Research and Evaluation in Education Journal
Volume 1, Number 1, June 2015 (73-83)
Available online at: http://journal.uny.ac.id/index.php/reid

AN ASSESSMENT MODEL OF HISTORICAL THINKING SKILLS BY MEANS OF THE RASCH MODEL

[1)]Ofianto; [2)]Suhartono
[1)]Padang State University, Indonesia; [2)]Gadjah Mada University, Indonesia
[1)]ofianto.anto@yahoo.com; [2)]suhartono@ugm.ac.id

Abstract

This study was conducted to produce a model and instruments of historical thinking skills in the history subject at the senior high school (SHS) and to identify SHS students' historical thinking skills. The study was conducted in two stages, namely model development and instrument development altogether with a small-scale tryout and a large-scale tryout. The test for each tryout consisted of six and five sub-test sets. Each test set contained 20 anchor items. The sample for each tryout comprised 1573 and 2613 testees. The data was analyzed by means of Partial Credit Model (PCM) using the QUEST program. The overall tryout results indicate that, based on the criteria for an INFIT MNSQ mean of 0.1 and a standard deviation of 1.0, the tests fit the PCM. The reliability coefficients of the tests for the tryouts are moderately good; the Cronbach's alpha coefficients are, respectively, 0.65 and 0.54. The lowest score of historical thinking skills is -.352 and the highest is +1.21 in an ideal range of -4.0 to +4.0. In overall, the testees' scores are not satisfactory. Only 5.89% of the testees are above the expected median.

Keywords: *instrument development, test, historical thinking skills, polytomous, PCM*

Introduction

Assessment is an important component in the operation of an education. An assessment is conducted in order to view and to monitor the development of educational quality from one period to another (Alen & Yen, 1997, p. 2; Griffin & Nix, 1991, p. 4). Therefore, in order to perform an assessment toward the educational quality, teachers might use multiple assessment tools. The assessment tools might be in the form of test and non-test (Mardapi, 2008, pp. 2-3). The use of multiple assessment tools is intended to portray the learning results comprehensively. Thereby, the assessment will be useful for viewing the educational quality in overall and the assessment will also provide important information for improving the learning process.

An assessment technique in the form of a test is a measurement activity because through a test, a teacher might attain numerical data for improving the learning participants' characteristics capability (Hargreaves & Schmidt, 2002, pp. 69-95). One of the learning subjects taught from the elementary schools to the senior high schools is history. The history subject in the schools aims to attain the historical thinking skills (Fogu, 2009, pp. 103-121), to encourage the learning participants to be critical-analytical (Winerburg, 2006, pp. 3-6) and to benefit the knowledge about the past in order to comprehend the life in the present time and in the future time.

According to the Ministry of National Education Regulation Number 20 Year 2007 (Depdiknas, 2007, pp. 1-2) regarding the assessment standards for the elementary and the high education, the assessment of history learning results contains three aspects: academic, historical awareness, and nationalism. In performing the assessment in the schools, teachers should pay attention to the compatibility between and among the standards (the competencies), the contents (the curriculum contents), the assessment and the learning strategies (Ashby & Shemit, 2005, pp. 150-163).

The analysis on the learning results is also important information for improving the learning process; therefore, the psychometric experts develop an analysis model known as test theory (Rasch, 1961, pp. 321-334; Rasch, 1977, pp. 58-93). The test theory that has been developed for a long period is the classical test theory (CTT) (Van der Linden & Hambleton, 1997, pp. 4-5; Hambleton & Swaminathan, 1985, p. 5). CTT, in its estimation, contains man erros and provides little information. Within the development, in order to overcome the fundamental weakness of CTT, the experts developed the item response theory (IRT) (Master, 1999, pp. 98-109). The IRT model provides more information with more assumptions. The IRT model consisted of three models namely the Rasch model or the 1 logistic parameter (1-LP), the 2 logistic parameter (2-LP) and the 3 logistic parameter (3-PL).

Based on the results of a survey which were conducted by the researchers, the researchers found that the assessment done by the history teachers had been an objective one and had tendency of demanding the learning participants to memorize the facts. Such fact has been investigated by several aspects such as Bain (2005, pp. 179-214), Barton & Levstik (2003, pp. 358-261) and Lee (2005, pp. 31-40). The results of their investigation show that the recent practice of history assessment had been lingering on the factual memory by means of multiple choice test provision. The other fact that these researchers found is that the written test, as one of the assessment tools that had been implemented up to date in order to uncover the students' capability or learning results, was constructed insystematically. As a result, many tests that the teachers provide cannot uncover the learning participants actual capability. The results of a study by Mardapi et al. (1999, p. 45) found that there had been many teachers who did not pay attention to the test guidelines while making the test items; instead, they tended to use the test items from the books circulated in the market.

In relation to that matter, the teachers should habituate themselves in implementing

the other test form, such as essay, that will be more appropriate for the subject characteristics and for the learning objectives that have been formulated. The demand within the formulation of one of the Basic Competence (BC) in the content standards of the national curriculum for the Senior High Schools/Madrasah Aliyahs is that the learning participants will be able to develop their ability in understanding and implementing the basic principles of inquiry, which has been the application historical thinking skills in the history subject.

Historical thinking skills might be defined as a scientific steps/process in studying the history (Seixas & Peck, 2004, pp. 109-117; Seixas, 2013, pp. 10-12). In each process of historical thinking skills, there will always be thinking process. Thereby, the historical thinking skills might also encourage the development of critical and creative thinking capabilities within the learning participants.

Based on the explanation, in order to measure the historical thinking skills, the researchers would like to provide an essay test. Therefore, the researchers should arrange an instrument of historical thinking skills that consists of a test and an assessment guideline. As a result, the researchers are encouraged to perform a study on the instrument development for measuring the learning participants' historical thinking skills that consists of a test and an assessment guideline.

Method

The study was a developmental one and its aim was to develop a test on senior high school students' historical thinking skills. The development procedures and phases implemented by the researchers in the study referred to the research and development proposed by Borg & Gall (1989, p. 227). However, the stages were made appropriate to the objectives and the importance of the study. Then, the stages of the research and development study were as follows: (1) needs analysis and preliminary investigation; (2) model planning and design; (3) model experiment; (4) evaluation; (5) implementation; and (6) dissemination.

The needs or problem analysis and the preliminary investigation were conducted in the form of direct observation/survey and literature or library study. The results of these activities were made as the basis of designing the initial draft of the test/assessment model.

In the model design, the researchers developed a test of senior high school students' historical thinking skills. According to Oriondo & Dallo-Antonio (1984, p. 34), the stages of test development include: (1) test design and (2) test experiment. The activities of test design were conducted until the drafting of the test that would be ready for the experiment.

The activities of instrument/ test designing included: (a) arranging the learning continuum (LC); (b) preparing the guidelines of historical thinking skills test/ instrument; (c) writing the test items; and (d) improving/ revising the test items and drafting the test/ instrument. The scales used were polytomous, adjusted according to the test form that would be taken, namely essay. For the polytomous scaling, the researchers implemented the scale from 0 to 2 for three categories.

The item revision was conducted after the researchers conducted a qualitative analysis toward the items that had been drafted. The qualitative analysis toward the items was not apart from the LC and the guideline. Therefore, first of all, the researchers performed a review toward the LC, the indicators, the guideline and the items by means of focus group discussions (FGD).

Next, the researchers performed a limited experiment toward the instrument of historical thinking skills that had been drafted in order to attain the empiric data. The results of the experiment were analyzed both by using classical approach and of item response theory (IRT). The analysis was performed in order to view the quality of the test items before the instrument would be re-arranged for the expanded experiment or the implementation.

Furthermore, the researchers performed the activities in the test, evaluation and revision stage. In the stage, the researchers performed an experiment toward the model that had been developed through the limited experiment. The data attained from the results of the experiment would be analyzed to decide whether the model developed had been fit or not.

The expanded experiment was performed after the limited experiment or after the instrument had been revised. The results of the expanded experiment would be analyzed to find how far the students had mastered the historical thinking skills. The final product of the model that would be developed would be disseminated to the users and the policy makers in the schools, namely: the teachers, the principals, the heads of education office in the city/ district, and the province. The dissemination would be conducted in the form of research distribution to the sample schools.

The product experiment would be performed twice namely: (1) in the form of limited experiment; and (2) in the form of expanded experiment. The activities that the researchers performed in the limited experiment were as follows: test implementation and results analysis. Then, the activities that the researchers performed in the expanded experiment were as follows: test implementation, results analysis, and results interpretation.

The study was conducted in West Sumatra province. The subjects were senior high schools students. The senior high schools involved in the study were the favorite ones located in the capitol of the province until the infavorite ones located in the capitol of the sub-district. The reason was that the researchers would like to attain maximum variability of measurement results.

The data that had been gathered in the study were quantitative one. The quantitative data were in the form of test results and the qualitative data consisted of the one from the limited experiment and the one from the expanded experiment. The data gathering in the study was performed by employing a set of test.

To measure the quality of the test instrument, the researchers performed both qualitative analysis by expert judgement from the aspects of contents (materials), construction and language and qualitative analysis by means of experimental process (empirical process). The data resulted from the experiment was analyzed with the Quest program. The objective of the analysis was to find the quality of test item parameter and the level of test reliability. The quality of test item parameter was only shown by the level of test item because the test item parameter was implemented the 1-PL model/ the Rasch model. On the other hand, the level of test reliability was performed by the score of Alpha coefficient.

The data resulted from the expanded experiment was analyzed with the Quest program. The analysis was conducted to attain information regarding the characteristics of the item parameter, the participants' capability parameter and the students' mastery toward the historical thinking skills in the school.

Findings and Discussions

Findings

The finding of this study is in the form of assessment model of historical thinking skills resulted in the study which belonged to the procedural model, namely the model that had procedures that should be performed sequentially. The phases included the test preparation, the limited experiment, and the expanded experiment.

Test Preparation

The activities of test preparation began with the formulation of learning continuum, the test guideline draft and test items composition for the historical thinking skills. Then, the researchers performed a review toward the instrument by involving several experts. The total test instruments made were six units. Those six units had 10 items as the anchor or the common items. The activities of limited experiment were performed toward the selected senior high schools and the experiment involved 1,572 learning participants from grade X and grade XI.

Table 1. The Characteristics of the Senior High Schools for the Limited Experiment of Historical Thinking Skills Test

No.	Name of Senior High School	Location	Popularity Based on the Graduates Accepted in the State University
1	1 Solok Senior High School	Solok City	Popular in Solok City
2	1 Payakumbuh Senior High School	Payakumbuh City	Popular in Payakumbuh City
3	1 Gunung Talang Senior High School	Solok County	Popular in Solok County
4	1 Batu Sangkar Senior High School	Tanah Datar County	Popular in Tanah Datar County
5	2 Solok Senior High School	Solok City	Unpopular in Solok City

Results of Limited Experiment

The scoring was performed by using the three categories and the 0-2 polytomous scale. The data were analyzed with QUEST program. The result was that there had been two test items that had not been fit with the model, namely test item number 23 and test item number 24. In both items, not all of the testees were able to attain the category-2 and there were very small number of testees who attained the category-3.

According to CTT, the reliability in the form of Cronbach Alpha, namely 0.65, is still the same after both items were eliminated from the analysis. Meanwhile, according to IRT, the estimated reliability based on the testees' (case/person) analysis in the form of person separation index is 0.82. Table 3 shows the average score for the increasing item difficulty level, starting from the easiest to the hardest one. The gradation for the aspect of fundamental capability is the chronological thinking skills, continuous and changing identifying skills and causal analyzing skills.

Table 2. Results of Item Estimation (I) and Testee Estimation (N) from the Limited Experiment

No.	Explanations	Before the Two Items were Eliminated (I=111)		After the Two Items were Eliminated (I=109)	
		Estimation for Item	Estimation for Testees	Estimation for Item	Estimation for Testees
1	Average and standard deviation scores	0.00 ± 1.08	-0.61 ± 0.86	0.00 ± 1.06	-0.58 ± 0.85
2	Average and standard deviation scores that had been made appropriate	0.00 ± 1.02	-0.61 ± 0.78	0.00 ± 1.00	-0.58 ± 0.77
3	Separation index	0.89	0.82	0.89	0.82
4	Cronbach Alpha scores		0.54		0.54
5	Average and standard deviation scores of INFIT MNSQ	0.98 ± 0.10	0.99 ± 0.47	0.98 ± 0.10	0.99 ± 0.48
6	Average and standard deviation scores of OUTIFTMNSQ	0.99 ± 0.15	1.00 ± 0.51	0.98 ± 0.13	1.00 ± 0.51
7	Average and standard deviation scores of INFIT t	-0.22 ± 1.06	-0.24 ± 1.09	-0.19 ±1.06	-0.24 ± 1.09
8	Average and standard deviation scores of OUTFIT t	-0.17 ± 1.07	-0.15 ± 1.05	-0.14±1.06	-0.14 ± 1.05
9	Item or testees of 0 score	0	0	0	0
10	Item or testees of perfect score	0	0	0	0

The aspects of historical thinking skills are, respectively, historical significant meaning establishing skills, historical source/information and data recording skills, historical research planning skills, historical results of research reporting skills and historical sources analyzing and benefitting skills. The average scores for the level of item

difficulty within the sub-aspect of historical sources analyzing and benefitting skills are the highest ones among the other historical thinking aspects; meanwhile, the average scores for the level of item difficulty within the sub-aspect of historical significant meaning establishing skills are the lowest ones. The item distribution, based on the level of difficulty in the form of difficulty value as the results of analysis by using the QUEST program, shows that 5.40% of the items of the basic skills are quite difficult (from 1.0 to <1.5) and that there has not been any item of basic skills that are difficult (from 1.5 to 2.0). From the items of historical research planning skills, the researchers found that there had been 5.40% of the items that were quite difficult (from 1.0 to <1.5) and that were difficult (from 1.5 to 2.0). The researchers also found that there had been 1.35% of the items that were very difficult (≥2.0).

Table 3. The Scores of Difficulty Level in the Aspects and the Sub-aspects of Historical Thinking Skills according to PCM based on the Results of Limited Experiment

No.	Aspects and Sub-Aspects of Historical Thinking Skills	Level of Item Difficulty Score		
		Difficulty	Delta	
1.	**Basic Skills**	**-0.989**	**-2.677**	**0.697**
a.	Chronological thinking skills	-1.776	-3.336	-0.221
b.	Continuity and change identifying skills	-1.027	-2.673	0.618
c.	Causal relationship analyzing skills	-0.348	-2.190	1.492
2.	**Historical research capabilities**	**0.508**	**-0.685**	**1.703**
a.	Significant meaning establishing skills	-0.450	-1.993	1.093
b.	Historical data/information/source recording skills	0.462	-0.862	1.788
c.	Historical sources benefitting and anakyzing skills	0.917	-0.405	2.238
d.	Historical research planning skills	0.689	-0.305	1.690
e.	Historical research results reporting skills	0.726	0.112	1.340

Table 4. The Item Distribution in the Aspects of Historical Thinking Skills based on the Scores of Difficulty Level in the Limited Experiment

Range on the Level of Difficulty	Basic Skills		Historical Research Capabilities	
	Absolute Frequency	Relative Frequency	Absolute Frequency	Relative Frequency
< -2.0	4	10.81%	0	0.00
-2.0 to <-1.5	5	13.51%	0	0.00
-1.5 to <-1.0	6	16.21 %	4	5.40%
-1.0 to <-0.5	11	29.72%	3	4.05 %
-0.5 to <0.0	5	13.51 %	9	12.16 %
0.0 to <0.5	3	8.10 %	16	21.62 %
0.5 to <1.0	2	5.40 %	23	31.08 %
1.0 to <1.5	1	2.70 %	16	21.62%
1.5 to <2.0	0	0.00%	2	2.70%
≥ 2.0	0	0.00%	1	1.35 %
Total	**37**	**100 %**	**74**	**100 %**

Results of Expanded Experiment

The summary was compiled by using the QUEST program and the results of the summary are presented in Table 5. Table 5 shows that overall, the items in the form of the test had been fit with the model which had been a prerequisite for the QUEST program.

Table 5. Results of Item Estimation (I) for the Historical Thinking Skills and of Testee Estimation (N) according to the Partial Credit Model (PCM) in the Expanded Experiment.

No.	Explanations	Estimation for Item	Estimation for Testees (Case/Person)
1	Average and standard deviation scores	0.00 ± 0.96	-0.58 ± 0.71
2	Average and standard deviation scores that had been made appropriate	0.00 ± 0.93	-0.58 ± 0.60
3	Separation index	0.93	0.72
4	Cronbach Alpha scores		0.41
5	Average and standard deviation scores of INFIT MNSQ	0.99 ± 0.05	0.99 ± 0.51
6	Average and standard deviation scores of OUTIFTMNSQ	0.99 ± 0.10	0.99 ± 0.56
7	Average and standard deviation scores of INFIT t	-0.16 ± 1.05	-0.25 ± 1.08
8	Average and standard deviation scores of OUTFIT t	-0.14 ± 1.04	-0.16 ± 1.05
9	Item or testees of 0 score	0	0
10	Item or testees of perfect score	0	0

According to CTT, the Cronbach Alpha index is 0.54. On the other hand, according to IRT (Wright & Masters, 1982, p. 106; Keeves & Masters, 1999, p. 276) the reliability that has been estimated based on the testee (case/person) analysis in the form of person separation index is 0.72.

Table 6. The Scores on the Level of Item Difficulty in the Aspects and the Sub-aspects of Historical Thinking Skills within the Expanded Experiment

No.	Aspects and Sub-Aspects of Historical Thinking Skills	Level of Item Difficulty Score		
		Difficulty	Delta	
			1	2
1.	**Basic Skills**	**-0.705**	**-2.307**	**0.897**
a.	Chronological thinking skills	-1.072	-2.641	0.488
b.	Continuity and change identifying skills	-0.698	-2.150	0.758
c.	Causal relationship analyzing skills	-0.420	-2.261	1.419
2.	**Historical research capabilities**	**0.369**	**-0.650**	**1.390**
a.	Significant meaning establishing skills	-0.13	-1.363	1.102
b.	Historical data/information/source recording skills	0.24	-1,00	1.481
c.	Historical sources benefitting and anakyzing skills	0.461	-0.552	1.475
d.	Historical research planning skills	0.643	0.135	1.153
e.	Historical research results reporting skills	0.933	0.178	1.691

Based on Table 6, most item analysis results in the expanded experiment are similar to those of the limited experiment. The average scores for the level of item difficulty from the basic skills to the historical research planning skills show an increasing gradation from the easiest ones to the hardest ones. The finding is similar to that of the limited experiment.

Results of Measurement for the Expanded Experiment

The results of measurement show that the range of raw scores is 2 as the lowest score and 39 as the highest one and the limit of maximum score is 50 (category-1 = 0, category-2 = 1 and category-3 = 2).

Table 7. The Absolute Frequency and the Converted Relative Scores of the Historical Thinking Skills in the Range between -2.00 and 2.00 with the Class Interval 0.5.

No.	Class of Interval for Converted Scores	Absolute Frequency	Relative Frequency	Cummulative Frequency
1	Score 0 (uncallibrated)	0	0.00	0.00
2	<-2.00	46	1.72	1.72
3	-2.00 s/d -1.50	244	9.12	10.84
4	>-1.50 s/d -1.00	321	12.00	22.84
5	>-1.00 s/d -0.50	738	27.60	50.44
6	>-0.50 s/d 0.00	1166	43.62	94.06
7	>0.00 s/d 0.50	122	4.56	98.62
8	>0.50 s/d 1.00	28	1.04	99.66
9	>1.00	8	0.29	100.00
	Total	**2673**	**100.00**	

After having been calibrated, the lowest converted score was -3.52 and the highest converted score was 0.09 from the range between -4.00 and +4.00. The calibrated scores were then grouped with the class interval 0.5. The results of the calibration show that there are 5.89% of the testees who earned the converted scores bigger than 0.00. Thereby, if the limit 0.00 was positioned as the mid-score, then 94.11% of the testees would be under the mid-score. As a result, most of the testees did not manage to earn 50% of the correct answers.

Discussions

Item Characteristics in the Activities of Limited Experiment

The results of analysis on the data of limited experiment, based on the Partial Credit Model, show that there are items that had delta-1 scores bigger than those of delta-2; however, in overall, the items had been fit with the model. The finding was not in contrary to the supporting theories, as having been proposed by Wright & Masters (1982, pp. 44-45) that according to PCM the analysis characteristics enabled the items that had the scores of delta-1 bigger than those of delta-2. The statement implied that the ability to improve from category-2 to category-3 might be lower than that of category-1 to category-2. The results of the analysis also showed that among 111 items that had been tested, there were 2 items that had not been fit to the Partial Credit Model (PCM), namely item number 23 and item number 24.

Level of Test Item Difficulty

The sub-aspects of basic skills and the questions of causal analyzing skills were the most difficult skills. Then, both of the skills were accompanied by the following skills: (a) change and continuity identifying skills; and (b) chronological thinking skills.

The causal analyzing skills were the skills that demanded in-depth comprehension from the learning participants. Analyzing the causal relationship should be followed by the learning participants' capabilities in the form of systematic historical data presentation so that the causal relationship in certain historical events would be easily comprehended. In this case, the students should not only memorize the facts presented in the textbooks and the lectures by teachers but also should present the causal relationship in certain historical events from many sources. In the same time, the students were also demanded to classify the historical presentations from the historical sources that they had. In other words, the students did not only summarize the results of their observation but also presented the results of their observation into multiple forms of data presentation such as tables, flowcharts, historical maps and alike.

In terms of historical research planning skills, the sub-aspects that had the highest level of difficulty was the historical research capabilities. The finding was common due to the lack of research reporting implementation in the schools. The level of difficulty in the

aspect was followed respectively by the following skills: historical research planning skills, historical sources benefitting/analyzing skills, historical sources/information/data recording skills and historical significant meaning establishing skills. The learning participants had difficulties when they had to think about alternative actions if the activities had been rarely conducted.

The Characteristics of Test Items in the Expanded Experiment

All of the items implemented in the expanded experiment had been fit with the model. The average scores for the level of item difficulty in the limited experiment, for the aspects of basic skills and of advanced skills, respectively, were -0.989 and 0.508. In the expanded experiment, the rank of the average scores for the level of difficulty, respectively, were -0.705 and 0.369. The data showed a similar pattern of responses between the results of limited experiment and those of expanded experiment and, based on the level of difficulty, still there had been a similar pattern of responses between the results of both experiments.

The average scores for the sub-aspect difficulty level from the basic skills aspect in the activities of limited experiment, starting from the most difficult, were -0.348 (sub-aspect c: analyzing causal relationship), -1.027 (sub-aspect b: identifying the continuity and change) and -1.776 (sub-aspect a: thinking chronologically). Then, the average scores for the sub-aspects difficulty level from the basic skills aspect in the measurement stage, starting from the most difficult one, were -0.420 (sub-aspect c: analyzing the causal relationship), -0.698 (sub-aspect b: identifying the change and the continuity) and -1.072 (sub-aspect a: thinking chronologically). Thereby, there have not been any difference in the pattern of testees' responses. Similarly, the easiest response was still the same, that is, thinking chronologically.

Results of Test in the Expanded Experiment

The results of the test in the expanded experiment show that the scores of historical thinking skills which were attained from 2,673 testees were unsatisfiying; there are only 5.89% of the testees who earned the scores above the mid-point. There were three factors that might cause the finding.

The first factor is that the historical thinking skills were not taught completely and integratedly in each subject topic. As a result, the opportunities of exercising the historical thinking skills became very small. The second factor is that the historical thinking skills in the subject topics of historical learning were not implemented especially in the strategies of applying the historical thinking skills for finding concepts instead of applying the historical thinking skills for clarifying the facts as a result of memorization. The historical learning that relied on the memorization of facts and concepts made the students unable to perform historical thinking appropriately. The third factors is that the historical thinking skills might have been taught in accordance with the demand of internal competence and standard competence as formulated in Curriculum 2013; however, the learning participants had not been habituated to work on the non-objective tests that enabled them to provide as many correct answers as possible.

Conclusions and Suggestions

Conclusions

Based on the results of the study and the discussions, the researchers draw several conclusions as follows. First, the assessment model that had been developed belongs to the procedural one. Second, the information attained from the assessment model of historical thinking skills was the formulation of learning continuum for the historical thinking skills, the item characteristics in the form of item difficulty and the testees' capability (*theta-θ*) and the test items that had empirical evidence that had been fit to the Partial Credit Model (PCM) based on the three category polytomous data. Third, the validity of test instrument for the historical thinking skills that had been designed had been met through the expert judgement and had been proven fit empirically to the Partial Credit Model (PCM) based on the three category of polytomous data.

Fourth, the reliability of test instrument for the historical thinking skills in the form of Cronbach Alpha index had been quite good, namely 0.64. Fifth, the overall results of assessment showed that the testees had not mastered the historical thinking skills that had been tested. The finding was apparent from the fact that only 5.89% of the testees who had been in the expected mid-scores based on the three-category polytomous data according to the Partial Credit Model (PCM). The reason was that the learning participants were lack of exercising the historical thinking skills in finding concepts and of working on the non-objective tests.

Suggestions

Based on the conclusions, the researchers formulate several suggestions as follows. First, the study only involved the state senior high schools as the samples; therefore, the researchers suggest that the future studies might involve larger sample size so that wider mastery of historical thinking skills in the related educational degree might be found. Future studies might also be developed in elementary schools or *madrasah ibtidaiyah*, senior high schools or *madrasah tsanawiyah* and even in universities.

Second, there should be further studies to find out the mastery of historical thinking skills as an inter-site comparison or an inter-year comparison with representative sample size. Further studies might also be conducted in order to find out the relationship between the historical thinking skills and the teaching strategy in the historical learning process.

Third, the researchers suggest teachers to train their students through appropriate learning process to develop their historical thinking skills. Fourth, it is suggested for the teachers to train historical thinking skills integratedly in every single learning activity in accordance to the characteristics of the subject topics. Thus, learning participants would habituate themselves to find facts, concepts and theories by utilizing historical thinking skills as having been performed by the historians specifically and social science experts in general.

Fifth, historical thinking skills in senior high schools should be measured periodically in order to find out the students' mastery level of historical thinking skills in the related year. Sixth, the teachers should utilize the mechanisms of assessment for learning using the results of measuring the historical thinking skills applied in the related senior high schools so that the results might be used for improving the quality of lesson plan design and even for providing remedy tests for the learning participants.

Seventh, there should be appreciations and also conducive atmospheres from the related parties in order to encourage the teachers to perform tests by employing open essay to stimulate the learning participants' development of historical thinking skills. Eighth, teachers should make the learning participants aware of the importance in identifying multiple test forms in order that they would have wider insights and comprehend the problems contained in the status of the test item kinds.

References

Allen, M. J. & Yen, W. M. (1979). *Introduction to measurement theory*. Belmont, CA: Wadsworth, Inc.

Ashby, R., Lee, P. J. & Shemit, D. (2005). Putting principles into practice: teaching and planning. In M.S. Donovan & J.D. Bransford (Eds.). *How students learn: History, mathematics, and science in the classroom*. Washington, DC: The National Academies Press.

Bain, R. B. (2005). Applying the principles of how people learning teaching high school history. In M.S. Donovan & J.D. Bransford (Eds.). *How students learn: History, mathematics, and science in the classroom*. Washington, DC: The Natio-nal Academies Press.

Barton, K. C. & Levstik, L. S. (2003). Why don't more history teachers engage students in interpretation?**.** *Research and Practice Social Education*, 67 (6), pp. 358-361.

Borg, W. R. & Gall, M. D. (1989). *Educational research: An introduction* (5th ed.). New York, NY: Longman.

Departemen Pendidikan Nasional (Depdiknas). (2007). *Peraturan Menteri Pendidikan Nasional Republik Indonesia Nomor 20, Tahun 2007, tentang Standar Penilaian Pendidikan untuk Satuan Pendidikan Dasar dan Menegah* [Indonesian National Education Minister's regulation number 20, in the year of 2007, about the standard of educational assessment for primary and secondary education].

Fogu, C. (2009). Digitalizing historical consciousness. *Journal History and Theory, 47* (1), pp. 103-121.

Griffin, P. & Nix, P. (1991). *Educational assessment and reporting: A new ap-proach*. Sydney: Harcourt Brace Jovanovich, Publishers.

Hambleton, R. K. & Swaminathan, H. (1985). *Item respons theory*. Boston, MA: Kluwer Inc.

Hargreaves, A., Earl, L. & Schmidt, M. (2002). Perspectives on alternative assesment reform. *American Educational Research Journal*, *39* (1), pp. 69-95.

Keeves, J. P. & Master, G. N. (1999). Introduction. In G. N. Masters & J. P. Keeves (Eds.). *Advances in measurement in education research and assess-ment*. Amsterdam: Pergamon, An imprint of Elsevier Science.

Lee, P. (2005). Putting principles into practice: understanding history. In M. S. Donovan & J. D. Bransford (Eds.). *How students learn: History, mathematics, and science in the classroom*. Washington, DC: The National Academies Press.

Mardapi, D. (1999). Estimasi kesalahan pengukuran dalam bidang pendidikan dan implikasinya pada ujian nasional [The estimation of miss-assessment in educational field and its implication to national examination]. *Proceeded in the inaugural speech of Professor on 4 May 1999*. Yogyakarta: Yogyakarta State University.

Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan nontes* [Technique of test non-test instrument arrangement]. Yogyakarta: Mitra Cendikia Press.

Masters, G. N. (1999). Partial credit model. In J. P. Keeves & G. N. Masters (Eds.). *Advances in measurement in educational research and assessment*. Amsterdam: Pergamon.

Oriondo, L. L. & Dallo-Antonio (1998). *Evaluating educational outcomes (test, measurement, and evaluation)* (5th ed.). Quezon City: REX Printing Company.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *The Danish Yearbook of Philosophy*, 4 (1), pp. 321-334.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *The Danish Yearbook of Philosophy*, *14* (3), pp. 58-93.

Seixas, P. & Peck, C. (2004). Teaching historical thinking. In A. Sears & I. Wright (Eds.), *Challenges and prospects for Canadian social studies*. Vancouver: Pacific Educational Press.

Seixas, P. (2013). *Linking historical thinking concepts, content and competencies*. Vancouver: Pacific Educational Press.

Van der Linden, W. J. & Hambelton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.

Winerburg, S. (2006). *Berpikir historis: Memetakan masa depan, mengajarkan masa lalu.* (M. Maris, Trans.). Jakarta: Yayasan Obor Indonesia.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.