P-ISSN: 1693-1246
E-ISSN: 2355-3812
June 2021

Jurnal Pendidikan Fisika Indonesia 17 (1) (2021) 60-71
DOI: 10.15294/jpfi.v17i1.27069

http://journal.unnes.ac.id/nju/index.php/jpfi

Developing An Achievement Test in Learning of Physics for Assessment

A. Dhien[1]*, Kumaidi[2]

[1]Program Studi Sistem Informasi STMIK PPKIA Tarakanita Rahmawati, Indonesia
[2]Universitas Muhammadiyah Surakarta, Indonesia

Received: 08 March 2021. Accepted: 03 May 2021. Published: June 2021

Abstract

The aims of the study are to know the advantage of achievement test in learning physics developed by utilizing item specifications and the ability distribution of tenth grade students at senior high school in Yogyakarta. The benefits of the research are to prepare test items, provide guidelines for teachers in writing quality items and produce learning achievement test that can be used on School Exams. This research is developmental research used the model suggested by Oriondo&Dallo-Antonio. The subjects were tenth grade students. The limited test was tried out to 354 students in school with categories of high, medium and low. The wide scale test was tried out to 644 students on package A and package B. Item analysis was conducted by using qualitative and quantitative analysis. This research shows developed test in physics by utilizing item specifications had the advantage in producing more equate test the ability distributions of tenth grade students at senior high school in Yogyakarta were mostly at the average category.

Keywords: *developing, achievement test, item specifications, physics*

INTRODUCTION

Teachers in the current era of the industrial revolution 4.0 are not only required to provide learning material, but also are required to be able to prepare students to become competent and quality graduates. This condition can only be achieved if the quality of education is continuously improved. Learning activities and assessment systems are one of the efforts to improve the quality of education. Learning is an activity that contains learning objectives, the teaching-learning process and the evaluation of learning outcomes. Learning objectives lead to assessments made by the teacher. The teaching and learning process in its implementation refers to the learning objectives and the evaluation of learning outcomes. These results are then used by the teacher to assess the extent to which the teaching and learning process has been carried out in accordance with the objectives of learning.

In carrying out their duties, a teacher must be able to assess learning outcomes, one of which is by preparing tests when starting and ending the teaching and learning process. This is in line with the opinion of Schunk (2012) that one way that can be used to find out the learning outcomes of students is through assessment. The test is a form of instrument used by educators to take measurements and consists of a number of questions that have true or false answers, or all are true or partially correct. The purpose of conducting the test is to determine the extent to which learning achievements or competencies have been achieved by students (Mardapi, 2012). The test results are information on the characteristics of a person or group of people. Characteristics can be cognitive abilities or skills. A test can be said to be good if it is able to measure what should be measured and can provide reliable results.

Assessment involves a formal effort to determine the achievement of learners. At school, what is most often used as a reference for assessment results is student achievement in the fields of: reading, writing, mathematics, science, and social sciences. The learning achievement test is the result of a student's work or study effort which shows a measure of the skills achieved in the form of grades (Wirantasa, 2017). Accountability that often leads to testing becomes an assessment tool that includes many measurement procedures in addition to testing to assess learning outcomes through written tests. Learning is often assessed in

*Correspondence Address:
E-mail: dhienastrini@gmail.com

tests based on students' written responses (Schunk, 2012). According to Sumintono & Widhiarso (2015) test results can be used by teachers to determine (1) the ability of students relative to other students in the same test; (2) shows the development of students' abilities over a period of time in knowledge and skills; (3) shows evidence of understanding of a particular subject matter, knowledge or idea; (4) predict the performance of students.

The ability of students to master certain learning materials at a level of education as a whole is expected to be in accordance with the predetermined Graduate Competency Standards (SKL). The results of competency attainment are expected to explain the actual abilities of students, one of which is learning physics.

Physics is a part of science that studies natural phenomena and their interactions. Physics learning is learning that does not ignore the nature of physics as a science (Prihatni, Kumaidi, & Mundilarto, 2016). One of the goals of the Physics Subject in SMA is that students are able to develop reasoning skills in thinking inductive and deductive analysis by using physics concepts and principles to explain various natural events and solve problems, both qualitatively and quantitatively (Istiyono, Mardapi, & Suparno, 2014). Another opinion was also expressed by Murdaka & Kuntoro (2013) that physics has characteristics regarding natural sciences which are fundamental and universal.

The laws of physics can be displayed in several representations, namely in the form of sentences, mathematical equations, graphs, and experimental data. This representation can be an obstacle for some students who have a tendency to certain representations. On the other hand, physics subjects require an understanding of the concepts that can be conveyed through representational diagrams. Especially for students of class X Senior High School (SMA) which is a transition from junior high school this is a big challenge. This means, in the preparation of test items it is necessary to accommodate the characteristics of the problems faced by students, so that they can reveal students' real understanding.

Based on the results of interviews conducted by several state high school physics subject teachers in the city of Yogyakarta, it was found that the resulting items first began with making the item questions, then the question grid was then compiled. In assembling the test questions, the item specifications have not been used. This is of course not in accordance with the rules of writing the questions and the characteristics of each item are not clearly known. In addition, it was also found that some of the test items were not fully made by the teacher. Most of them only took from books available in schools, so it was feared that they were not well calibrated.

Another problem is obtained at the State Senior High School in Yogyakarta city by looking at the items in the Mid-Semester Examination (UTS) and Semester Final Examination (UAS). Several items were found to have completed pictures, but the presentation did not clearly reveal the physics problems that were asked for in the items. This of course will have an impact on the quality of the test, where if the items used are not good it can cause a bigger measurement error or in other terms the Standard Error of Measurement (SEM) becomes significantly large. Based on the field data obtained, it is important to review the items, both qualitatively and quantitatively, so that the characteristics of the good items can be identified. As for the items that are not good, they must be repaired or replaced. Thus, it is hoped that the development of the test can affect the skills of teachers to produce quality items.

The results of Megawati's (2012) study suggest that there are two weaknesses of educators in measuring using tests. First, the test equipment that is made each time will perform an assessment which is not tested statistically, especially on the type of multiple choice questions. This is because it does not go through trials to test validity and reliability, including testing the difference power and effectiveness of a distractor on the type of multiple choice questions. In addition, the validity of the constructs still needs to be questioned because the preparation is improper and less planned. Second, if you want to obtain a really good quality test, it certainly takes a lot of time to make an assessment. Item Response Theory (IRT) or what is commonly called Modern Test Theory is a review of the items using the item answer theory. This theory uses a mathematical function to connect the odds of answering a scale correctly with students' abilities (Fatkhudin, Surarso, & Subagio, 2014).

The use of logistic functions in IRT not only allows estimation of item parameters and capability parameters but also allows consideration of the accuracy of each parameter estimated (Umar, 1999). According to Russell & Airasian (2012) Stanine is the second type of standardized test score. Stanine uses a scale of nine, with each stanine representing 1 lowest performer and 9 highest. A common score that is often used is the percentile rank. The percentile ratings range from 1 to 99 and show the percentage of the student cohort as a reference. Stanine (Standard nine) is designed to show student performance.

Based on the facts presented, it can be concluded that the item specification is a specific guideline used by the teacher in writing the item which consists of the requirements for graduate skills, indicators, scope of material, form of questions and sample items to produce test items that match the indicators. Information on the ability of students needs to be known clearly through the learning achievements obtained. If the accuracy of the information obtained is greater, the measurement error will be smaller so that the items used can actually measure students' abilities at a certain level. The purpose of this study was to determine the advantages of learning achievement tests for physics subjects that have been developed by utilizing item specifications and to obtain information on the distribution of abilities of class X SMA students in Yogyakarta. The benefits of this study are to prepare test items that will then be made, provide guidelines for teachers in writing quality question items and produce learning achievement tests that can be used on school level exams.

METHOD

This research is a development research about tests with reference to the Oriondo & Dallo-Antonio model. The first step of this research is the design of the test which includes: formulation of physics material for class X, preparation of test item guidelines, preparation of item specifications, review of items, writing of items in parallel packages. The second step is research trials and the third stage is research trials.

The procedure for developing a learning achievement test for physics subjects used the test development step according to Oriondo & Dallo-Antonio (1998). The stages in developing this test are based on a systematic procedure, namely:

(a) Setting goals. The purpose of the measurements taken to produce the test;

(b) Prepare a specification table. The specification table based on the formulation of the specified material then determines the percentage of the number of grains produced;

(c) Select an appropriate item format. The item format used in this study was multiple choice;

(d) Writing test items. The writing of the test items was carried out based on the question preparation guide;

(e) Editing test items. Prior to testing, an item was edited so that there was no mistake in the concept of the question;

(f) Conducting test trials. The test items that have gone through the editing stage are then tested;

(g) Preparing for the test. The final form of the test includes items that meet the criteria for good item characteristics;

(h) Determine the validity test. The validity of the test is determined based on the content (topic);

(i) Determine the reliability test. The development of the test required the consistency of a test to be used so that a high reliability coefficient was needed;

(j) Interpreting the test. Test interpretation is the final stage in the research undertaken.

The test subjects in this study were students of class X SMA in the city of Yogyakarta. The research trial subjects were obtained based on the rankings of the results of the National Examination (UN) SMA which had an average UN score including high, medium and low categories. The average ranking of UN results is seen based on data from the Education Office in the city of Yogyakarta. The research sample was taken by using stratified random sampling technique. The trial consisted of 354 students and the broad-scale test consisted of 644 students.

The research data was obtained through the answer sheets that the test takers worked on in the parallel test package, namely package A and package B. The data collection instruments were in the form of several instruments used in the study, namely the physics learning achievement test guide, the learning achievement test in the form of multiple choices with five alternative answers as many as 40 items for package A and package B, test instrument assessment sheets, test review sheets and test answer sheets. Data collection techniques in this study through tests and documentation during testing on limited trials and large-scale research trials in SMA Yogyakarta.

The data analysis technique in this study used qualitative and quantitative analysis. The qualitative analysis was carried out through the study sheet, the learning achievement test was carried out before the test was tested. The test items that have been reviewed are then corrected based on the input and suggestions for improvement given by the validator.

Validity is the ability of the instrument to measure what the constructed constructs should measure (Elvira & Sainuddin, 2020). The quantitative analysis was carried out in 2 (two) stages, namely First, reviewing the suitability of the items based on the content through the assessment of ten validators to determine the extent to which the items represented the construct being measured. As disclosed by Azwar (2014), items are said to represent the construct being measured, meaning that the item is relevant to the indicator. This is because the indicator is an operational translation of the attributes being measured. Assessment is carried out by giving a number between 1 and 4. An analysis that can be used to determine the index of the content validity of the test instrument is Aiken's formula (Aiken, 1980).

$$V = \frac{\sum s}{[n(c-1)]} \qquad (1)$$

Based on the results of the analysis, the items that meet the content validity can be determined. Items are declared valid if they meet the criteria V> 0.73 for 10 assessors with a rating category of 4 scales (Aiken, 1985). Second, empirical analysis uses the item response theory (IRT) approach through the BILOG-MG program. Empirical analysis was carried out after the questions were tested. Quantitative analysis was carried out twice, namely limited-scale and large-scale group research trials.

RESULTS AND DISSCUSSION

The development of learning achievement tests for physics subjects is carried out based on the specifications of the items arranged in reference to two Competency Standards (SK), Basic Competencies (KD) and the indicators are further developed into 40 question indicators contained in the physics learning achievement test grid for class X Optical Material and heat. The test is equipped with a manual for specifying items.

Guidelines and item specifications are reviewed qualitatively and then used as guidelines for question writers to produce a product in the form of two parallel test packages. The parallel test consists of package A and package B. The resulting product is then tested on class X high school students in the city of Yogyakarta. The items in each test package consist of 40 multiple choice questions. Package A and package B were done by test takers who were designated as subjects in the study.

The results of the review of the items showed that the test had met the criteria for the review of the item, namely 75% to 100%. Thus, most of the items had met the criteria for the aspects studied theoretically. Guidelines for test items and specification of the resulting items are first analyzed quantitatively. Good question items certainly fulfill the validity of the content, for that the questions that have been written must be validated by experts in the field, or by allied subject teachers (Rusilowati, 2015).

The quantitative analysis was carried out by measurement experts, physics education experts and eight physics teachers at Yogyakarta City State Senior High School. The validator provides an assessment of the learning achievement test instrument for physics subjects. The items are checked and matched to see whether the items reflect the predetermined indicator domain. The test instrument assessment sheet is one part of the content validity to assemble the test. Content validity aims to ensure whether the items are appropriate and relevant to the question preparation guidelines, learning objectives and learning indicators that you want to achieve (Elvira & Hadi, 2016).

Test assessment can be done by looking at the suitability of the item indicators with the aim of developing the instrument, the suitability of the indicators with the material coverage or the suitability of the theory, the suitability of the instruments and the item indicators; analyzing the truth of the concept of question items, examining the truth of the content, the key truths of the test, language and culture (Retnawati, 2016).

The effectiveness of the item distractor must be considered in producing a quality test. A good distractor is a distractor chosen by a few low-ability students, but the number of students is significant. Problems that are too difficult can also cause the distractor to not function or there is a misunderstanding of concepts in students (Elvira & Hadi, 2016). In this case, Package A and Package B were designed by several authors with attention to the functioning of the fraudsters. Each answer option is obtained based on the concept of the correct equation to the problem presented. However, in this study, a distractor was made by considering several possible ways in which students performed equation operations. This is done so that if students perform an incorrect equation operation, they will choose the answer for the distractor.

The results of the analysis showed that there were 33 items that were declared valid, namely V> 0.73. This score is interpreted as a fairly high coefficient for this item. This means that these items have good content validity and support the validity of the overall test content (Azwar, 2014). The 7 items stated that they did not meet the content validity were presented in Table 1.

Table 1.The results of the calculation of aiken's index formula

Item number	V
5	0.6
6	0.67
7	0.6
8	0.6
33	0.7
35	0.7
40	0.7

The reliability of test results scores is important information in the analysis of test items or test development (Sarwiningsih, 2017). The reliability index is a measure of the consistency of a test score from one measurement to another. Reliability also describes the consistency of scores obtained by the same test taker when given the same test at different times (Purnama & Alfarisa, 2020). The output results in phase 3 of BILOG-MG 3.0 model 2 Logistic Parameters (PL) show that package A has a reliability coefficient of 0.860 and package B has a reliability coefficient of 0.826. The correlation ranges between r = 0.81 for various and inhomogeneous items including the alignment of the answers on each item (Anderson, Irvin, Alonzo, & Tindal, 2015).

The results of the trial analysis showed that the model that produced the most good items was the 2-PL model in package A and package B. The criteria for good items according to the 2-PL model were based on distinguishing power and grain difficulty level. Grain differentiation is said to be very high if it ranges from 0-2. The difficulty level of a good item ranges from -2 logit ≤ bi ≤ 2 logit. Values that are close to -2 logit indicate that the items are getting easier and values that are closer to +2 logit indicate that the items are getting more difficult (Amelia & Kriswantoro, 2017).

The results of the analysis output using BILOG-MG show that package A has good grain characteristics according to model 2 with 35 logistical parameters and 39 items for package B, then items that do not meet the criteria are subject to product revision. Product revision is not done by removing bad items but by improving grammar in the question sentences. The product revision in package A contains 4 items and package B has 1 item. The three assumptions underlying the item response theory are unidimensional, local independence, and parameter invariance.

Unidimensional item requirements are intended to maintain invariance in IRT. If a test item measures more than one dimension, then the answer to that item is a combination of the various abilities of the participants (Kriswantoro, Amelia, & Irwanto, 2016). Unidimensional states that each test item is only able to measure one ability. In practice, this assumption is very difficult to fulfill because of the many factors that influence the test. One way to test this assumption is by performing factor analyzes that yield KMO, eigenvalues, explainable variance and factor components.

The results of the analysis also contain the dominant factors measured in test package A and graphs that contain steep or sloping lines. Package A has a total variance based on the quotient between the main factor and the tenth factor, a percentage of 48% of the total variance is obtained where the graph only shows two steeples while the other graph is in the form of a slope. This means that there are two dominant factors measured in the test. Based on the exploratory analysis, it is found that the instrument in the form of a learning achievement test kit can be said to be valid for measuring physical abilities and is proven empirically.

Measuring tool that produces a valid score, then the measuring tool is reliable. Therefore, to be able to make the right instrument for its use, a valid and reliable instrument is needed in the model test analysis (Elvira & Sainuddin, 2020). The results of the analysis can be seen in the scree plot presented in Figure 1.

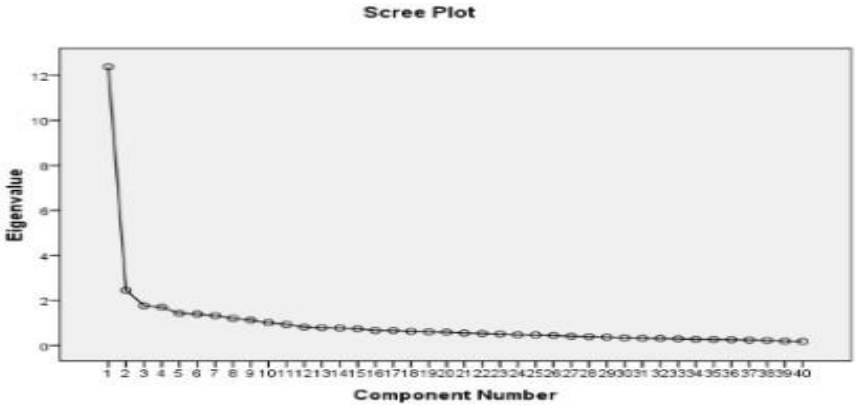

Figure 1. Scree Plot of Package A Exploratory Factor Analysis Results

Package B has a total variance based on the main factor quotient with the tenth factor obtained by 32.4% of the total variance greater than 20%. Thus, it can be said that based on the components of the variance results, the results of the factor analysis have met the unidimensional assumptions. In addition, it can be seen that the scree plot is presented in Figure 2 which shows the eigenvalues that are starting to slope in the third factor. This shows that there is 1 dominant factor in the physics learning achievement test and other factors contribute significantly to the variance that can be explained. The results of the unidimensional analysis according to Retnawati (2016), the function of the test equipment measures at least 2 factors with the first factor being the dominant factor.

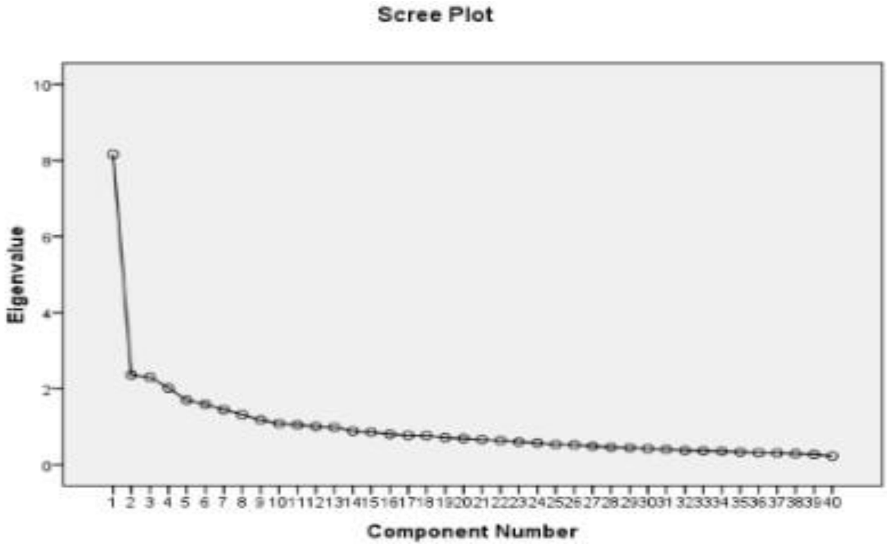

Figure 2 .Scree Plot of Analysis Results of Exploratory Factors Package B

Local independence is met if the covariance in each group approaches zero. The local independence test aims to see whether the abilities of participants in the same sub-population are independent of items (Elvira & Hadi, 2016). Local independence is intended as a location at a point on the participant's characteristic continuum. In practice, points on the participant's continuum can be in the form of intervals and within points or within the interval of the participant's characteristic parameters to a homogeneous sub-population (Kriswantoro, Amelia, & Irwanto, 2016).

Based on the results of the analysis using Microsoft Excel, it shows that package A and package B are proven to fulfill the local independence assumption with the covariance value in each group close to zero. Thus, it means that there is no correlation for each item in the group.

Parameter invariance is divided into two, namely item parameter invariance and student parameter invariance. The item parameter invariance test aims to determine whether the item characteristics will change even though it is done by students with different abilities, while the student parameter invariance test aims to observe whether the student's ability changes if given questions with different levels of difficulty and different power (Elvira & Hadi, 2016).

Based on the results of the analysis of grain parameter estimates, it shows that the mapping forms a diagonal line. In the presentation of Figure 3, it can be seen that some points on the diagram are not on the diagonal line, this indicates that there is an indication of grain parameter invariance. This means that the characteristics will change if done by students with different abilities.

The invariance of the item difference parameter of package A with a coefficient of determination of 0.684. This reflects that the item difference has a sufficient correlation of 68.4% to the level of information provided to be able to determine the level of the respondent's ability. Therefore, the value of a is an indicator of how much an item provides information about the test taker's ability level (Kriswantoro, 2016).

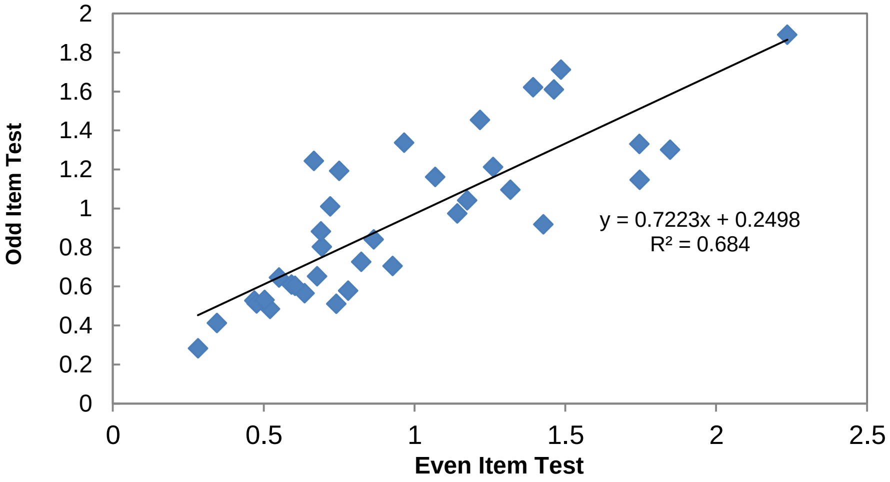

Figure 3. Invariance of Item Package A Distinguishing Power Parameters

Figure 4 shows that the points on the diagram are relatively close to the diagonal lines. This informs that there is no variation in the estimated parameters for the odd and even student groups so that it is proven that the item parameter invariance in package A is fulfilled. The invariance of the parameter of the difficulty level of the package A items has a determination coefficient of 0.6087. This means that the level of difficulty of the items has a sufficient correlation of 60.87% to the level of information provided to be able to determine the level of the respondent's ability. The results of the item parameter analysis in package B required testing the assumption of item differentiation parameter that can be done by separating the two groups of test participants, namely even and odd, using the help of Microsoft Excel so that the correlation is presented in Figure 5.

Based on the presentation in Figures 5 and 6, it can be seen that the mapping results form a diagonal line, however, some points on the diagram are not on the diagonal line, this indicates that there is an indication of grain parameter invariance. This means that the characteristics will change if done by students with different abilities. The invariance of package B grain difference power parameter with determination coefficient of 0.6651 and 0.8627. This reflects that the difference between the items and the difficulty level of the items have a sufficient correlation of 66.51% and 86.27% to the level of information provided to be able to determine the level of the respondent's ability. The invariance of the ability parameters is proven by grouping the even and odd items done by all test participants. Furthermore, a scatter diagram is made, then the proximity is compared to the slope line 1 (Retnawati, 2014). The analysis of the estimation of the ability parameters is obtained based on the results of BILOG-MG in phase 3 which contains the ability (selanjutnya) which is then calculated with the help of Microsoft Excel. Overall, the estimation of the test takers' ability who worked on package A as in Figures 7 and 8 is in the form of a scatter diagram.

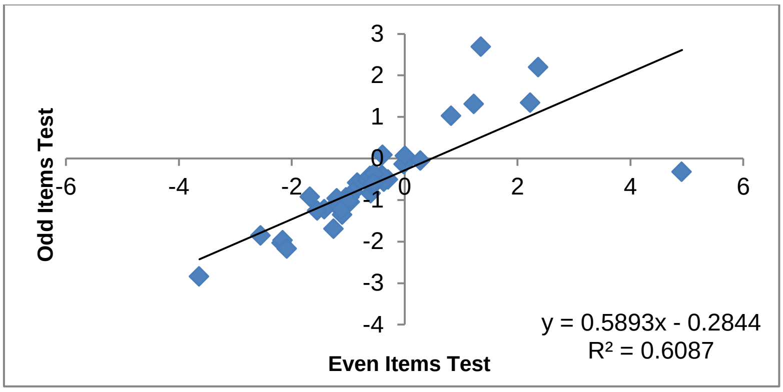

Figure 4. Invariance of Item Hardness Level Parameters of Package A

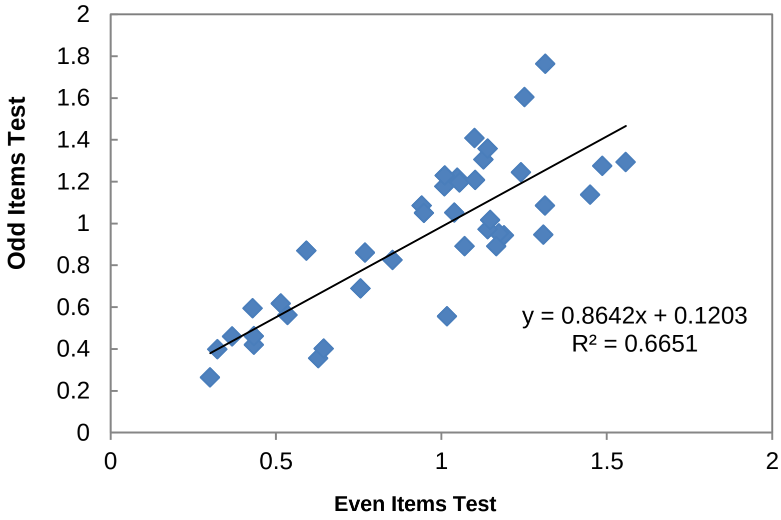

Figure 5. Invariance of Package Item Distinguishing Power Parameters. B

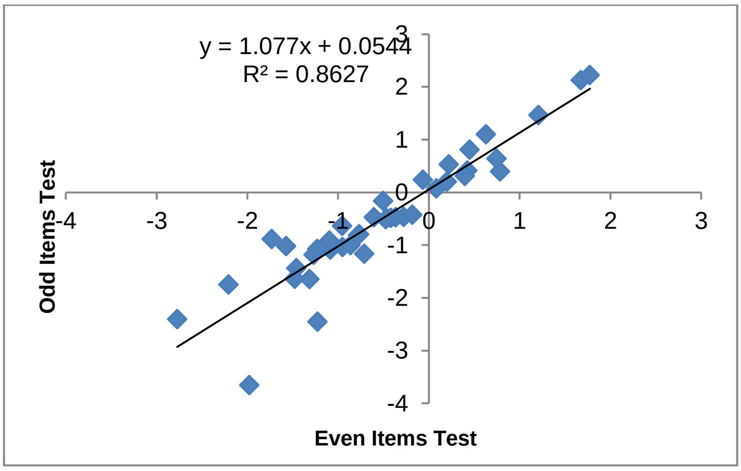

Figure 6. Invariance of Item Package B difficulty parameters

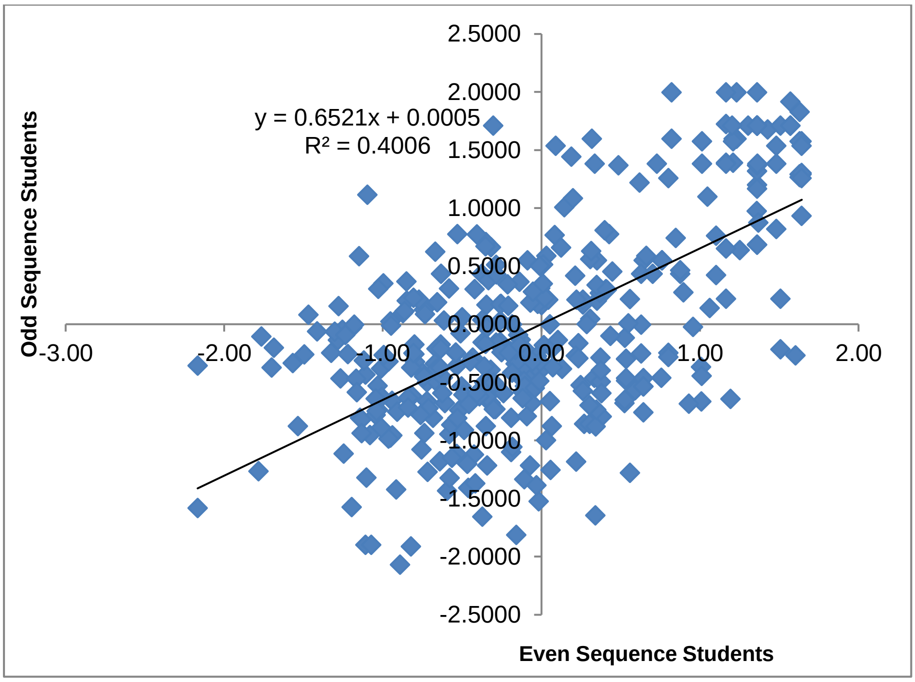

Figure 7. Invariance of Package A Capability Parameters

Based on the results of the analysis, it is found that package A has been proven to meet the invariance of the ability parameters can be seen in Figure 7 which shows the points on the scatter diagram are relatively close to the diagonal line, meaning that the assumption test of the student's parameter invariance is fulfilled. The coefficient of determination obtained is 0.4006. This reflects that 40.06% is sufficiently correlated with the estimated ability of test takers in the odd group and test takers in the even group.

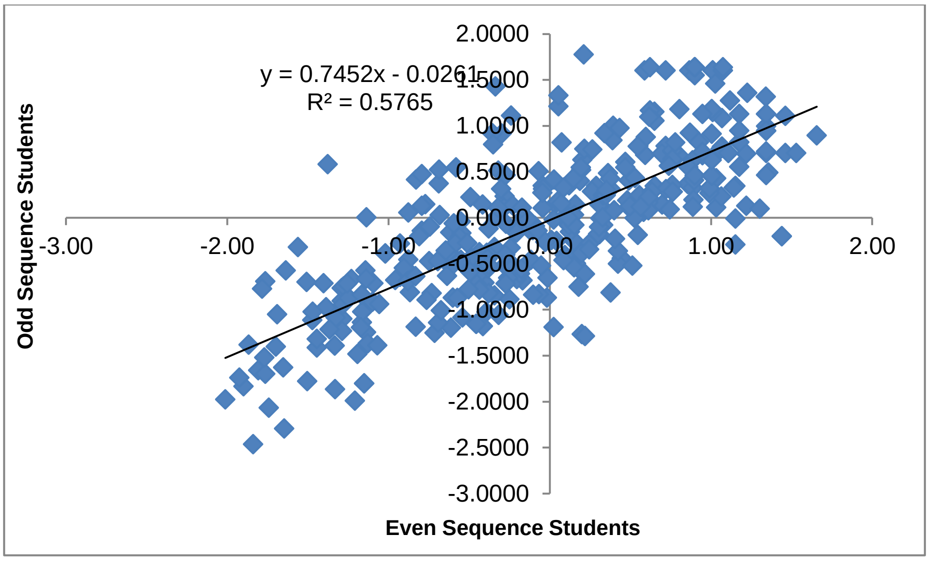

Figure 8. Invariance of Package B Capability Parameters

Based on the results of the analysis, it is found that package B has been proven to meet the invariance of the ability parameters can be seen in Figure 8 which shows the points on the scatter diagram approaching the diagonal line. The coefficient of determination of 0.5765 means that the correlation is sufficient to estimate the ability of test takers in the odd group and test participants in the even group with a proportion of 57.65%.

Packages A and B are proven to fulfill the three assumptions in the item response theory. The results of the analysis of the 2 logistic parameter model obtained that 35 items of test items that met the criteria for item differentiation, the level of difficulty and suitability of the items against the

Package A model and 36 items in Package B. In estimating the parameters, a logistic model is used with the highest number of fit items. Item fit is an item with a calculated Chi-Square value smaller than the table Chi-Square value or a p value above 5%. Goodness of fit in the test aims to determine whether the items used are in accordance with the model applied (Purnama, 2017).

The results of the analysis show that the average level of difficulty of the items in the learning achievement test items in high school physics subjects in the city of Yogyakarta in package A is -0.712 and package B is -0.514. The average distinguishing power of the items on the learning achievement test for high school physics subjects in the city of Yogyakarta in package A was 0.956 and package B was 0.938. The test information function measures the strength of the test in determining the ability of students and plays a role in assessing the condition of an item that is functioning optimally or not optimally. This is of course a further reference regarding items that are feasible or not worthy of being included in the test (Rukli, 2016).

Based on the results of the item characteristic analysis using the BILOG-MG program in phase 2 for model 2. Logistic parameters obtained different power (ai) and item difficulty level (bi) which are then entered into Microsoft Excel to calculate the value of the information function of each item. The information function in package A.

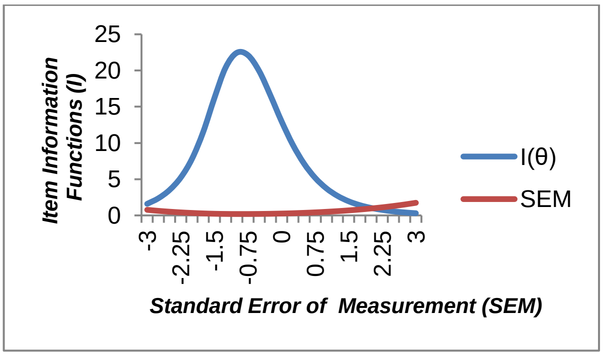

Figure 9. Curve Information Function of Package A Test

The highest test information function in package A lies in the ability (θ) = -1 with the highest test information function value of 22.4 (SEM = 0.211). This means that the physics learning achievement test in package A is suitable for test participants with moderate and low abilities. Based on the results of the analysis, it is found that the highest test information function lies in the ability (θ) = -0.75 with the highest test information function value of 15.6 (SEM = 0.252). The physics learning achievement test in Package B is suitable for moderate and low ability test takers. It can be concluded that the two test packages provide information on ability that is almost close.

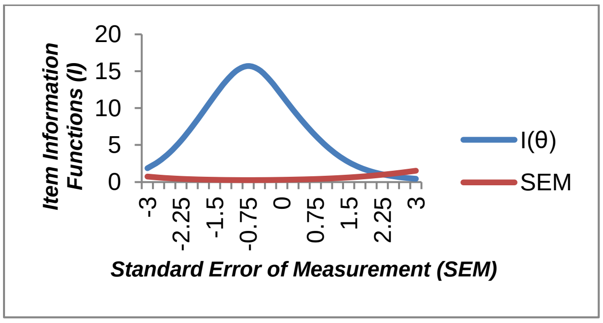

Figure 10. Curve Information Function Test Package B

Standard nine (stanine) is a standard value with a scale of nine where the value ranges from 1 to 9 (there is no value 0 and no value 10). Based on the results of the BILOG-MG output in phase 3, the ability of each test participant was obtained as many as 372 respondents for package A and package B. The abilities of the test participants based on the calculation of a scale of 9 (nine) are presented in Table 2.

Table 2. Calculation of Capability Category Package A

Percentile Rank (θ)	Stanine	Percentage of Participants	Criteria
16.17 to upper	9	3.4	Higher
1.154 — 1.616	8	10.2	Above of average
0.691 — 1.153	7	9.6	Above average
0.228 — 0.690	6	15.3	Average
-0.24 — 0.227	5	15.8	Average
-0.70 — -0.23	4	26.8	Average
-1.16 — -0.69	3	4.8	Lower of average
-1.62 — -1.15	2	10;2	Lower of average
-1.61 to lower	1	3.4	Lower

The results of the calculation showed that 26.8% of the students' ability categories who worked on package A were in the average category. The ability (θ) of the test takers is in the range -0.70 to -0.23. Thus, the distribution of the abilities of students who worked on package A was mostly done by students who were included in the average ability.

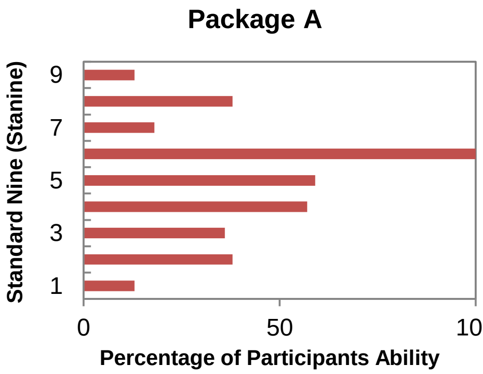

Figure 11. Ability Distribution of Package A Students

Based on the results of calculations in Table 2, the test taker's ability categories can be shown in Figure 11 which shows that the percentage of test takers who worked on package A almost formed a normal curve. The abilities with the highest percentages were ranked 6, 5, and 4 included in the average ability category.

The ability category of students who worked on package B showed that 34.1% of respondents were in the average category. The test taker's ability (θ) is in the range -0.70 to 0.24. Thus, the distribution of the abilities of students who worked on package B was mostly done by students who were included in the average ability.

Table 3. Calculation of Capability Category Package B

Percentile Rank (θ)	Stanine	Percentage of Participants	Criteria
1.63 to upper	9	9.9	Higher
1.163 — 1.629	8	6.1	Above of average
0.697 — 1.162	7	6.1	Above of average
0.229 — 0.696	6	16.3	Average
-0.237 — 0.228	5	12.6	Average
-0.70 — -0.24	4	26.6	Average
-1.171 — -0.69	3	9.6	Lower of average
-1.64 — -1.16	2	7.5	Lower of average
-1.63 to lower	1	4.8	Lower

Based on the results of the calculations in Table 3, the test taker's ability categories can be shown in Figure 12 which shows that the percentage of test takers who worked on Package B almost formed a normal curve. The abilities with the highest percentages were ranked 6, 4, and 5 included in the average category. Thus, overall package A and package B were tested on class X high school students in the city of Yogyakarta who have average abilities.

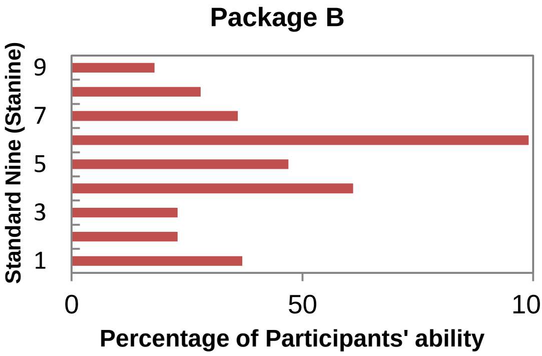

Figure 12. Ability Distribution of Package B Students

The results of this study indicate that a physics learning achievement test has been successfully developed by utilizing the item specifications for SMA in the city of Yogyakarta. The test consisting of optical material, temperature and heat has resulted in a parallel test form that has been tested for quality and equivalence both theoretically and empirically. Thus, the parallel test package generated through the item specifications can be used to measure the same competence for each indicator.

CONCLUSION

Based on the results of the research and data analysis, it shows that the average level of difficulty in the learning achievement test items in high school physics subjects in the city of Yogyakarta between package A and package B shows that the level of difficulty index is almost equal without going through an equivalent process. The distribution of the ability of students in class X SMA in the city of Yogyakarta who took the test was mostly in the average ability category.

Schools are expected to be accustomed to making item specifications so that when making the same indicators they produce relatively the same items. Thus, item similarity is more guaranteed so that if there is no item specification, it is feared that it will result in test items that tend to be free from one item to another even though they come from the same indicator and test the same thing.

REFERENCES

Aiken, L.R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psycological Measurement*, *40*, 955-959.

Aiken, L.R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psycological Measurement, 45*, 131-142.

Amelia, N.R.,& Kriswantoro. (2017). Implementasi Item Response Theory sebagai Basis AnalisisKualitasButirSoal dan Kemampuan Kimia Siswa Kota Yogyakarta. Jurnal Kima dan Pendidikan Kimia. 2(1), 6.

Anderson, D., Irvin, S., Alonzo, J., & Tindal. (2015). *Gauging Item Alignment Through Online Systems While Controlling for Rater Effe*cts. *Journal Educational Measurement: Issues and Practice*, 34(1), 22–33.

Azwar, S. (2014). *Reliabilitas dan validitasedisi 4*. Yogyakarta: Pustaka Pelajar.

DOI: https://doi.org/10.18860/preschool.v1i2.9091.

DOI: https://doi.org/10.21831/pep.v18i1.2120

Elvira, M., & Hadi, S. (2016). Karakteristik Butir Soal Ujian Semester dan Kemampuan Siswa SMA di Kabupaten Muaro Jambi. *JurnalEvaluasi Pendidikan*, 4(1), 62.

Elvira, M., & Sainuddin, S. (2020). Uji Model Instrumen The Mathematical Development Beliefs Survey (MDBS) pada Pendidikan Prasekolah. *Jurnal Perkembangan dan Pendidikan Anak Usia Dini*, 1(2), 98.

Fatkhudin, A, Surarso, B, &Subagio, A. (2014). Item Response Theory Model Empat Parameter Logistik Pada Computerized Adaptive Test. *Jurnal Sistem Informasi Bisnis.* 2, 122.

https://media.neliti.com/media/publications/173335-ID-pengembangan-tes-diagnostik-sebagai-alat.pdf.

Istiyono. E., Mardapi. D., & Suparno. (2014). Pengembangan Tes Kemampuan Berpikir Tingkat Tinggi Fisika (Pysthots) Peserta Didik SMA. *Jurnal Penelitian dan Evaluasi Pendidikan*. 18 (1), 2.

Kriswantoro., Amelia, N.R., & Irwanto. (2016). Peningkatan Kompetensi Calon Pendidik Kimia Melalui Item Response Theory: Strategi Menghadapi Masyarakat Ekonomi ASEAN. *Makalah* Presentasi Seminar Nasional Kimia Dan Pendidikan Kimia VIII, Surakarta.

Mardapi, D. (2012). *Pengukuran penilaian dan evaluasi pendidikan*. Yogyakarta: Nuha Litera.

Megawati, E. (2012). *Pengembangan perangkat tes kimia dalam rangka pembentukan bank soal di kabupaten paser kalimantan timur.* Tesis magister, tidak diterbitkan, Universitas Negeri Yogyakarta, Yogyakarta.

Murdaka, B.E.J. & Kuntoro, T.P. (2013). *Fisika dasar edisi 2*. Yogyakarta: CV Andi offset.

Oriondo, L.L., & Dallo-Antonio, E.M. (1998). *Evaluation educational outcomes*. Manila: Rex Printing Compagny, inc.

Prihatni, Y., Kumaidi., & Mundilarto. (2016). Pengembangan Instrumen Diagnostik Kognitif Pada Mata Pelajaran IPA di SMP. *Jurnal Penelitian dan Evaluasi Pendidikan*, 20(1), 112. DOI: https://doi.org/10.21831/pep.v20i1.7524

Purnama, D.N. (2017). Characteristics and equation of accounting vocational theory trial test items for vocational high schools by subject-matter teachers' forum. *Research and Evaluation in Education* (REID). 3(2), 154.

Purnama, D.N., & Alfarisa, F. (2020). Karakteristik Butir Soal Try Out Teori Kejuruan Akuntansi SMK berdasarkan Teori Tes Klasik dan Teori Respons Butir. *Jurnal Pendidikan Akuntansi Indonesia.* 18(1), 39.

Retnawati, H. (2014). *Teori respon butir dan penerapannya untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Nuha Medika.

Retnawati, H. (2016). *Validitas, reliabilitas & karakteristik butir panduan untuk peneliti, mahasiswa, dan psikometrian*. Yogyakarta: NuhaMedika.

Rukli. (2016, May). *Analisys Item Information Function on the Test of Mathematics*. *Proceeding* of International Conference on Educational and Research Evaluation, Yogyakarta.

Rusilowati, A. (2015). Pengembangan Tes Diagnostik sebagai Alat Evaluasi Kesulitan Belajar Fisika. *Prosiding* Seminar Nasional Fisika dan Pendidikan Fisika. 6(1), 6.

Russell, M.K., & Airasian, P.W (2012). *Classroom assessment concepts and applications (7thed.).* New York: McGraw-Hill Companies, Inc.

Sarwiningsih, R. (2017). Komparasi Ketepatan Estimasi Koefisien Reliabilitas Tes Ujian Nasional Kimia Provinsi Jambi TahunAjaran 2014/2015. *Jurnal Kima dan Pendidikan Kimia*, 2(1), 35.

Schunk, D.H. (2012). *Learning theories an educational perspective (6th ed.).* The University of North Carolina at Greensboro: Pearson Education, Inc.

Sumintono, B. & Widhiarso, W. (2015). *Aplikasi pemodelan RASCH pada asessment pendidikan*. Cimahi: Trim Komunikata.

Umar, J.(1999).Item banking.Dalam Masters, G.N.,& Keeves, J.P (Ed), *Advances in measurement in educational research and* assessment. New York: Pergamon