the students have the ability of representation at the average level.

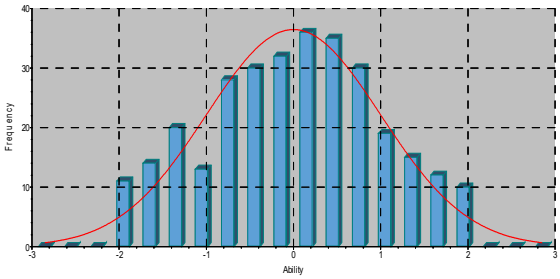

Fig. 4. Distribution ability estimation with PCM model.

E. Discussion

Assessment in the Physics learning is very important to know the achievement of learning competence.Therefore, it is important to know the characteristics of assessment instruments. This study aims to determine the characteristic assessment of the ability of students' diagrams and verbal representation. The ability of diagram and verbal representations can use the assessment instrument in the form of an essay test. Scoring on the description test was usually done partially based on steps to be taken to correctly answer a point. The scoring of student responses to the classical test theory is done

by summing the overall score obtained by the students. Such scoring is not entirely appropriate. This is because the level of difficulty of each step is not taken into account. In addition, the likelihood of someone answering a particular matter correctly is also unpredictable [23]. For this reason, another approach is needed such as the item response theory (IRT) approach. IRT overcomes the weakness in the classical test theory where the IRT model has an independent between item and ability parameters [26]. According to this rule, item parameters (difficulty, discrimination and guessing) are not dependent on the distribution of students ability in test so as ability parameters independent on a specific set of test items. Test item with more than two response options can be modelled with polytomous IRT models. The model used is an extension of the Rasch model called the Partial Credit Model (PCM) [29]. PCM is a scoring model of polytomous. The assumption on PCM is that each item has the same discrimination.

The test consists of two test sets i.e. A and B test set. Each question in the test set consists of 10 items containing two anchor items. Student response data are polytomous data with 4 categories. This assessment instrument has passed the validation process by experts. The assessment instrument was tried on subject of test. The data to be used in this study is primary data derived from the assessment of the ability of representation diagrams and verbal students given to students of XI IPA class in senior high school. The response data analyzed using PCM model according to IRT approach. The results showed that (1) all items have good item difficulty; (2) all of test items fit with PCM model; (3) there is a correlation between the ability to test information function (TIF) and standard error of measurement (SEM); and (4) the estimation of student ability is diverse.

The results of this study were in line with some previous research. Supahar examines the estimation of inquiry performance items which analyzed based on PCM model [30]. The result is the parameter estimate of the degree of difficulty of questions *b* in range -2 to +2 so test item fulfill the criteria of a good test item. The difference with the research that has been done is the ability of students who are measured. This study measures the ability of students' verbal and diagrams representation. There are some similarities found in the study. The result of student's response is the data of polytomous with category four which is analyzed using PCM model. The important thing is the result of the parameter estimation of this item has in common, where the index of difficulty point *b* about the ability of the representation of diagrams and verbs has a value in range -1.02 to +1.17. The whole item has a difficulty index (*bi*) ranging from -2 to +2, thus there is no item difficulty index which passes the boundary so as to satisfy the criteria of a good item [21].

According to the research conducted by Istiyono, the test item with polytomous data response checked the goodness of fit against the PCM model in the process of developing the instrument [31]. The goodness of fit to the PCM model is obtained by looking at the mean value of INFIT MNSQ. This and the standard deviation are developed by Adam and Khoo [28]. If the mean of INFIT MNSQ is about 1.0, and the standard deviation is 0.0, the whole test fits with the model. The similarity with the research conducted is the same using PCM model. The result shows the INFITMNSQ value in range 0.80 to 1.17 and standard deviation of 0.12, which means whole item fit with PCM 1 PL model. In contrast to previous research, the Maydeu-Olivares study was concerned the goodness of fit assessment of item response theory models with an asymptotic chi-square distribution [32].

The function of test information relates to the strength of the contribution to each item on the test. The function of the test information was used to discovering the latent trait measured by the test. The test information function (TIF) is the sum of item information function at a given ability level. The research that Supahar and Zuhdan have done in the development of assessment instruments is also concerned with the function of test information [33]. The assessment instrument on high school Physics subjects is exactly tested on the respondents with the ability of between -2.5 to +2.5. This assessment instrument will provide maximum information on students' abilities and low error rates when tested to students who have the ability of -1 in logit scale. The same is also found in the study, where the function of the test information provides the maximum information of 25 with error measurement of 0.06 if done by students with ability of -1.2. This test is suitable for test takers with moderate category ability in range -2.30 $\leqslant \theta \leqslant$ +2.18. Other findings in both studies are the best test information function with measurement error. The higher test information function (TIF) will result in a lower measurement error (SEM) which means the more accurate the scoring model is in estimating ability. High-ability students have a wider understanding so that the possibility of solving the problem is greater [34].

Retnawati describes the estimation of the ability of the test participants with the scoring of polytomous on the Generalized Partial Credit Model (GPCM) model following the IRT approach [35]. GPCM is an extension of the PCM model [24]. When estimating participants' abilities, the difficulty level at each stage of problem solving needs to be taken into account. Unlike the case with the assumption on PCM, in which each item has the same power difference. The results show the similarity between the distributions of student abilities with parameters θ at intervals -4 to +4 logit scale. In order to be exploited for a better interpretation, it is necessary to proceed with the usual linear transformation. The test items provide information on the latent trait (ability) which fit with the given item's difficulty level.

V. CONCLUSION

The conclusion of the research on the characteristics of assessment instrument is as follows: 1) all test items fit with the PCM model; 2) the item difficulty index ranged from -1.02 to +1.17, meaning that the test items can be categorized as good; 3) the information function is 10 and SEM is 0.21; and 4) the ability estimation ranged from -2.08 to +2.10, indicating that the student abilities are varied. This assessment instrument is appropriate for measuring the diagram and verbal representation skill on students with average ability. The characteristics of test items with IRT model facilitate the development of assessment instruments that provides useful data for both descriptive and parametric statistics.

Further research is needed on the characteristics of the rating instrument with GPCM 2-PL and 3-PL model which estimates the difficulty parameter of item *b*, discrimination of item *a*, and guessing. The number of respondents used should also be wider, so that more can describe the characteristics of the test and the ability of students. Finally, it is recommended that teachers validate the characteristics of the assessment instrument to measure students' abilities.

ACKNOWLEDGMENT

The authors thank the Directorate of Research and Community Service for funding this study through the *Penelitian Tim Pascasarjana* (*PTP*) 2018 program.

REFERENCES

[1] Mardapi D 1999 *Pengukuran Penilaian dan Evaluasi* (Yogyakarta: Nuha Medika).

[2] Widoyoko 2014 *Penilaian Hasil Pembelajaran di Sekolah* (Yogyakarta: Pustaka Pelajar).

[3] Schunk D H 2011 *Learning Theories : an Educational Perspective* (Boston: Pearson Education).

[4] Mardapi D 2008 *Teknik Penyusunan Instrumen Tes dan Nontes* (Yogjakarta: Mitra Cendikia).

[5] Cohen R J and Mark E S 2010 *Psychological Testing and Assessment: An Introduction to Test and Measurement* (New York: McGraw-Hill Company).

[6] Lissitz W R and Samuelsen K 2007 *J. Educ. Res.* **36** 437–448.

[7] Zitzewitz P W 2011 *The Handy Physics Answer Book* (Canton: Visible Ink Press).

[8] Halliday D, Resnick R and Waker J 2001 *Fundamentals of Physics* (New York: John Wiley).

[9] Docktor J L, Strand N E, Mestre J P and Ross B H 2015 *Phys. Rev. Phys. Educ. Res.* **11** 1–13.

[10] Jonassen D H 2011 *Learning to Solve Problems: A Handbook for Designing Problem Solving Learning Environments* (New York: Routledge).

[11] Rosengrant D E, Etkina and Van Heuvelen A *Phys. Educ. Res. Conf.* **883** 149–153.

[12] Kalathil R and Sherin M G 2000 *Fourth Int. Conf. Learn. Sci* (Mahwah, NJ: Erlbaum) p 27–28.

[13] Mansyur J 2015 *Int. Educ. Stud.* **8** 1–13.

[14] Arslan A S and Kurnaz M A 2014 *Procedia-Social Behav. Sci.* **116** 627–632.

[15] Berge M and Weilenmann A 2014 *Eur. J. Eng. Educ.* **39** 601–615.

[16] Rosengrant D, Van Heuvelen A and Etkina E 2009 *Phys. Rev. Phys. Educ. Res.* **5** 1–13.

[17] Maries A and Singh C 2017 *Eur. J. Phys.* **39** 1–19.

[18] Meltzer D E 2002 *Conf. on Ontological, Epistemological, Linguistic and Pedagogical Considerations of Language and Science Literacy* (Canada: University of Victoria) p 1–18.

[19] Van Der Meij and De Jong T 2003 *European Association for Research on Learning and Instruction* (Pandova, Italy) p 1–17.

[20] Setyani N D and Handhika J 2017 *Int. J. Sci. Appl. Sci.* **1** 162–169.

[21] Hambleton R K and Swaminathan H 1985 *Item Response Theory* (Boston: Kluwer Nijhoff Publishing).

[22] Demars C E 2010 *Item response theory* (New York: Oxford University Press).

[23] Keeves J P and Alagumalai S 1999 *Measurement in Eduactional Research and Assessment*, (Amsterdam: Pegamon).

[24] Muraki E and Bock R D 1997 *Parscale 3: IRT Based Test Scoring and Item Analysics for Graded Items and Ratting Scales* (Chicago: Scientific Software Inc).

[25] Retnawati H 2014 *Teori Respons Butir dan Penerapannya: untuk Peneliti, Praktisi Pengukuran dan Pengujian, Mahasiswa Pascasarjana* (Yogyakarta: Nuha Medika).

[26] Hambleton R K, Swaminathan H and Rogers H J 1991 *Fundamentals of Item Response Theory* (United Kingdom: Sage Publications Inc)

[27] Seon H S 2009 *Pract. Assessment, Res. Eval.* 14 1–8

[28] Adams R J and Khoo S T 1996 *Quest: The Interactive Test Analysis System Version 2.1*. (Camberwell, Vitcoria: The Australian Council for Educational Research).

[29] Masters G N 1982 *Psychometrika* **47** 149–174.

[30] Supahar 2014 *Proc. of Int. Conf. on Research, Implementation and Education of Mathematics and Sciences* (Yogyakarta State University) p 137–144.

[31] Istiyono E, Mardapi D and Suparno *J. Penelit. dan Eval. Pendidik.* **16** 2.

[32] Maydeu A 2013 *Measurement* **11** 71–101.

[33] Supahar and Prasetyo Z K 2015 *J. Penelit. dan Eval. Pendidik.* **19** 104–107.

[34] Ramos J L S, Dolipas B B and Villamor B B 2013 *Int. J. Innov. Interdiscip. Res.* **4** 48–60.

[35] Retnawati H 2011 *Pros. Sem. Nas. Pendidikan dan Penerapan MIPA* (Yogyakarta State University) p 53–62