Universal Journal of Educational Research 7(12): 2857-2866, 2019 http://www.hrpub.org
DOI: 10.13189/ujer.2019.071237

Final Examination Test Instruments for History Subject in Yogyakarta, Indonesia: A Quality Analysis

Aman

Faculty of Social Sciences, Universitas Negeri Yogyakarta, Indonesia

Received October 6, 2019; Revised November 19, 2019; Accepted November 25, 2019

Abstract This study aims to: 1) analyze the quality of the final examination test instrument for the Grade 10 History subject in the 2013 curriculum used in Yogyakarta; and 2) determine the number of items in the final examination for the Grade 10 History subject in the high school level that use the 2013 curriculum in Yogyakarta that can be recommended for the procurement of test item bank. This research is an explorative descriptive study that uses a quantitative descriptive approach to determine the quality of the test instrument of the History subject final examination based on the 2013 curriculum. Analysis of the data in this study was carried out through two theories with three stages, namely theoretical analysis, which is item analysis by a team of experts (expert judgment), and empirical analysis, which consists of two stages, i.e. classical test theory with the Iteman computer software, and item response theory using the of the BilogMG/Winsteps programs (for essay questions). The results of the qualitative analysis of the final examination test instrument found that 44 items (88%) were declared Good, and six items (12%) were declared Poor based on the aspects of material, construction, and language. The quantitative analysis found that 32 items (64%) were classified as Good, and 18 items (36%) were classified as Poor. The results of the overall analysis of the test instrument found that there were 32 items that can be recommended for the procurement of test item bank for the History subject.

Keywords Final Examination Test, Test Item Analysis, Item Bank

1. Introduction

Every Indonesian citizen regardless of social status, race, ethnicity, religion, and gender, has the right to obtain quality education in accordance with their interests and talents. The existence of quality education is a prerequisite for the existence of quality Human Resources, namely intellectual citizens who are superior, competent in mastering Science and Technology, and productive in work. In addition, they should have a good moral by applying noble character in daily life, a high commitment to various social roles, and competitiveness with other nations in the global era.

To improve the quality of education, the development of national education needs to be directed at improving human dignity as a whole. Educational institutions are expected to be able to become a strategic place in an effort to develop all potential individuals, including building character and national insight for students who will become an important foundation for maintaining national oneness and unity within the framework of the Unitary State of the Republic of Indonesia. Through this, the ideals of building the character of a complete Indonesian human can be carried out properly in accordance with the objectives of the national education.

A learning experience is a learning process that involves the selection of methods used by the teacher in delivering the material, the design of classroom activities, and the achievement of the final targets that can be achieved by students. The evaluation procedure is an activity to collect the information about the learning process systematically to determine whether changes occur to students and the extent to which these changes affect the lives of students. When seen as a process, education contains three main elements which are interrelated, namely learning objectives, learning experiences, and evaluation procedures. Learning objectives refer to the state philosophy as outlined in the education curriculum. Each defined educational curriculum has been formulated with the objectives of each subject and instructional objectives in general on each subject.

Conceptually, evaluation, assessment, and measurement are different terms. Evaluation is broader in scope than assessment, while assessment is more focused on certain aspects. For example, if the aspect assessed is a learning system, then the scope is all components of learning. The

right term for assessing the learning systems is evaluation. However, if the objects are some components of learning, for example, learning outcomes, then an assessment is the more appropriate term. Evaluation and assessment are qualitative, while measurements are quantitative (scores / numbers). Measurements are obtained by using a measuring instrument in the form of either a test or non-test. Nevertheless, if an assessment has been carried out, then it can be assumed that measurement has basically been conducted, as well.

The Regulation of Minister of Education and Culture No. 23 of 2016 on Educational Assessment Standards explained that assessment of education is the process of collecting and processing information to determine the achievement of students' learning outcomes. Assessment of student learning outcomes is carried out based on educational assessment standards that apply in the national scope. Educational assessment standards are national education standards which are related to the mechanisms, procedures, and instruments for evaluating students' learning outcomes.

The test instrument is used to measure student competency achievement. The result of the measurements is an overview of learning outcomes regarding the degree of achievement of student competencies. A test instrument is considered to be good if it meets certain rules, can provide accurate data in accordance with its functions, and only measures certain behavioral samples. One of the characteristics of a good instrument is that it at least meets the validity and reliability requirements. An instrument is considered valid if they measure what they want to measure precisely, and it is reliable if it has consistent test results.

Test instruments can be in the form of test or non-test. If it is a test, teachers must make a set of question items, while if it is a non-test, the teacher can make a questionnaire, observation guidelines, interview guidelines, study documentation, talent assessment, and so on. In summary, a test is a device for obtaining a sample of an individual's behavior [1]. Likewise, a test is a way or tool to conduct an assessment in the form of a task or series of tasks that must be done by students or groups of students in order to produce grades about student behavior or achievement [2] [3].

Tests are usually administered by giving questions in the form of multiple choice questions and essay questions. Developing question items is an elaboration of indicators into questions in which the characteristics are in accordance with the blueprint guidelines. The quality of the items will determine the quality of the test as a whole. Before the test is used, it needs to be tested first in the field to find out which questions need to be changed, corrected, even discarded, and which questions are good to use [4] [5].

The specific requirements of a test instrument will be used as an evaluation tool so that the test instrument can provide an accurate picture of the students' competency achievement. A good test is composed of good items [6]. To get good items, an analysis of the items in a test instrument is needed. Item analysis can determine the item's difficulty level, item discrimination, and the effectiveness of the deception; in addition, the validity and reliability of the test instrument can be identified, as well [7] [8]. There are two types of item analysis, namely qualitative (theoretical) analysis and quantitative (empirical) analysis. The qualitative (theoretical) analysis is also called the item review. Studies are carried out by covering material, construction, and language aspects [9].

Quantitative analysis (empirical) is an analysis aimed to find information about the difficulty level, item discrimination, deception function, and reliability of the items. There are two main theories used to examine the quality of questions, namely classical test theory and item response theory [10]. Although classical test theory has a simple approach in analyzing items, it has a disadvantage as the item's characteristics depend on the ability of the group working on the problem. The item response theory, which is also considered the modern test theory, is present to complement the shortcomings of the classical test theory. The parameters measured in item response theory are the student abilities and the test items. The item parameters consist of the item discrimination index, item difficulty index, and guessing. One of the criteria for a good test instrument is that it has relatively small guessing [11].

As one of the techniques for evaluating student learning outcomes, a test has an important role in measuring student achievement. Tests can be done periodically in the form of daily tests, midterm examination (UTS), and final examination (UAS). A semester evaluation of student learning outcomes is conducted through the final examination. The grade achieved in the final examination is a depiction of the mastery of competencies that students learn for one semester. Therefore, a more extensive problem is needed [12].

After the enactment of regional autonomy, the implementation of educational evaluation is carried out by the Department of Education in the provincial or regional level. For example, the implementation of a joint semester examination, where the Department of Education or the Principal Work Association (MKKS) provides the final examination questions and distributes them to all the public schools in the region to measure the success of student learning outcomes [13]. Sometimes, the Department of Education assigns the development of the final examination questions to the Subject Teachers Association (MGMP) assisted by the MKKS.

Referring to the results of the preliminary survey on one of the high school History subject teachers in Yogyakarta who joined the MGMP, the test instrument of the final examination or grade passing examination used in senior high school is made by MGMP. However, the development of the test instrument is not done through a predetermined test development procedure because after the test is developed, the test instrument is immediately used for the

final examination. There is no information on the characteristics of each test item because no analysis whether qualitative (theoretical) or quantitative (empirical) is made on the test instrument, especially the one used in the school with the 2013 curriculum.

The preliminary study also obtained information that there was a lack of good quality test items of the History subject in the city of Yogyakarta, especially those used for schools with the 2013 curriculum. The lack of availability of good quality items would make it difficult for History teachers in Yogyakarta to conduct an assessment of student learning outcomes. The final examination test instrument for the History subject for the 2013 curriculum compiled by MGMP is based on the existing content standards, but it is still necessary to analyze the quality of the items. The purpose is to ensure that the quality of the test items made by MGMP have good characteristics so that they can truly measure the desired competencies.

Analysis of items in a test instrument aims to identify the shortcomings of the test instrument or to improve learning. Based on the objective, the item analysis have many benefits, including 1) helping test users in the test evaluation; 2) providing a relevant source for the development of informal and local tests; 3) supporting the development of effective test items, improving the test in terms of materials, improving the item's validity and reliability; (4) determining whether the function of an item is on the right track; 5) providing input to the teacher about students' difficulties; 6) providing input on certain aspects for the curriculum development and revision on the assessed or measured materials [14] [15] [16].

In response to the limitations of a good test instrument quality on the eleventh-grade History subjects, especially those used for the 2013 curriculum in Yogyakarta, the research is very important to do. The test instrument used to evaluate historical learning achievements needs to be analyzed so that the quality is identified. The analysis results of the test instrument can later be the first step in developing the item bank on high school History subjects. Thus, it is necessary to analyze the quality of high school History Final Exam in the 2013 curriculum to help teachers improve the assessment quality of the student learning achievements. Based on the description above, this study focuses on the quality analysis of the Final Exam test instruments in high school History subject in the Yogyakarta city. The problems formulated in this study are as follows: 1) how is the quality of the Final Exam test instrument for Grade 10 History subject in the 2013 curriculum used in Yogyakarta, 2) how many items are covered in the Final Exam questions in Grade 10 History subject in the 2013 curriculum in Yogyakarta, which can be recommended for the development of the item bank.

2. Materials and Methods

The research is an explorative descriptive study that uses a quantitative descriptive approach to determine the quality of test instruments for the final exam used in the 2013 curriculum in the 2016/2017 academic year in Yogyakarta. The aim is to obtain items that meet the criteria of validity and reliability so that they can be recommended for the development of the item bank.

The research was carried out through the following stages: 1) collecting final exam test instruments for History subjects made by Subject Teacher Association in Yogyakarta. The test instruments used in the final exam for the odd semester of the 2016/2017 academic year was obtained by collecting the archives from the department of education and high schools implementing the 2013 curriculum, 2) collecting the students' final exam answer sheets, 3) test item validation done by several assessors or raters to measure the coefficient of content, using the Aiken's formula, 4) test item qualitative analysis through expert judgment by practitioners such as high school History subject teachers and instrument experts, 5) conducting quantitative analysis using a classical and response test theories, and 6) selecting good quality questions analyzed to be developed into a question bank. The good item categories should be based on the expert judgment and should be accepted by Iteman and Bilog MG or Winsteps (essay questions); questions that were not accepted in a stage of analysis were considered null and were not developed in the item bank, and 7) labeling test items with question cards.

There were three sources of data used in this study. Firstly, the data were obtained from final exam test items of History subjects of Grade 10 in the 2013 curriculum taken from the odd semester of the 2016/2017 academic year. There were three types of questions in the test instrument, namely multiple choice, short answer, and essay questions. Secondly, the data were taken from students' responses to the test instruments describing the ability of Grade 10 high school students in the History subject. For students' responses, the sample was taken using the Stratified Random Sampling technique. At last, the data are from the results of the study and theoretical validation from 12 experts consisting of evaluation experts and teachers.

The variables in this study were test instruments for Grade 10 senior high school History subjects and students' responses to the test. The sub-variables of this study were: 1) item difficulty (a point scale on the participant's ability); 2) item discrimination (questions to distinguish students with high abilities and students with low abilities; 3) effectiveness of distractors (function as deceptive items) seen from a minimum of 5% of participants or students who selected the right choices of test items; and 4) question information function, which is a measure that shows the reliability index on an item response theory. The information function presents the contribution of test items in revealing the latent trait measured by the test.

The data in this study were collected using the documentation technique. Validation sheets were used to

obtain the expert judgment as well as the results of an empirical analysis of students' responses to the test instruments of History subject. The data collection consisted of 45 multiple choice questions and five essay questions, review and validation sheets analyzed by experts, and responses of the test instrument from the students.

The data analysis in this study was carried out through two theories with three stages, i.e. theoretical analysis (reviewing the items by the expert judgment) and empirical analysis. Theoretical analysis of content validation with the Aiken's formula item test history tool. Aiken's V formula can be used to calculate the content-validity coefficient based on the results of an expert panel's evaluation of n people on an item in terms of the extent to which the item represented the measured construct. The research decision was based on data support from three reviewers, with the following decision provisions: a) Good, if all the criteria for reviewing the items in the material, construction, and language aspects were all in accordance with the prescribed rules, and supported by all reviewers, b) Good Enough , if the whole criterion for reviewing the items in the material aspect is in accordance with the specified rules, as much as possible there is one criterion in the construction aspect and one criterion in the language aspect that is not in accordance with the prescribed rules, at least supported by two reviewers. If the answer key is wrong, or there are criteria on material aspects that are not in accordance with those determined by more than one criterion in the construction aspect and more than one criterion in the language aspect that is not in accordance with established rules, it is supported by at least two reviewers. The statistical analysis with the formula proposed by Aiken is as follows.

$$V = \sum s / [n (C-1]$$

where

S = r - lo

Lo = lowest rating score (e.g. 1)

C = highest rating score (e.g. 4)

R = score given by reviewer

Table 1. Example of Aiken's V Content Contents Validity

Reviewer	Item 1	
	Score (R)	S = R - Lo
1	3	3 – 1 = 2
2	4	4 – 1 = 3
3	4	4 – 1 = 3
4	3	3 – 1 = 2
5	3	3 – 1 = 2
6	4	4 – 1 = 3
7	4	4 – 1 = 3
	∑s	18
	V	0.857

The S value for reviewer 1 is obtained from the score of appraiser 3 minus the lowest score (Lo), so that 3 - 1 = 2, and so on. The value ∑s is the sum of the S score which is 2 + 1 + 3 + 2 + 2 + 3 + 3 = 18. Thus, the value of V can be calculated as follows:

$V = \sum s / [n (C-1]$

$V = 18 / [7 (4-1)] V$

$= 0.857$

The coefficient value of Aiken's V ranges from 0-1. This coefficient of 0.857 can already be considered to have adequate content validity.

Empirical analysis with the Iteman Program (Classical Test Theory) is used to estimate the magnitude of the level of difficulty, different items, the functioning of the deception, and the reliability index with the following conditions:

1) The level of difficulty of the items received to express good items is between 0.25 0.75. The level of difficulty is expressed in the proportion of correct answers, with the following formula:

$P = (\Sigma B) / N$

Where:

P = Difficulty level

Σ B = Number of students who answered Right

N = Number of Students taking the test

Criteria:

p <0.25 = Difficult

0.25≤p≤ 0.75 = Medium

p> 0.75 = Easy

2) The magnitude of the Power of Difference in items to state good items is one that has a value greater than 0.2 with the following formula:

$DP = (WL-WH) / n$

Where:

D P = Distinguishing Power

WL = Number of students who failed from the Lower group

WH = Number of students who failed from the Above group

n = 27% x N

N = Number of test takers

Criteria:

0.40-1 = Very Good

0.30-0.39 = Good

0.20-0.29 = Good Enough

0-0.19 = Bad

3) The deception function is said to be good if it is responded to at least 0.02 or at least 2% on the deception and the biserial value is negative. The formula is as follows:

$IP = P / ((N-B) / (n-1)) \times 100\%$

Where:

IP = Deception index

P = Number of students who choose deception
B = Number of students who answered correctly in each question
n = Alternative answer (option)
N = Number of test takers
Criteria:
76% - 125% = Very Good
51% - 75% or 126% -150% = Good
26% - 50% or 151% -175% = Not Good
0% -25% or 176% -200% = Poor
> 200% = Very Bad

4) The Reliability Index is said to be good if it is greater than 0.70. This relates to the magnitude of standard error of measurement, the greater the reliability value of the items the smaller the level of measurement error. The formula is as follows:

r ii = [k / (k-1)] [1- 〖∑ab〗 ^ 2 / (∑ 〖at〗 ^ 2)]
Where:
r ii: Reliability of the item
k: The number of items
〖∑ab〗 ^ 2: Number of item variants
∑ 〖at〗 ^ 2: Total number of variants
Kappa statistical value criteria:
<0 Poor agreement
0.0 - 0.20 Slight agreement
0.21 - 0.40 Fair agreement
0.41 - 0.60 Moderate agreement
0.61 - 0.80 Substantial agreement
0.81 - 1.00 Almost perfect agreement

Empirical analysis (Item Response Theory) using Bilog MG will show the following results: 1) Stope shows different power, 2) Treshold shows difficulty level, 3) Asymtote shows pseudo guessing, 4) Outfit item states mismatch of response to difficulty level, and 5) point biserial point The questions state the correlation coefficient between the students' answers to each item from all students and the total score. In addition to the MG Bilog also used the Winsteps program where the scoring model used is a PCM (partial credit model) with the Rasch model approach, because only the level of item difficulty is seen. In the first stage, classical test theory analysis was done using Iteman computer program. Then, in the second stage, BilogMG/Winsteps computer program was used to conduct response theory analysis for essay questions.

3. Results and Discussion

Content validity is used to measure the test instrument. One of the frequently used validity formulas is Aiken's. The Aiken's validity is based on the results of the expert panel's assessment item in terms of the extent to which it represents the measured construct. This study measures the content validity of the items using twelve-panel assessors consisting of seven experts of History learning evaluation and five History teachers of senior high school. The aim of the Aiken's content validity analysis is to measure how accurate the test items are in performing their measuring functions.

The results of the content validity analysis using the Aiken's formula can be seen in the following table.

Table 2. Results of the Aiken's Validity

Test Item	Valid Item	Invalid Item
Multiple Choice	42	3
Essay	5	0

Based on the results, the item decisions are valid, quite valid, and invalid, i.e., if an item has a validity value of>0.7, it is declared valid; the value of 0.5 - 0.7 is declared quite valid, while the value of <0.5 is declared invalid. It is found through the analysis of the test items in History subject that 28 multiple choice questions, item number: 1, 4, 5, 6, 9, 10, 12, 14, 15, 17, 19, 20, 22, 23, 26, 27, 28, 31, 33, 34, 35, 37, 38, 39, 40, 42, 43, and 44 and three essay questions, items number: 46, 47, and 49 are declared valid (high). Meanwhile, 14 multiple choice questions, item number: 2, 3, 7, 8, 11, 16, 18, 21, 24, 25, 30, 32, 36, and 45 and two essay questions, item number: 48 and 50 are claimed quite valid (moderate). Then, three multiple choice questions, item numbers: 13, 29, 41 are categorized as invalid (low).

3.1. Theoretical Analysis Results

A theoretical (qualitative) analysis is carried out to see three aspects of the test, namely material, construction, and language. This analysis involves 12 people to examine the test items. For aspects of material appropriateness, construction, and language, there are seven evaluation experts and five History subject teachers of senior high school. The research decision is based on the supporting data from 12 reviewers, with the following conditions.

Test items are accepted if they meet the good and sufficient criteria. The criteria of a good item involve all assessment indicators such as material, construction, and language aspects. For items with sufficient criteria, they should meet the criterion of the material aspect. Moreover, they should have, in maximum, one sub-indicator not meeting the criteria in terms of construction and language aspects, and they should be at least supported by seven reviewers.

The test items are rejected if the ones included in the criteria are not good. The test items categorized as not good are the ones in which answer key is wrong, or there are criteria on material aspects that are not in accordance with what is specified. More than one criterion in the aspects of construction and language should be at least supported by seven reviewers.

The results of the expert judgment in this study can be seen in Table 3.

Table 3. Summary of the Review by Experts

Test Item	Valid Item	Invalid Item
Multiple Choice	41	4
Essay	5	0

Based on the summary, the results of a theoretical review by the experts show that of the 46 out of 50 items with two types of questions are acceptable. The items are accepted by 12 assessors consisting of evaluation experts and teachers. They conform the standards in terms of material, construction, and language aspects, while the other four items, item numbers 4, 7, 13, and 42, are rejected. The rejected items are considered poor according to some aspects of the study. They do not meet the material criteria and are not in accordance with the indicators of competency achievement and construction of developing good questions. Thus, only 46 items consisting of 41 multiple choice items and five essay items have met the standard test item development based on the material, construction, and language aspects.

3.2. Empirical Analysis Results

In this study, the characteristics of the Final Exam test instrument in History subject are identified in two ways by using three computer programs, namely the Iteman 3.00 to analyze the classical test theory, Bilog MG to determine the characteristics of test instruments by using an item response theory, specifically for the type of essay questions in data, using a Winsteps computer program with the Rasch Model approach.

3.3. Analysis Using Classical Test Theory

The data analysis using Iteman 3.00 is used to find out information on the characteristics of the test instrument in the form of the difficulty level of the items (indicated by the proportion of students having correct answers), discrimination index (indicated by the biserial point correlation) and the effectiveness of distractors. The data analysis with Iteman 3.00 is done separately for multiple choice and essay questions. This is done on the consideration that the data obtained from the two types of questions have a different pattern. Moreover, the data can actually represent each item one by one [17].

The aim of the analysis using the classical test theory approach is to find out the characteristics and empirical quality of items. This analysis shows the characteristics of the items and test instrument statistics, namely: 1) the statistics of the questions including the level of difficulty, discrimination index, and distractor effectiveness; and 2) the statistics of the test instrument including mean, median, reliability index, skew, and standard error of measurement [18].

3.4. Multiple Choice Questions

Based on the results of the analysis using the Iteman program, the difficulty level of the test items for the History subject of Grade 10 of senior high schools in Yogyakarta can be seen in Table 4.

Table 4. The Level of Difficulty of Multiple Choice Questions

Category	Item Number	Total	Percentage
p>0.75			
Easy	3, 4, 9, 11, 13, 22, 27, 35, 39, 41, 45	11	24.44
0.25≤p≤ 0.75			
Medium	1, 6, 7, 8, 14, 15, 16, 17, 18, 19, 23, 24, 25, 29, 30, 31, 32, 33, 34, 36, 37, 40, 42, 44	24	53.34
p<0.25			
Difficult	2, 5, 10, 12, 20, 21, 26, 28, 38, 43	10	22.22
Total	45	100	

Based on the table above, it can be concluded that there are eleven items (24.44%) in the Easy category, 24 items (53.34%) in the Medium category, and ten items (22.22%) in the Difficult category. This means that most of the test items have a medium level of difficulty. In general, the index of difficulty of items should be at the intervals of 0.3 - 0.7. Index of difficulty or P is the abbreviation of “proportion” which shows that the higher the item difficulty index is, the easier the items are [16].

The result of item discrimination analysis using Iteman program can be seen in Table 5 below.

Table 5. The Item Discrimination Analysis of Multiple Choice Questions

Category	Item Number	Total	Percentage
>0,20			
High	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45	42	93.3
<0,20			
Low	11, 21, 34	3	6.7
Total	45	100	

Based on the table above, 42 items (93.3%) have high item discrimination index, while three items (6.7%) have low item discrimination index. The value showing the item discrimination is presented in letter D. The item discrimination analysis is intended to distinguish between the high achievers and the low ones [19] [20]. The minimum value of the discrimination index is 0.3. Then,

the higher the discrimination index is, the better the item is. In determining the discrimination index, biserial correlation index and alignment index may be used.

The results test item distractor analysis using Iteman program can be seen in Table 6 as below.

Table 6. Multiple Choice Item Distractor Analysis

Category	Item Number	Total	Percentage
>2% and r_pbis negative			
(Good)			
	1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 40, 41, 42, 43, 44	41	91.11
<2% and r_pbis positive			
(Poor)	8, 16, 39,45	4	8.89
Total	45	100	

Based on the item analysis result presented in the table above, 41 items (91.11%) have Good distractors because the distractors chosen by the test participants are more than 2% and the r_pbis is negative. Meanwhile, Poor distractors (8.89%) are chosen by less than 2% of the test participants or r_pbis positive (except the answer key). There is no test item with "check the key" caution due to the students' lack understanding of statements given in the formulation of the questions. The analysis of distractor effectiveness is used to determine whether the distractors work well. The distractors maybe categorized effective if they are chosen by 5% of the test participants from the lower group.

3.5. Essay Questions

Based on the results of the analysis with the Iteman program, the item difficulty of the test items are shown in Table 7 below.

Table 7. Level of Difficulty of Essay Questions

Category	Item Number	Total	Percentage
p>0,75			
Easy	-	-	-
0,25≤p≤ 0,75			
Medium	1,2,3,5	4	80
p<0,25			
Difficult	4	1	20
Total	5	100	

The table above shows that there are four items (80%) in the Medium category, one item (20%) in the Difficult category, and no item in the Easy category. It indicates that all essay questions have a Medium level of difficulty.

The summary of the accepted and rejected test items analyzed using the classical theory approach, and Iteman program is presented in Table 8 below.

Table 8. Summary of Item Analysis Using Classical Test Theory

Test Item	Accepted Item	Rejected Item
Multiple Choice	33	12
Essay	5	0

Based on the summary of the item analysis, the number of items accepted and rejected is presented in details. The accepted items are 38, consisting of 33 multiple choice questions and five essay questions. The rejected items are 12, and they are all in the form of multiple choice questions. The rejected ones are item numbers: 4, 7, 13, 14, 16, 19, 21, 24, 29, 33, 41, and 42.

3.6. Analysis Using Item Response Theory

The analysis using the item response theory approach for multiple choice questions is done with Bilog MG (a computer program). Meanwhile, the essay questions are analyzed using Winsteps (Rasch Model) computer program. As stated in the previous part of this paper, the analysis models used in this research are reviewed before item analysis is conducted. Thus, the model with a certain kind of parameters conforms to analysis. To determine the suitability of the model, the criteria value of prob >0.05 is used. This method is used to check the output of analysis using Bilog MG with parameter 1, 2, and three models.

The results of the analysis show that the number of items matching the one-parameter model is 16. Moreover, 31 items match the two-parameter model, and 42 items match the three-parameter model. Thus, the test instrument analyzed in this research is suitable for the three parameter model. This model has three parameters, namely item difficulty, item discrimination, and guessing factor. Previous study [21] reveals that the three-parameter model is more comparable and more highly correlated with classical test theory statistics than the one parameter or 2 parameter models when the classical test theory and item response theory are used together to analyze the test items. The next analysis is finding the levels of item discrimination, item difficulty, and pseudo-guessing.

The analysis using Bilog MG is through two stages. The first stage is initial analysis in which items having biserial prob of <0.5 are deleted. In the second stage, items having prob value of >0.5 proceed to check the level of difficulty, item discrimination index, and pseudo guessing chance [22] [23]. Of 45 items being analyzed, there are only 45 items having prob > 0.5, and five items are deleted.

There are 40 fit items (items having output more than 0.5). All those items are included in the next stage of analysis. The results of the analysis obtained in this

second stage can be seen in stage 2. In stage 2, information about item discrimination, level of difficulty, and pseudo guessing chance of the items are obtained.

3.7. Multiple Choice Questions

To interpret the results of the analysis in terms of the level of difficulty, criteria of the item response theory approach are developed. The ranges of value meeting the criteria are -2 to 2. Items having the difficulty value less than -2 are categorized as too easy, while items having value more than two are too difficult [22] [23]. If the item is too difficult, the test participants tend to guess. Based on the results of the analysis using the Bilog MG program, in terms of the level of difficulty, the test items are in the Medium category, which is in the Good category.

The decision criteria for the value of item discrimination index ranges from 0.0-2.0 [24]. The ones in which the value of the discrimination index is more than 2.0 are Good items, while the lower discrimination index indicates that the items are Poor. Based on the result of analysis using Bilog MG, it is found that 37 out of 40 items are in the Good category. Three items, numbers 3, 7, and 20, are in the Poor category. Therefore, most of the multiple choice questions have Good discrimination indices. The high discrimination value shows a better discrimination index [25].

The criteria for pseudo guessing values range from 0 to 0.20. A high pseudo guessing value indicates that test participants guess the answer correctly. On the other hand, the low pseudo guessing value means that it is less likely that test participants guess the answer correctly. The results of the analysis with the Bilog MG program show that 34 out of 37 items analyzed using Bilog MG, pseudo guessing, or guessing factors are correctly answered. There are three items (numbers 10, 13, and 21) in the Poor category. It can be said that most of the multiple choice questions have little probability to be answered correctly.

3.8. Essay Questions

In the test instrument, there are five essay questions analyzed using Winsteps program with the Rasch model approach. The scoring model, PCM (Partial Credit Model) with the Rasch model approach, is applied to check the level of difficulty of those five essay questions.

Item scoring for multiple choice questions ranges from 1 to 5. Zero or no score is given when item number 1 is not answered. The wrong answer is worth 1 point. The correct but incomplete answer is worth 3 points. At last, the correct and complete answer is worth 5 points. For item number 52, no answer is given 0, score 1 is for one answer, and two is for two answers. For item number 53, 0 is given to those with no answer, score two is given if one theory is mentioned, score four is given if two theories are mentioned, and a completely correct answer is scored 5. Scoring model for items number 54 and 55 is similar to that of item number 53. The scoring model provided for the multiple choice questions maybe different, but they basically have the same minimum and maximum scores.

The results of the analysis by Winsteps generate information related to the scores obtained, estimation of the test takers' ability (measure of ability), estimated difficulty index of the items (measure of difficulty) in the form of a standard scale known as logit (log add unit), standard error of measurement (SEM), the conformity of the model (infit and outfit), and its test information function (TIF) [9]. The followings are the conformity analysis results of the essay questions appearing in the final semester test instrument based on the analysis with Winsteps as presented in Table 9.

Table 9. Essay Questions that Conform and do not Conform to the Rasch Model

Criteria	Item Number.	Total Item
Outfit < 2.00 & Z-standard is positive (Items conform to the model)	51, 52, 54	3
Outfit > 2.00 & Z-standard is negative (Items do not conform to the model)	52, 55	2
Total	5	

It is apparent from the analysis results that three items conform to the model, namely, item numbers 51, 53 and 54, and items that do not conform to the model are 2, numbers 52 and 55. Such conformity of an item with a model can be seen from the chi-square value of the item compared to the critical values of the chi-square distribution in accordance with the corresponding item at the significance level of 0.01 or 0.05. Were its chi-square value smaller than the critical values of the chi-square distribution, the item is found to conform to the model [18].

The distribution of the essay questions' difficulty index as analyzed through the Winsteps program is presented in Table 10.

Table 10. The Difficulty Index of Essay Test Items Conforming to Rasch Model

Criteria	Item Number	Total Item
Difficult (> 2.00)	-	-
Medium (-2.00 s/d 2.00)	51, 54	2
Easy (< -2.00)	53	1
Total	3	

Based on the table above, two of three items conforming to the model have Medium difficulty index or categorized as Good test item, but one, number 53 is relatively easy, so it is not in the Good category. The difficulty index is the same as a matrix to one's expertise or trait [20] and a good difficulty index ranges between -2 to +2 [23].

As found by the item response theory, Bilog MG for multiple choice questions, and Winsteps for the essay items, the analysis results can be seen in Table 11.

Table 11. The Item Analysis Results Using Item Response Theory

Test Item	Accepted Item	Rejected Item
Multiple Choice	40	5
Essay	3	2
Total	43	7

As clearly seen in Table 10, 40 multiple choice items have a good discrimination index while five others indicate faulty items. Additionally, three out of five essay questions have fulfilled the criteria of accepted items. Concerning this, a research suggests that there is an inverse relationship between test information function (TIF) and standard error of measurement (SEM) [26]. This means that in the item response theory, the higher the TIF numerical value, the smaller the SEM. In the analysis through this item response theory, the accepted items are 43 in total, consisting of 40 multiple choice questions and three essay questions, while seven items are rejected, namely five multiple choice questions and two essay questions. Each item might have an information function, and its number seems to represent an information function too so that the TIF value will be high if the constituent items have certain information function as well [27].

Based on the validity analysis by the Aiken's, 47 out of 50 multiple choice questions met the validity criteria were in the Poor and Good categories. Meanwhile, based on the theoretical (qualitative) approach, 46 out of 50 items analyzed are categorized as Good in terms of the material, construct, and language. Furthermore, 50 multiple choice questions and descriptions, after going on such qualitative analysis, were deliberately analyzed by means of quantitative technique to see how the information of the items in their entirety to be taken into consideration by the test developers in the future. With the same question format for the same participant ability, an overview of information about the item difficulty index and discrimination index are provided. This is in line with the notion that information about item quality is very crucial for teachers to motivate students [14] [15].

Additionally, the Iteman results show that 38 (76%) of the total of 50 items analyzed are considered as Good. The Bilog MG program analysis for multiple choice questions similarly suggests that 41 out of 45 multiple choice questions analyzed with three parameter model are Fit. Moreover, in the Rasch model (Winsteps), 3 out of 5 essay items analyzed are found to be Fit. After going through all stages of analysis, at the end of the results, a total of 32 items are suggested for the procurement of the item bank, consisting of 29 multiple choice and three essay questions.

The importance of the item bank has been put forward by several research, namely to ensure the quality of test instruments used by teachers and to monitor the quality of education [28] [29] because it is composed of items with identified, predicted, and ensured as well as reliable quality so that they can be used to compile new tests or sub-tests [30]. These items are not merely a collection of questions, but calibrated questions that can be used to provide information about the trait of the test takers [31].

4. Conclusions

Based on the results and discussion in the previous parts of the paper about the analysis of the quality of the multiple choice and essay test items both theoretically and empirically description questions, three conclusions are formulated as follows.

Firstly, the analysis on the final semester test items of the History school subject in Grade 10 senior high schools in Yogyakarta in the 2013 curriculum which was analyzed qualitatively (theoretically) suggests that out of 45 the multiple choice and five essay questions 44 questions are stated to be Good by 12 experts. Items declared Poor are rejected since they do not conform to the indicators of learning outcomes or test construct.

Secondly, the quantitative analysis results suggest a couple of points, namely Aiken's validity and Iteman analysis which found that most of the items (at least 76%) can be perceived as having Good discrimination index and effective distractors. Meanwhile the rests may be too easy or hard or have low discrimination index and ineffective distractors so that they fail to differ test takers' trait and ability. Further Bilog MG for the multiple choice questions and Winsteps for the essay questions find that 86% of the items are accepted while others (14%) have low discrimination index, difficulty index, and guess chance.

Lastly, as suggested by both (theoretical) qualitative and (empirical) quantitative analyses of Aiken's validity, classical test and item response theory, 29 multiple choice and 3 essay questions can support the procurement of item bank for the History subject in grade 10.

Acknowledgements

The author is very grateful to Universitas Negeri Yogyakarta for funding the research and to experts for their appropriate and constructive suggestions to improve this paper.

REFERENCES

[1] M. J. Allen, W. M. Yen. Introduction to Measurement Theory, Brooks/Cole Publishing Company, Monterey, CA., 1979.

[2] Djaali. Hasil belajar evaluasi dalam evaluasi pendidikan: Konsep dan aplikasi. Uhamka Press, Jakarta, 2006.

[3] C.R. Prihantoro. The perspective of curriculum in Indonesia on environmental education, International Journal of Research Studies in Education, Vol. 4, No. 1, 77-83, 2015.

[4] A. Garino, J. Van Rhee. Test item analysis for the physician assistant educator, The Journal of Physician Assistant Education, Vol. 20, No. 3, 22-27, 2009.

[5] D. Mardapi. Pengembangan Instrumen dan Kisi-kisinya. Universitas Negeri Yogyakarta, Yogyakarta, 2011.

[6] H. Retnawati. Teori Respon Butir dan Penerapannya. Nuha Medika, Yogyakarta, 2014.

[7] D. Mardapi. Pengukuran Penilaian dan Evaluasi Pendidikan, Nuha Medika Yogyakarta, 2012.

[8] R.J. Cohen, M.E. Swerdlik. Psychological Testing and Assessment (6th Edition), The Mc Graw-Hill, Boston, 2005.

[9] T. Kubiszyn, G. Borich. Educational Testing and Measurement: Classroom Application and Practice. John Wiley & Sons, Hoboken, NJ, 2009.

[10] N. Guler, G.K. Uyanik, G.T. Teker. Comparison of classical test theory, an item response theory in terms of item parameter. European Journal of Research on Education: International Association of Social Science Research, Vol. 2, 1-6, 2014.

[11] N.S. Aminah. Karakteristik metode penyetaraan skor tes untuk data dikotomos, Jurnal Penelitian dan Evaluasi Pendidikan, Vol. 16, 88-101, 2012.

[12] T. Takeshi. Developing communication skill in teaching of history as part of social sciences, International Journal of Social Studies, Vol. 2, No. 1, 37-48, 2006.

[13] T.R. Moon, C.M. Brighton, C.M. Callahan, A. Robinson. Development of authentic assessment for the middle school classroom. The Journal of Secondary Gifted Education, Vol. 16, No. 2-3, 119-135, 2005.

[14] W.A Mehrens, I.J. Lehman. Measurement and Evaluation in Education and Psychology, Holt, Rinehart, and Winston, Inc., New York, 1973.

[15] P.W. Miller. Measurement and Teaching, Patrick W. Miller & Associates, Indiana, 2008.

[16] C.R. Reynolds, R.B. Livingston, V. Willson. Measurements and Assessment in Education (2nd Edition), Pearson, New York, 1999.

[17] S. Azwar. Reliabilitas dan Validitas, Pustaka Pelajar, Yogyakarta, 2015.

[18] R.L. Linn. Educational Measurement, McMillan Publisher, New York, 1989.

[19] B.A. Uno, H. Sofyan, I.M. Candiasa. Pengembangan Instrumen untuk Penelitian. Delima Press, Jakarta, 2001.

[20] D. Mardapi. Teknik Penyusunan Instrumen Tes dan Non Tes, Mitra Cendikia, Yogyakarta, 2008.

[21] N. Abdelaziz, Leng, C.H. The relationship between CTT and IRT approaches in analyzing item characteristic, The Malaysian Online Journal of Educational Science, Vol. 1, No. 1, 64-70, 2013.

[22] C. Demars. Item Response Theory, Oxford University Press, New York, 2010.

[23] R.L. Hambleton, H. Swaminathan. Item Response Theory: Principles and Application, Kluwer, Boston, 1985.

[24] S. Azwar. Tes Prestasi: Fungsi dan Pengembangan Pengukuran Prestasi Belajar (Edisi Kedua), Pustaka Pelajar, Yogyakarta, 2015.

[25] J.S. Kim. Using the distractor categories of multiple-choice items to improve IRT linking, Journal of Educational Measurement, Vol. 43, No. 3, 193-213, 2006.

[26] A. Moghadamzadeh, K. Salehi, E. Khodaie. A comparison of the information function of the item and test in one, two, and three parametric models of the item response theory (IRT), Procedia - Social and Behavioral Sciences, Vol. 29, 1359-1367, 2011.

[27] R.L. Hambleton, H. Swaminathan, J.H. Rogers. Fundamentals of Item Response Theory. Sage Publications, London, 1991.

[28] R.G. MacCann, G. Stanly. Item banking with embedded standards, Practical Assessment, Research, and Evaluation, Vol. 14, No. 17, 1-8, 2009.

[29] J. Umar. Item banking, in Advances in Measurement in Educational Research and Assessment, 207-219, 1999.

[30] G.K. Eid. The effect of sample size on the equating test items, Education, Vol. 126, No. 1, 2005.

[31] J.B. Bjorner, C.H. Chang, D. Thissen, B.B. Reeve. Developing tailored instruments: item banking and computerized adaptive assessment. Quality of Life Research, Vol. 6, No. Supplement 1, 95-108, 2007.