Received : 16 December 2019
Revised : 18 April 2020
Accepted : 23 April 2020
Online : 6 June 2020
Published: 30 June 2020

DOI: doi.org/10.21009/1.06105

The Development of Horizontal Anchor Items Test Tool by Rasch Model for Physics National Examination using Macromedia Flash

Yetti Supriyati[a)], Raihanati, Wirda Nilawati

Physics Education, Faculty of Mathematics and Natural Science, Universitas Negeri Jakarta
Jl. Rawamangun Muka, Jakarta Timur, Indonesia

✉: [a)]y_supriyati@yahoo.com

Abstract

This study aims to develop a Macromedia Flash-based test device using Horizontal Anchor Items for the National Examination in High School Physics lessons. The research method used in this study is research and development. This research and development uses a qualitative and quantitative approach and uses the Research and Development (R & D) development model of the Dick and Carey model. The results of the validation test of the National Examination test instrument by material experts showed that the National Examination test instrument that had been developed had very good quality (91.25%) when viewed in terms of material. Besides, the results of the validation test of the National Examination test instrument by media experts showed that the National Examination test instruments that had been developed had very good quality (87.5%) when viewed in terms of media. In the first stage of the empirical test, the first reliability test device was quite good, while the second test device showed sufficient reliability. Besides that, the two test kits showed that all items were declared valid. The results of the Person fit for the test respondent of the first test device, and the second test device showed that there were no respondents who were inconsistent, careless, lucky, or cheating. In the second stage of the empirical test, the first test device showed quite good reliability, while the second test device showed sufficient reliability. Besides that, both the test kits show that all items are declared valid. The results of the Person fit for the test respondent of the first test device, and the second test device showed that there were no respondents who were inconsistent, careless, lucky, or cheating. In the third empirical test, the first test device showed good reliability, while the test device showed sufficient reliability. Besides that, the two test kits showed that all items were declared valid. The results of the Person fit for the test respondent of the first test device, and the second test device showed that there were no respondents who were inconsistent, careless, lucky, or cheating so that the two test kits were declared feasible and could be used for the implementation of the Computer-Based National Examination in Physics.

Keywords: horizontal anchor items, test tool, Rasch models, physics, national examination

INTRODUCTION

Assessment is one part of the learning process to determine the success rate of a student. In Indonesia, the assessment is divided into three, namely the assessment by educators, educational units, and the government. The research carried out by the government is known as the National Examination. Exams are one of the crucial instruments used by each teacher to determine the level and extent to the achievement students by the desired teaching goals (Adow et al. 2010). The National Examination is a type of evaluation conducted at school and adjusted to national achievement standards. This is held at

the end of the learning process. This is used to obtain, analyze, and interpret student learning processes and their learning outcomes systematically and sustainably (Jamiludin et al. 2017). The results of the exam in the school will provide information about the level of success students achieve in the learning objectives The results of the national exam are used as one of the considerations for (a) mapping the quality of programs and educational units; (b) determine the graduation of students from educational programs and units; (c) assistance or services to educational units, the focus of research is the results of National Examinations as a basis for mapping the quality of education (Irawan 2015).

Today, the examination process in the education level shows rapid changes. Based on the Government Regulations on Education and Culture Number 23 of 2016, Assessment of Learning Outcomes by the Government is carried out in the form of National Examinations and other forms needed. The National Examination was conducted in order to map the quality of programs and educational units, consideration of selection into the next level of education, and fostering and providing assistance to education units in their efforts to improve the quality of education (Kementerian Pendidikan Nasional 2016). National Examination based on Paper and Pencil Test has made several problems such as test innovations, test making, greater security, standardization, and test efficiency. That is why National Examination based on paper and pencil test become contradictory in Indonesia (because of those disadvantages).

National Examination has developed into the Computer-Based Test (CBT) National Examination. CBT has begun to gain popularity compared to traditional PPT because of the many benefits it can offer (Abubakar & Adebayo 2016; Balogun & Olanrewaju 2016). The use of CBT in measuring the quality of education graduates in Indonesia has shown success (Darmawan & Harahap 2016). In one study, CBT was preferred by test participants compared to CBT (Reid et al. 2016). Computer-based tests (CBT) provide a number of advantages such as more accessible and more precise testing and reporting of tests, test innovations, test making, greater security, standardization, and test efficiency, test books and deletion of answer sheets, more flexible scheduling, reduce measurement errors, etc. (Khoshsima et al. 2017; Olawuyi et al. 2018). Besides, the advantages of other CBTs are lower administrative costs and ease, high accuracy, the suitability of scoring and reporting, and flexible schedules (Hosseini et al. 2014). However, the results of observations show that the implementation of the National Examination is still often experiencing obstacles, including limited computer units in high school. The use of Macromedia flash for National Examination is one solution to this problem. By using the Macromedia flash-based National Examination instrument, students can directly access and carry out the National Examination through a personal computer or smartphone. Besides that, the National Examination instrument based on Macromedia flash will also enable the development of questions that are rich in multimedia, so that the questions will seem monotonous as in a paper and pencil test.

Characteristics of Physics questions must be logical and systematic. The problems raised in the problem originate from a phenomenon, then are identified. In the identification stage, the Physics problems are analyzed to look for any Physics concepts and what Physics problems arise so that solutions can be found. Therefore, a stimulus is needed on every Physics question that can show a Physics phenomenon, then analyze it, until a solution is obtained. By using Computer Based Tests, stimulus problems can be easily resolved.

The implementation of the National Examination in practice is carried out by making parallel test kits. The use of various test instruments often has problems, including tests that one does not necessarily have the same level of difficulty. There is a possibility the first test tends to have easy questions, but the second test has most problems that are difficult (Aşiret & Sünbül 2016). Using scores from several different tests is not directly usable, first requiring equal testing (Gübeş & Kelecioğlu 2016; Nguyen et al. 2018). Test equations are a method used to adjust scores on two or more versions of a test so that scores from different tests can be used on different test instruments (Gonz 2014; Lin et al. 2016).

This pair of parallel test devices have to anchor items with each other, the same items on one test device with the other test devices. Other items in this pair of test instruments are similar non-anchor items through the equivalence of items. Furthermore, the score of the students' work using a pair of parallel test devices which have been equalized is the problem, followed by equalizing the scores

between the pairs of test devices. Anchor items Design is commonly used if a security test problem is one of the crucial considerations and allows us to hold several subtests at the same time.

The use of anchor items on the test device, in general, gives better results to the test takers (Arıkan & Gelbal 2018; Lu et al. 2015). In carrying out equalization tests, at least 20% of the total questions are about anchor items (Uysal & Kilmen 2016). Anchor items are built specifically to represent a total test, where anchor items have a strong correlation with the total test score (Lu & Guo 2018).

The Rasch model can be used to analyze the quality of tests that have been developed. The Rasch model is easy to use and has many interpretations to present and provide data meanings (Mursidi & Soeharto 2017). Unlike the classical test theory, the Rasch Model explains not only the quality of the item questions but also the quality of the standard of the individual (respondent) based on the response given by the respondent (test taker). It provides more information on the quality of the item items, and the respondents (test-takers) are more precise than others, but with smaller standard errors (Hagell 2014). The use of the Rasch model provides good construction and test design (Chan et al. 2014).

Based on the descriptions above, the purpose of this study is to develop a test tool based on Macromedia Flash using Horizontal Anchor Items for the National Examination in High School Physics lessons.

METHOD

The research method used in this research is research and development (research and development) or better known as R&D. This research and development use qualitative and quantitative approaches and uses the Research and Development (R & D) development model of the Dick and Carey model, which can be used to design new products through various stages systematically. Dick and Carey's model is one of the most widely used models of educational research and development. This model has ten steps, they are: 1) assess need to identify the goal, 2) conduct instructional analysis, 3) analyze learners and contexts, 4) write performance objectives, 5) develop assessment instruments, 6) develop instructional strategy, 7) develop and select instructional materials, 8) design and conduct a formative evaluation of instruction; 9) revise instruction; and 10) design and conduct a summative evaluation. (Gall et al. 2003)

Research Introduction

The research introduction conducted the first step until the fourth step of the Dick and Carey Model. The data obtained at this preliminary stage are the results of documentation and literature, the results of observations regarding the condition of the National Examination. Before the R & D process can be applied, it is necessary to describe specifically the educational products that will be developed. This study tries to describe the assessment products that will be developed. After describing the model that will be developed, various kinds of literature studies are used that are useful in developing the product. The study of literature that can be done is to do a literature study in examining various theories and through studies that have been carried out by previous researchers.

Model Development Planning

After conducting research at the preliminary stage and completing the study of literature and related information, the next step is to do the planning. Planning was done so that it can make it easier to make a clear line of steps in implementing development research in the field. The critical elements that are in the planning phase are an estimation of money, human resources, and the time needed to develop the product.

Research Product

This stage was conducted with the fifth step until the eighth step of Dick and Carey Model and also the ninth step of Dick and Carey to make a revision for the product that has been developed. The products developed in this study are based on assessment instruments, instructional strategy,

instructional materials, and formative evaluation of instruction. When the product has been developed, it will take to the ninth step to make make a revision.

The initial product was developed by preparing the material needed to support the model to be developed by gathering various information. Products that have been made are validated by several experts, namely material experts and media experts, to review the initial product and provide input for product improvement. If there is still a deficiency in the product, then the second revision is made, but if it does not revise it, the product can be said to be the final product or final product.

The final product needs to be tested, and various kinds of revisions are repeated so that the final product is ready to be used as a test device for the National Examination. Validation is carried out on the model that will be developed through testing experts related to the product being developed. Suggestions and input from experts and supervisors are used for improvements and revisions in improving the design of the models made. The product results were developed in the form of a test model in the form of a test that will be used for the National Examination so that it will be able to help students in class XII in the future.

The initial field test is a validation test of the assessment model developed, then explains the procedures in developing the model to be developed, and the experts are tasked to observe and provide input on the model that has been produced. Based on input from these experts, the existing model was revised. Assessment is carried out by two experts or experts, each of whom is a physics material expert and media expert.

The researcher conducted a trial of the product developed by gathering respondents as many as ten students of class XII who would take the National Examination Physics lesson. Then the Respondents were asked to try the test device and provide input on the model developed. Based on the results of the trials and these inputs, the model was revised again by the researchers.

The medium group trial is an operational field test involving 40 respondents. Respondents who will be included are based on the number of classes which will take the National Examination Physics lesson. Large group trials will be conducted by outside parties to maintain the objectivity of the conclusions that will be produced. This input from the field trial will be the basis for the improvement and improvement of the final product. After being repaired according to input from the field, the products developed can be considered final and ready to be implemented.

The large group trial is an operational field test involving more respondents, namely 100 respondents. Respondents who will be included are based on the number of classes which will take the National Examination Physics lesson. Large group trials will be conducted by outside parties to maintain the objectivity of the conclusions that will be produced. This input from the field trial will be the basis for the improvement and improvement of the final product. After being repaired according to input from the field, the products developed can be considered final and ready to be implemented.

The result of the validation and field trial will declare the quality of products that have been developed. It was the last step of the Dick and Carey model, summative evaluation.

RESULTS AND DISCUSSION

Expert Validation

The expert review is carried out at the initial stage of validation. Validation by experts was carried out quantitatively by material experts and the media on two sets of tests that had been developed, of which 20% of each test device (8 questions) were anchor items.

The results of the expert validation of the material were carried out to determine the quality of the National Examination test instrument that had been developed in terms of material. The feasibility test of the material expert involved two Physics lecturers at the Jakarta State University who were professionals in their fields. The results of the validation test of the National Examination test instrument by material experts showed an average percentage of 91.25% with very good interpretations. The details of the validation results of the National Examination test instruments by material experts are as follows: (1) The material is 91.68%, while 8.32% is caused by a number of questions in test instruments that have incorrect answers, such as in the question about the half-life of radioactive elements; (2) The construction of the question is 92.5%, while 7.5% is because some questions do not have the same relatively long answer choices; and (3) Language is 87.5%, while 12.5%

is caused by a number of questions using non-simple language. This shows that the National Examination test instrument that has been developed has met the eligibility requirements in terms of material.

The results of the media experts validation were carried out to determine the quality of the National Examination test instruments that had been developed in terms of media. The validation test of media experts involves a media expert lecturer at the Jakarta State University who has been a professional in his field. The results of the validation test of the National Examination test instrument by media experts showed an average percentage of 87.5% with very good interpretations. The details of the validation test results of the National Examination test instruments by media experts are as follows: (1) Initial page of 91.67%, while 8.33% is because the use of color is still monotonous; (2) Instructions for use page of 75%, while 25% is caused because the layout on the usage instructions page is not proportional; (3) Question page of 91.67%, while 8.33% is because the media which used on several questions is not proportional and cannot be observed; and (4) Final page 85%, while 15% is caused by the function of the button on the final page does not work correctly. This shows that the National Examination test instrument that has been developed has met the requirements of feasibility in terms of media.

First Stage of Empirical Test Results

The empirical test was carried out in one of the high schools in Tangerang. The first stage empirical test was carried out with a total of 10 respondents. The data were analyzed with Winsteps software to determine the overall quality of the instrument, the quality of the items, and the ability of students. Through these analyses, it will be known whether the instrument has been compiled accordingly or not. If it is appropriate, then the instrument can be continued with a second stage empirical test, but if it is not appropriate, further analysis needs to be done. The first analysis was conducted on 72 items (8 questions about anchor items and 64 questions for non-anchor items) and ten respondents. The analysis show that both the item and the respondent are good and are fitted with the model. Besides, based on the unidimensional value obtained from Win-steps software analysis, the quality of the instrument is good. Then the results of the summary statistical analysis also show the quality of the instrument, which has a good consistency.

The statistical summary shows overall info about the quality of the overall response pattern, the overall quality of the instrument, and the interaction between Person and item. The summary statistics for the first test instrument show that the value of person reliability is 0.39, while the value of item reliability is 0.71.

Based on the data, it appears that the average Person measure shows a value of -0.51 logit, which shows the value of all respondents in working on the items given. The average value that is smaller than the logic value of 0.0 indicates the tendency of respondents to be smaller than the difficulty level of the question. Besides, person reliability and item reliability in the table is used to determine the level of reliability of the Person and items. If less 0.67 means having weak reliability, 0.67 to 0.80 is enough, 0.80 to 0.90 is good; 0.91 to 0.94 is very good, and more than 0.94 is said to have special reliability. Based on the data, person reliability shows a value of 0.39, where good consistency of answers from respondents has weak reliability. In contrast, the reliability item shows a value of 0.71, where the quality of the items in the first test device shows pretty good reliability. The summary statistics for the second test instrument show that the value of person reliability is 0.44, while the value of item reliability is 0.70.

Based on the data, it appears that the average Person measure shows a value of -0.67 logit, which shows the value of all respondents in working on the items given. The average value that is smaller than the logic value of 0.0 indicates the tendency of respondents to be smaller than the difficulty level of the question. Besides, person reliability and item reliability in the table is used to determine the level of reliability of the Person and items. If less 0.67 means having weak reliability, 0.67 to 0.80 is enough, 0.80 to 0.90 is good; 0.91 to 0.94 is very good, and more than 0.94 is said to have special reliability. Based on the data, person reliability shows a value of 0.44, where good consistency of answers from respondents has weak reliability. In contrast, the reliability item shows a value of 0.70, where the quality of the items in the second test device shows pretty good reliability.

Item fit explains whether the item functions normally in making measurements or not. This information will be used as a reference by researchers to improve the quality of questions if items that

are not fit are obtained. Both the first test device and the second test device show that all items are declared valid.

The approach to using the Rasch Model also provides an interpretation of the difficulty level of an item. Through the Winsteps application, the difficulty level of the problem can be seen from the logic value. A high logic value indicates a high level of difficulty and vice versa. Positive logit values show difficult and negative items for easy items. While the standard deviation (SD) value can also be combined to classify the difficulty level. 0.00 logit + 1SD is a group of difficult questions, more than + 1SD is a very difficult group of questions; 0.00 logit - 1SD is a group of easy questions, and smaller than 1SD is a group of questions very easy. On the first test, questions number 36, 25, 2, and 20 show very difficult questions, questions number 32, 21, 19, 11, 7, 5, 1, 40, 39, 38, 35, 34, 31, 13 , 12, 9, and 8 including difficult questions, questions number 3, 4, 14, 18, 30, 10, 22, 24, 29, 37, 16, 23, 17, 26, 33, 6, 15, and 27 including easy questions, and number 28 questions including very easy questions. The difficulty level test results for the first test device show several items that are categorized as very easy, easy, difficult, and very difficult. The problem that is categorized very easily is number 28 about Simple Harmonic Motion, one of which is categorized as easy one of them is number 3 concerning One-Dimensional Motion, one of which is categorized as difficult one of them is number 32 about Lorentz Force. One that is categorized as very difficult is number 20 about Temperature Scale Conversion. While for the second test device, questions number 12, 31, 34, 35, 1, 3, 9, 13, 18, 36, 38, and 40 show very difficult questions, questions number 8, 21, 2, 10, 11, 14, 20, 25, 30, and 32 including difficult questions, questions number 4, 23, 26, 27, 29, 37, 5, 19, 22, 33, 7, and 39 including easy questions, and questions number 6 , 16, 17, 24, 28, and 15 including very easy questions. The difficulty level test results for the second test device show several items that are categorized as very easy, easy, difficult, and very difficult. The problem that is categorized as very easy one of them is number 16 about Mechanical Energy, one of which is categorized as easy one of them is number 5 concerning Vertical Upward Motion, one of which is categorized as difficult one of them is number 30 about Optics, and one that is categorized as very difficult is number 34 about Magnetic Induction.

The average logit value is used as the identifier of the respondent group. The standard deviation (SD) value is combined to classify the respondent's ability level. In the range of "average logit - 1SD" up to "average logit + 1SD" is a group of respondents with moderate ability, more than "average logit + 1SD" is a group of highly capable respondents, and smaller than "average logit - 1SD" is a group of low-ability respondents. On the first test device, the group of highly capable respondents must have a logit value greater than -0.01, where respondents who have high abilities are respondents numbers 2 and 10. Groups of respondents who are capable have logit values ranging from -1.01 to -0, 01, where respondents who have the moderate ability are respondents number 3, 9, 1, 4, and 5 while groups of respondents with low ability have a logit value of less than -1.01, where respondents who have the low ability are respondents number 6, 7, and 8. Whereas in the second test device, the group of highly capable respondents must have a logit value greater than -0.12, where respondents who have the high ability are only respondent number 3. Groups of capable respondents have logit values ranging from -1, 22 to -12.12, where respondents who have moderate abilities are respondents number 1, 9, 4, 2, 5, 6, and 8. While the group of respondents with ability low numbers have a logit value of less than -1.22, where respondents who have the low ability are respondents numbers 7 and 10.

Person Fit is used to detect the presence of individuals who have inappropriate response patterns. The different response pattern is the pattern of the incompatibility of answers given based on their ability. This is used to indicate the consistency of the respondent's thinking as well as to find out the respondents who are careless or lucky and respondents who cheat / cheating. The results of the Person fit for the test respondent in the first test showed that there were no respondents who were inconsistent, careless, lucky, or cheating. However, when viewed on a scalogram, it appears that respondents 1 included respondents who were careless because the second easiest item (no. 6) could not answer it, but the second hardest item (no. 36) could be answered, where no other respondent could answer question number 36. The results of the Person fit for the second trial test respondent showed that there were no respondents who were inconsistent, careless, lucky, or cheating. However, when viewed on a scalogram, it appears that respondents five including respondents were lucky because item number 9 could be answered, even though the response could not answer easy questions, and item 9 only answered successfully.

Second Stage of Empirical Test Results

The second stage of the empirical test was carried out in one of the high schools in Jakarta. The second stage empirical test was carried out with a total of 40 respondents. The data were analyzed with Winsteps software to determine the overall quality of the instrument, the quality of the items, and the ability of students. Through these analyzes, it will be known whether the instrument has been compiled accordingly or not. If it is appropriate, the instrument can be continued with the third stage of empirical testing, but if it is not appropriate, further analysis is needed. The analysis was carried out on 72 items (8 question anchor items and 64 non-anchor items) and ten respondents. The analysis show that both the item and the respondent are good and are fitted with the model.

The statistical summary shows overall info about the quality of the overall response pattern, the overall quality of the instrument, and the interaction between Person and item. The summary statistics for the first test instrument show the value of person reliability is 0.48, while the value of item reliability is 0.79.

Based on the data, it appears that the Person measure shows an average value of -0.89 logit, which shows the value of all respo7ndents in working on the items given. The average value that is smaller than the logic value of 0.0 indicates the tendency of respondents to be smaller than the difficulty level of the question. Besides, person reliability and item reliability in the table is used to determine the level of reliability of the Person and items. If less 0.67 means having weak reliability, 0.67 to 0.80 is enough, 0.80 to 0.90 is good; 0.91 to 0.94 is very good, and more than 0.94 is said to have special reliability. Based on the data, person reliability shows a value of 0.48 where good consistency of answers from respondents has weak reliability, while the reliability item shows a value of 0.79, where the quality of the items in the first test device shows pretty good reliability. The summary statistics for the second test instrument show that the value of person reliability is 0.39, while the value of item reliability is 0.80.

Based on the data, it appears that the average Person measure shows a value of -0.87 logit, which shows the value of all respondents in working on the items given. The average value that is smaller than the logic value of 0.0 indicates the tendency of respondents to be smaller than the difficulty level of the question. Besides, person reliability and item reliability in the table is used to determine the level of reliability of the Person and items. If less 0.67 means having weak reliability, 0.67 to 0.80 is enough, 0.80 to 0.90 is good; 0.91 to 0.94 is very good, and more than 0.94 is said to have special reliability. Based on the data, person reliability showed a value of 0.39, where good consistency of answers from respondents had weak reliability. In contrast, the reliability items showed a value of 0.80, where the quality of the items in the second test device showed good reliability.

Item fit explains whether the item functions normally in making measurements or not. This information will be used as a reference by researchers to improve the quality of questions if items that are not fit are obtained. Both the first test device and the second test device show that all items are declared valid.

The approach to using the Rasch Model also provides an interpretation of the difficulty level of an item. Through the Win-steps application, the difficulty level of the problem can be seen from the logic value. A high logic value indicates a high level of difficulty and vice versa. Positive logit values show difficult and negative items for easy items. While the standard deviation (SD) value can also be combined to classify the difficulty level. 0.00 logit + 1SD is a group of difficult questions, more than + 1SD is a very difficult group of questions; 0.00 logit - 1SD is a group of easy questions, and smaller than 1SD is a group of questions very easy. On the first test device, questions number 32, 38, 20, 10, 25, and 40 show very difficult questions; questions number 1, 11, 34, 3, 19, 13, 21, 31, 35, 39, 18, 2, 7, and 9 included difficult questions; Question number 17, 24, 36, 37, 30, 8, 12, 14, 22, 23, 33, 5, 27, and 29 including easy questions, as well as questions number 4, 16, 15, 6, 28 and 26 including questions which is very easy. The difficulty level test results for the first test device show several items that are categorized as very easy, easy, difficult, and very difficult. The problem that is categorized as very easy one of them is number 16 about Mechanical Energy, one of which is categorized as easy one of them is number 17 about Impulse, one of which is categorized as difficult one of them is number 11 about Rigid Equilibrium, and one that is categorized as very difficult is number 32 about Lorentz Force. While for the second test device, questions number 38, 34, 12, 3, 40, and 19 show very difficult questions; questions number 9, 18, 32, 35, 11, 25, 31, 20, 36, 39, 21, 30, 10, and 13 included difficult

questions; questions number 1, 2, 7, 8, 14, 17, 37, 24, 28, 5, 23, 27, and 29 including easy questions, as well as questions number 22, 33, 26, 4, 15, 16, and 6 including very easy question. The difficulty level test results for the second test device show several items that are categorized as very easy, easy, difficult, and very difficult. The problem that is categorized as very easy one of them is number 26 about Stationary Wave, one of which is categorized as easy one of them is number 1 concerning Vernier Calliper, one of which is categorized as difficult one of them is number 10 about One-Dimensional Motion Dynamics (Newton's Law), and one that is categorized as very difficult is number 40 about Half Time.

The average logit value is used as the identifier of the respondent group. The standard deviation (SD) value is combined to classify the respondent's ability level. In the range of "average logit - 1SD" up to "average logit + 1SD" is a group of respondents with moderate ability, more than "average logit + 1SD" is a group of highly capable respondents, and smaller than "average logit - 1SD" is a group of low-ability respondents. In the first test device, the group of highly capable respondents must have a logit value greater than -0.36, where respondents who have the high ability are respondents number 11, 4, 17, 5, 1, 10, 33, and 40. The group of respondents is capable while having logit values ranging from -1.42 to -0.36, where respondents who have moderate ability are respondents number 7, 37, 9, 30, 34, 39, 14, 27, 20, 21, 25, 28, 31, 24, 3, 8, 12, 23, 29, 35, 38, 13, 22, 32, 15, 19, 26. While groups of respondents with low ability have a logit value of less than -1.42, where respondents who have the low ability are respondents number 18, 2, 6, 16, and 36. Whereas in the second test device, the group of highly capable respondents must have a logit value greater than 0.91, where respondents who have the high ability are respondents number 38, 34, 12, 3, 40, and 19. Groups of respondents with moderate ability have logit values ranging from -0.91 to 0.91, where respondents who has moderate ability is respondent number 9, 18, 32, 35, 11, 25, 31, 20, 36, 39, 21, 30, 10, 13, 1, 2, 7, 8, 14, 17, 37, 24, 28, 5, 23, 27, and 29. While the group of respondents with low ability has a logit value of less than -0.91, where respondents who have the low ability are respondents number 22, 33, 26, 4, 15, 16, and 6.

Person Fit is used to detect the presence of individuals who have inappropriate response patterns. The different response pattern is the pattern of the incompatibility of answers given based on their ability. This is used to indicate the consistency of the respondent's thinking as well as to find out the respondents who are careless or lucky and respondents who cheat / cheating. The results of the Person fit for the test respondent in the first test showed that there were no respondents who were inconsistent, careless, lucky, or cheating. However, when viewed on a scalogram, respondents 5 included respondents who were careless because the second easiest item (no. 28) could not answer it, but the second hardest item (no. 32) could be answered. The results of the Person fit for the second trial test respondent showed that there were no respondents who were inconsistent, careless, lucky, or cheating. But when viewed on a scalogram, it shows that respondents 13 was careless because the easiest item (no. 6) could not be answered, while the hardest item (no. 38) could be answered.

Third Stage of Empirical Test Results

The third stage of the empirical test was carried out in one of the high schools in Bekasi. The third stage of the empirical test was carried out with a total of 100 respondents. The data were analyzed with Win-steps software to determine the overall quality of the instrument, the quality of the items, and the ability of students. Through these analyzes, it will be known whether the instrument has been compiled accordingly or not. If it is appropriate, the instrument can be continued with a conclusion, but if it is not appropriate, further analysis needs to be done. The analysis was carried out on 72 items (8 question anchor items and 64 non-anchor items) and ten respondents. The analysis show that both the item and the respondent are good and are fit with the model. Besides, based on the unidimensional value obtained from Winsteps software analysis, the quality of the instrument is good. Then the results of the summary statistical analysis also show the quality of the instrument with good consistency.

The statistical summary shows overall info about the quality of the overall response pattern, the overall quality of the instrument, and the interaction between Person and item. The summary statistics for the first test instrument show the value of person reliability is 0.56, while the value of items reliability is 0.89.

Based on the data, it appears that the Person measure shows an average value of -0.88 logit, which shows the value of all respondents in working on the items given. The average value that is smaller

than the logic value of 0.0 indicates the tendency of respondents to be smaller than the difficulty level of the question. Besides, person reliability and item reliability in the table is used to determine the level of reliability of the Person and items. If less 0.67 means having weak reliability, 0.67 to 0.80 is enough, 0.80 to 0.90 is good; 0.91 to 0.94 is very good, and more than 0.94 is said to have special reliability. Based on the data, person reliability shows a value of 0.56, where good consistency of answers from respondents has weak reliability. In contrast, the reliability item shows a value of 0.89, where the quality of the items in the first test device shows good reliability. The summary statistics for the second test instrument show that the value of person reliability is 0.33, while the value of item reliability is 0.92.

Based on the data, it appears that the Person measure shows an average value of -0.89 logit, which shows the value of all respondents in working on the items given. The average value that is smaller than the logic value of 0.0 indicates the tendency of respondents to be smaller than the difficulty level of the question. Besides, person reliability and item reliability in the table is used to determine the level of reliability of the Person and items. If less 0.67 means having weak reliability, 0.67 to 0.80 is enough, 0.80 to 0.90 is good; 0.91 to 0.94 is very good, and more than 0.94 is said to have special reliability. Based on the data, person reliability showed a value of 0.33, where good consistency of answers from respondents had weak reliability. Then, the reliability items showed a value of 0.92, where the quality of the items in the second test device showed good reliability.

Item fit explains whether the item functions normally in making measurements or not. This information will be used as a reference by researchers to improve the quality of questions if items that are not fit are obtained. Both the first test device and the second test device show that all items are declared valid.

The approach to using the Rasch Model also provides an interpretation of the difficulty level of an item. Through the Win-steps application, the difficulty level of the problem can be seen from the logic value. A high logic value indicates a high level of difficulty and vice versa. Positive logit values show difficult and negative items for easy items. While the standard deviation (SD) value can also be combined to classify the difficulty level. 0.00 logit + 1SD is a group of difficult questions, more than + 1SD is a very difficult group of questions; 0.00 logit - 1SD is a group of easy questions, and smaller than 1SD is a group of questions very easy. On the first test device, questions number 38, 11, 34, 40, 1, 2, and 20 show very difficult questions; questions number 39, 3, 25, 31, 12, 21, 32, 35, 10, 13, 36, 8, 19, and 14 including difficult questions; questions number 9, 24, 7, 18, 30, 37, 12, 22, 28, 5, and 33 including easy questions, as well as questions number 23, 27, 29, 6, 4, 16, 26, and 25 including the questions that very easy. The difficulty level test results for the first test device show several items that are categorized as very easy, easy, difficult, and very difficult. The problem that is categorized as very easy one of them is number 4 about Circular Motion, one of which is categorized as easy one of them is number 9 about Rotational Kinetic Energy, one of which is categorized as difficult one of them is number 39 about Mass Defect, and one that is categorized as very difficult is number 38 about RLC Series. As for the second test device, questions number 38, 34, 35, 40, 3, 12, 31, 32, and 36 show very difficult questions; questions number 11, 9, 25, 19, 21, 39, 18, 10, 20, 30, 1, and 13 include difficult questions; questions number 2, 7, 24, 14, 8, 27, 5, 17, 29, 37, 22, 28, and 23 include easy questions, as well as questions number 33, 4, 26, 16, 15, and 6 including questions that very easy The difficulty level test results for the second test device show several items which are categorized as very easy, easy, difficult and very difficult. The problem that is categorized as very easy one of them is number 16 about Mechanical Energy, one of which is categorized as easy one of them is number 5 concerning Vertical Upward Motion, one of which is categorized as difficult one of them is number 30 about Optics, and one that is categorized as very difficult is number 34 about Magnetic Induction.

The average logit value is used as the identifier of the respondent group. While the standard deviation (SD) value is combined to classify the respondent's ability level. In the range of “average logit - 1SD” up to “average logit + 1SD” is a group of respondents with moderate ability, more than “average logit + 1SD” is a group of highly capable respondents, and smaller than “average logit - 1SD” is a group of low-ability respondents. In the first set of tests, the group of highly capable respondents must have a logit value greater than -0.30, they are respondents number 41, 40, 92, 17, 97, 32, 55, 67, 81, 28, 33, 38, 89, 90, 23, 37; the group of respondents with moderate ability have logit values ranging from -1.46 to -0.30, they are respondents number 22, 34, 44, 64, 78, 61, 69, 70, 72, 93, 7, 25, 66, 94,

3, 6, 24, 36, 43, 51, 83, 99, 10, 26, 31, 42, 50, 52, 53, 54, 62, 76, 88, 5, 15, 39, 48, 63, 73, 85, 86, 2, 14, 16, 21, 56, 57, 60, 91, 95, 12, 20, 30, 46, 65, 71, 79, 96, 18, 29, 45, 58, 74, 75, 98, 1; and groups of respondents with low abilities have a logit value of less than -1.46, they are respondents number 4, 11, 27, 35, 47, 49, 68, 77, 80, 84, 100, 8, 9, 13, 19, 82, 59, 87.

Meanwhile, based on the data for the second test instrument, the group of highly capable respondents must have a logit value greater than -0.42, they are respondents number 47, 39, 80, 74, 56, 72, 1, 28, 31, 37, 54, 71, 25, 36, 53, 76, 95; the group of respondents with moderate ability have logit values ranging from -1.36 to -0.42, they are respondents number 2, 16, 78, 94, 10, 44, 83, 90, 97, 14, 18, 22, 24, 33, 40, 50, 57, 61, 63, 75, 13, 26, 46, 64, 68, 73, 77, 88, 92, 93, 11, 12, 48, 52, 62, 66, 70, 82, 84, 99, 6, 8, 17, 20, 21, 32, 35, 49, 65, 81, 85, 96, 41, 55, 58, 59, 67, 98, 3, 4, 5, 9, 15, 19, 23, 27, 38, 42, 43, 45; and groups of respondents with low abilities have a logit value of less than -1.36, they are respondents number 29, 30, 51, 60, 69, 91, 100, 7, 34, 86, 89, 79, 87.

Person Fit is used to detect the presence of individuals who have inappropriate response patterns. The different response pattern is the pattern of the incompatibility of answers given based on their ability. This is used to indicate the consistency of the respondent's thinking as well as to find out the respondents who are careless or lucky and respondents who cheat / cheating. The results of the Person fit for the test respondent in the first test showed that there were no respondents who were inconsistent, careless, lucky, or cheating. However, when viewed on a scalogram, it appears that 97 respondents, including respondents, were careless because the third easiest item (no. 16) could not answer it, but the hardest item (no. 38) could be answered. In addition, respondents 11 included respondents who were lucky because they were able to answer the most difficult items (no. 38). The results of the Person fit for the second trial test respondent showed that there were no respondents who were inconsistent, careless, lucky, or cheating. However, when viewed on a scalogram, it appears that 74 respondents included respondents who cared because the easiest item (no. 6) could not be answered, while the hardest item (no. 38) could be answered. Besides, respondents 89 included respondents who were lucky because they were able to answer the most difficult items (no. 38).

Discussion

The test kit developed consisted of two packages, with each package having several items anchor horizontal with other types of test devices. The use of horizontal anchor items in the development of these test devices aims to generalize the quality measured between one package and another. Each question package has eight items of anchor items and 32 items of non-anchor items.

After the test device was developed, then it was examined by several experts who were experts in their fields. This expert study aims to find out the quality of instruments that have been developed in terms of material and media. In terms of material, two panelists have provided suggestions and input on the test kits that have been developed. While in terms of material, a panelist has also provided advice and input on the test kits that have been developed.

The results of the validation test of the National Examination test instrument by material experts showed an average percentage of 91.25% with very good interpretations. The details of the validation results of the National Examination test instruments by material experts are as follows: (1) The material is 91.68%, while 8.32% is caused by a number of questions in test instruments that have incorrect answers, such as in the question about the half-life of radioactive elements; (2) The construction of the question is 92.5%, while 7.5% is because some questions do not have the same relatively long answer choices; and (3) Language is 87.5%, while 12.5% is caused by a number of questions using non-simple language. This shows that the National Examination test instrument that has been developed has met the eligibility requirements in terms of material.

The results of the validation test of the National Examination test instrument by media experts showed an average percentage of 87.5% with very good interpretations. The details of the validation test results of the National Examination test instruments by media experts are as follows: (1) Initial page of 91.67%, while 8.33% is because the use of color is still monotonous; (2) Instructions for use page of 75%, while 25% is caused because the layout on the usage instructions page is not proportional; (3) Question page of 91.67%, while 8.33% is because the media which used on several questions is not proportional and cannot be observed; and (4) Final page 85%, while 15% is caused by the function of

the button on the final page does not work properly. This shows that the National Examination test instrument that has been developed has met the requirements of feasibility in terms of media.

The results of the expert review then become an input to be improved to become better. Then, devices that have been validated by experts are then tested in three types of groups, namely small, medium, and large groups.

The first stage of the empirical test was carried out in one of the high schools in Tangerang. The first stage empirical test is carried out with a total of 10 respondents. The results of the summary statistical analysis also show the quality of the instrument, which has a quite good consistency. In the first stage of the empirical test, based on the data, it appears that the average Person measure shows a value of -0.51 logit, which shows the value of all respondents in working on the items given. The average value that is smaller than the logic value of 0.0 indicates the tendency of respondents to be smaller than the difficulty level of the question. Based on the data of the first test device, person reliability showed a value of 0.39 where both the consistency of the answers of the respondents had weak reliability, while the reliability items showed a value of 0.71, where the quality of the items in the first test device showed good reliability. Based on the data, it appears that the average Person measure shows a value of -0.67 logit, which shows the value of all respondents in working on the items given. The average value that is smaller than the logic value of 0.0 indicates the tendency of respondents to be smaller than the difficulty level of the question. Based on the data of the second test device, person reliability shows a value of 0.44 where good consistency of answers from respondents has weak reliability, while the reliability item shows a value of 0.70, where the quality of the items in the second test device shows enough reliability. Both the first test device and the second test device show that all items are declared valid because at least they fulfill one of the three criteria above. The results of item fit for the first and second test devices can be seen in the attachment. The results of the Person fit for the test respondent of the first test device, and the second test device showed that there were no respondents who were inconsistent, careless, lucky, or cheating. Nevertheless, when viewed on a scalogram, it appears that several respondents are careless and lucky in answering questions.

The second stage of the empirical test was carried out in one of the high schools in Jakarta. The second stage of the empirical test was carried out with a total of 40 respondents. The results of the summary statistical analysis also showed the quality of the instrument, which had a quite good consistency. The second stage of the empirical test showed that the average Person measure of the first test device showed a value of -0.89 logit, which shows the value of all respondents in working on the items given. The average value that is smaller than the logic value of 0.0 indicates the tendency of respondents to be smaller than the difficulty level of the question. Based on the data, person reliability shows a value of 0.48 where good consistency of answers from respondents has weak reliability, while the reliability item shows a value of 0.79, where the quality of the items in the first test device shows pretty good reliability. Whereas for the second test device, the Person measure value obtained shows an average value of -0.87 logit, which shows the value of all respondents in working on the items given. The average value that is smaller than the logic value of 0.0 indicates the tendency of respondents to be smaller than the difficulty level of the question. Based on the data, person reliability showed a value of 0.39, where good consistency of answers from respondents had weak reliability. In contrast, the reliability items showed a value of 0.80, where the quality of the items in the second test device showed good reliability. Both the first test device and the second test device show that all items are declared valid because they meet at least one of the three criteria above. The results of item fit for the first and second test devices can be seen in the attachment. The results of the Person fit for the test respondent of the first test device, and the second test device showed that there were no respondents who were inconsistent, careless, lucky, or cheating. But when viewed on a scalogram, it appears several respondents are careless and lucky in answering questions.

The third stage of the empirical test was carried out in one of the high schools in Bekasi. The third stage empirical test was carried out with a total of 100 respondents. The results of the summary statistical analysis also showed the quality of the instruments with good consistency. In the third stage of the empirical test, the average Person measure for the first test device shows a value of -0.88 logit, which shows the value of all respondents in working on the items given. The average value that is smaller than the logic value of 0.0 indicates the tendency of respondents to be smaller than the difficulty level of the question. Based on the data, person reliability shows a value of 0.56, where good

consistency of answers from respondents has weak reliability. Then, the reliability item shows a value of 0.89, where the quality of the items in the first test device shows good reliability. While on the second test device, it appears that the Person measure shows an average value of -0.89 logit, which shows the value of all respondents in working on the items given. The average value that is smaller than the logic value of 0.0 indicates the tendency of respondents to be smaller than the difficulty level of the question. Based on the data, person reliability showed a value of 0.33, where good consistency of answers from respondents had weak reliability. Then, the reliability items showed a value of 0.92, where the quality of the items in the second test device showed good reliability. Both the first test device and the second test device show that all items are declared valid because they meet at least one of the three criteria above. The results of item fit for the first and second test devices can be seen in the attachment. The results of the Person fit for the test respondent of the first test device, and the second test device showed that there were no respondents who were inconsistent, careless, lucky, or cheating. But when viewed on a scalogram, it appears several respondents are careless and lucky in answering questions.

Based on the results of a review of several experts in terms of material and media, as well as the results of empirical trials carried out in three stages, the anchor items-based test kits can be declared feasible and capable of being used properly for the public interest, especially in the interests of education. The expert review is carried out at the initial stage of validation. Validation by experts was carried out quantitatively by material experts and the media on two sets of tests that had been developed, of which 20% of each test device (8 questions) were anchor items. It is in line with some of the results of previous studies which explained that Rasch Model provides more in-depth information about the quality of students (Zamri & Nordin 2015), thus helping educators in evaluating the learning process in the classroom (Suranata et al. 2018; Rahmani 2018; Mursidi & Soeharto 2017). Besides, the quality of the items (problems) that are informed can be more accountable because the quality of the questions is analyzed using the difficulty level.

CONCLUSION

In general, this study aims to develop a test device for horizontal anchor items based on Macromedia flash that is feasible to use for National Examination in high school physics lessons and able to be used appropriately for the public interest, especially in the interests of education. The quality of test instruments that have been developed can be declared feasible to be used for the implementation of the National Examination. In addition to providing an analysis of the quality of the problems that have been developed, the implementation of the Rasch Model also provides information on the quality of students who carry out the test. This informs educators in improving and improving the quality of their students.

REFERENCES

Abubakar, AS & Adebayo, FO 2014, 'Using computer-based test method for the conduct of examination in Nigeria: Prospects, challenges and strategies', *Mediterranean Journal of Social Sciences*, vol. 5, no. 2, pp. 47–56.

Adow, IM, Alio, AA, & Thinguri, R 2015, 'An assessment of the management of KCSE examination and its influence on irregularities among students: A case of secondary schools in Mandera County, Kenya', *Journal of Education and Practice*, vol. 6, no. 28, pp.15-22.

Arıkan, ÇA & Gelbal, S 2018, 'The effect of mini and midi anchor tests on test equating', *International Journal of Progressive Education*, vol. 14, no. 2, pp.148-60.

Aşiret, S & Sünbül, SÖ 2016, 'investigating test equating methods in small samples through various factors', *Educational Sciences: Theory & Practice*, vol. 16, no. 2, pp. 647-68.

Balogun, AG & Olanrewaju, AS 2016, 'Role of computer self-efficacy and gender in computer-based test anxiety among undergraduates in Nigeria', *Psychological Thought*, vol. 9, no. 1, pp. 58-66.

Chan, SW, Ismail, Z, & Sumintono, B 2014, 'A Rasch model analysis on secondary students' statistical reasoning ability in descriptive statistics', *Procedia-Social and Behavioral Sciences*, vol. 129, pp. 133-139.

Darmawan, D & Harahap, E 2016, 'Communication strategy for enhancing quality of graduates nonformal education through computer based test (CBT) in West Java Indonesia', *International Journal of Applied Engineering Research (IJAER)*, vol. 11, no. 15, pp. 8641-5.

Gall, MD, Gall, JP, & Borg, WR 2003, *Educational research: an introduction*, Pearson Education Inc, Boston.

Gonz, J 2014, 'SNSequate : Standard and nonstandard statistical models and methods for test equating', *Journal of Statistical Software*, vol. 59, no. 7, pp. 1-30.

Gübeş, NÖ & Kelecioğlu, H 2016, 'The impact of test dimensionality, common-item set format, and scale linking methods on mixed-format test equating', *Educational Sciences: Theory & Practice*, vol. 16, no. 3, pp. 715-734.

Hagell, P 2014, 'Testing rating scale unidimensionality using the principal component analysis (PCA) / t-test protocol with the Rasch model: The primacy of theory over statistics', *Open Journal of Statistics*, vol. 4, no. 6, pp. 456-465.

Hosseini, M, Abidin, MJZ, & Baghdarnia, M 2014, 'Comparability of test results of computer based tests (CBT) and paper and pencil tests (PPT) among English language learners in Iran', *Procedia-Social and Behavioral Sciences*, vol. 98, no. 6, pp 659-667.

Irawan, C 2015, 'The national examination and the quality of education mapping', *Indonesian Journal of Educational Review*, vol. 2, no. 1, pp 97-105.

Jamiludin, D & Uke, WAS 2017, 'Students' perception towards national examination 2017: computer-based test or paper-based test', *Mediterranean Journal of Social Sciences*, vol. 8, no. 4, pp. 139-144.

Kementerian Pendidikan Nasional 2016, *Government regulations on education and culture number 23 of 2016*, Kemdikbud, Jakarta.

Khoshsima, H, Hosseini, M, & Toroujeni, SMH 2017, 'Cross-mode comparability of computer-based testing (CBT) versus paper-pencil based testing (PPT): An investigation of testing administration mode among iranian intermediate EFL learners', *English Language Teaching*, vol. 10, no. 2, pp 23-32.

Lin, P, Dorans, N, & Weeks, J 2016, 'Linking composite scores: effects of anchor test length and content representativeness', *ETS Research Report Series*, vol. 2016, no. pp. 1–21.

Lu, R & Guo, H 2018, 'A simulation study to compare nonequivalent groups with anchor test equating and pseudo-equivalent group linking', *ETS Research Report Series*, vol. 2018, no. 1, pp. 1-16.

Lu, R, Haberman, S, Guo, H, & Liu, J 2015, 'Use of jackknifing to evaluate effects of anchor item selection on equating with the nonequivalent groups with anchor test (NEAT) design', *ETS Research Report Series*, vol. 2015, no.1, pp. 1-12.

Mursidi, A & Soeharto, S 2017, 'An introduction: Evaluation of quality assurance for higher educational institutions using rasch model', *Journal of Education, Teaching and Learning*, vol. 1, no. 1, pp. 1-6.

Nguyen, C, Griffin, P, & Wu, M 2018, 'Test equating for measuring system progress in longitudinal surveys of student academic achievement', *Journal of Physics: Conference Series*, vol. 1044, no. 1, p. 012064.

Olawuyi, OF, Tomori, RA, & Bamigboye, OO 2018, 'Students' suitability of computer based test (CBT) mode for undergraduate courses in nigerian universities : a case study of university of ilorin', *International Journal of Educational Sciences*, vol. 20, no. (1-3), pp.18–24.

Rahmani, BD 2018, 'Differential item functional analysis on pedagogic and content knowledge (PCK) questionnaire for Indonesian teachers using RASCH model', *Journal of Physics: Conference Series*, vol. 948, no. 1, p. 012061.

Reid, J, Robinson, D, & Lewis, C 2016, 'Assessing the evidence: Student response system versus computer based testing for undertaking multiple choice question assessment in undergraduate nursing education', *Pediatrics and Neonatal Nursing-Openventio Publishers,* vol. 3, no. 1, pp. 10-14.

Suranata, K et al. 2018, 'Diagnosis of students zone proximal development on math design instruction : A Rasch analysis', *Journal of Physics: Conference Series*, vol. 1114, no. 1, p. 012034.

Uysal ,İ & Kilmen, S 2016, 'Comparison of item response theory test equating methods for mixed format tests', *International Online Journal of Educational Sciences*, vol. 8, no. 2, pp. 1-11.

Zamri, A & Nordin, 2015, 'Modeling a multiple choice mathematics test with the rasch model', *Indian Journal of Science and Technology*, vol. 8, no. 12, p. 70650.