Research Article https://doi.org/10.12973/eu-jer.11.2.609

European Journal of Educational Research
Volume 11, Issue 2, 609 - 619.
ISSN: 2165-8714
http://www.eu-jer.com/

The Development of Historical Thinking Assessment to Examine Students' Skills in Analyzing the Causality of Historical Events

Ofianto* Universitas Negeri Padang, INDONESIA
Aman Universitas Negeri Yogyakarta, INDONESIA
Tri Zahra Ningsih Universitas Sebelas Maret, INDONESIA
Nur Fatah Abidin Universitas Sebelas Maret, INDONESIA

Received: September 2, 2021 ▪ Revised: November 8, 2021 ▪ Accepted: December 30, 2021

Abstract: This research aimed to develop a historical thinking assessment for students' skills in analyzing the causality of historical events. The development process of Gall and colleagues and Rasch analysis models were used to develop an assessment instrument consisting of two processes, including the analysis of the framework of cause and consequence, the validity, reliability, and difficultness test. This research involved 150 senior high school students, with data collected using the validation sheet, tests, and scoring rubric. The results were in the form of an essay test consisting of six indicators of analyzing cause and consequence. The instruments were valid, reliable, and suitable for assessing students' skills in analyzing the causality of historical events. The developed instruments were paired with a historical thinking skills assessment to improve the accuracy of the information about students' level of historical thinking skills in the learning history.

Keywords: *Causality, historical events, historical thinking skills.*

To cite this article: Ofianto, Aman, Ningsih, T. Z., & Abidin, N. F. (2022). The development of historical thinking assessment to examine students' skills in analyzing the causality of historical events. *European Journal of Educational Research*, *11*(2), 609-619. https://doi.org/10.12973/eu-jer.11.2.609

Introduction

Historical thinking is important in the theory and practices of history education. This thinking is a higher-order skill that must be possessed by history students (Barton, 2011). Historical thinking skills involve more than transferring knowledge from teacher to student, resulting in inactive students in the class (Kesuma, 2020). Students' ability to use procedural knowledge to investigate and interpret historical events is known as historical thinking (Duquette, 2015; Laksana, 2020; Seixas & Peck, 2004). Therefore, students should think and act as historians (Carroll, 2019; Gestsdóttir et al., 2018; Seixas, 2017). Historical thinking skills are grouped into six aspects, namely establishing historical significance, using primary source evidence, identifying continuity and change, analyzing cause and consequence, taking perspectives, and understanding the moral aspect of historical interpretations (Seixas, 2006).

Students' skills in analyzing the causality of historical events are an important aspect. Historical explanation demands an advanced skill to use cause and consequence logic thinking (Seixas, 2017). This skill requires a higher level of thinking to memorize the historical facts and understand the causality of historical events through primary source analysis (Ofianto & Suhartono, 2016). The ability to understand the causality of historical events is important to students (Seixas, 2006). They can identify interaction and constraint on intentional human action that triggers change and continuity. Additionally, students can identify multiple causes and understand the counterfactuals of historical events.

Teachers should be able to assess students' skills level in analyzing historical events causality. Student's ability to understand the causality of historical events is unnoticed by teachers due to a lack of necessary instruments. Seixas et al. (2015) developed a historical thinking assessment model called the One-Hour Test consisting of short answers and multiple-choice items. However, this instrument focuses on assessing two aspects of historical thinking skills, namely the use of primary source evidence and understanding the moral aspect of historical interpretation. Smith (2017) adopted Seixas' (2006) model of historical thinking by developing multiple-choice instruments. Moreover, recent research of Ningsih et al. (2019) developed a portfolio instrument consisting of an essay test and interview sheet to assess students' skills in using primary source evidence.

* **Corresponding author:**
Ofianto, Universitas Negeri Padang, Kota Padang, Sumatera Barat 25173, Indonesia. ✉ ofianto@fis.unp.ac.id

Developments of historical thinking skills instruments, including multiple-choice and portfolio, have several weaknesses. The multiple-choice items cannot assess students' cognitive processes outside their memory or knowledge (Cantor et al., 2015). The multiple-choice weakness is seen in its inability to diagnose students' partial or impartial answers (Zhai & Li, 2021). Moreover, it can only drive students to choose conjecture answers (Zhai & Li, 2021), making them unable to give detailed responses to the questions (Slepkov & Godfrey, 2019). In contrast, portfolio instruments research takes more time to assess students' skills, particularly in the large classroom (Soifah & Pratolo, 2020). Its data cannot be analyzed (Fuller, 2017) and have a biased score because it depicts the best students' scores (Lombardi, 2008).

The results show that students' skills in analyzing causality of historical involve two practical problems including: (i) The existing instrument can only assess some historical thinking skills such as student's skill in using primary source evidence and understanding the moral aspect of historical interpretations. This means that other historical instruments' thinking aspects, specifically students' ability to analyze the causality of historical events, are not yet developed. (ii) The existing instruments consisting of multiple-choice and portfolio have several weaknesses to be evolved to improve the assessment of historical thinking skills in learning history.

This research used the Rasch Model to develop a specific instrument for assessing students' skills in analyzing the causality of historical events. Teachers assessed students' skills in analyzing the causality using the developed instrument. This model provides a framework explaining humans' knowledge comparing data. Formulations and models are more detailed and suitable to be described by analogy (Ramadhan et al., 2019). The Rasch Model fulfills the five principles of the measurement model thus used to develop the instrument (Rasch, 1977; Wright & Masters, 1982). The developed instrument consists of essay questions constructed by polytomous or more than two alternative responses and its scoring rubric to assess students' skills. This model gives accurate information about the level of students' historical thinking.

Literature Review

Cause and Consequence Analysis Skills

Students with higher-order thinking skills in the 21st century should have analytical skills, which are part of historical thinking skills (Amzaleg & Masry-Herzallah, 2021; Anderson & Krathwohl, 2001; Araya, 2020; Seixas, 2017). Analytical skills help students to build historical interpretations through reading evidence or historical sources, distinguishing patterns, linking events, identifying causal relationships, and concluding (Gibson & Peck, 2020; Keleşzade et al., 2018; Monte-Sano, 2010; Schoemaker, 2020; Seixas, 2017). Analytical skills require one to understand the causes and consequences of a historical event. Therefore, the cause and consequence are how an event occurred and its outcome. The causes represent events, circumstances, actions, or beliefs related directly or indirectly to the event (Woodcock, 2011). The cause and consequence emerge from the continuity and change in history. It is characterized by network causes and consequences that interact to produce far-reaching results (Boadu & Donnelly, 2020; Seixas & Morton, 2013).

Moreover, cause and consequence result from the interplay between humans and their conditions. The cause and consequence explanation is crucial in history because it is limited to telling the past and understanding current events (Boadu & Donnelly, 2020). Students with good cause and consequence analysis skills comprehensively understand historical events (Alcoe, 2015). The results indicate that cause and consequence analysis skills are the primary capital for students to build historical interpretations through reading several historical sources, connecting events, and analyzing the causes and consequences of events to conclude. Analytical skills can improve higher-order thinking skills including reasoning abilities, causal understanding, and historical thinking through reading historical sources, distinguishing patterns, connecting events, identifying cause-and-consequence relationships, and drawing conclusions.

Essay Test

Student classroom achievement and learning effectiveness are measured using teacher assessment (Black & Wiliam, 2018; Maba, 2017; Mehany & Gebken, 2021; Suhaini et al., 2021). Teachers use tests to obtain students learning outcomes and assess them for completing specific tasks through a systematic procedure. This is conducted to know about students' particular mastery of skills or knowledge (Adom et al., 2020; Osadebe & Nwabeze, 2018). A good test should have high validity and reliability, be objective, non-discriminatory, comprehensive, easy to use, and the results can be justified (Hughes et al., 2000). The tests were classified into objectives and essay (Adom et al., 2020). The objective test focused on the correct answer from several choices. In contrast, essay tests challenged students to respond to questions presented as sentences or narratives series.

Students are free to respond to questions by providing ideas and information related to the questions asked during the essay test (Ogunka & Iweka, 2021). Research indicates that essay tests are suitable for measuring higher-order thinking skills such as analytical, critical, and historical thinking skills (Alifah et al., 2020; Gómez et al., 2020). During essay tests, students should present answers in the form of descriptions or narratives. This allows them to integrate information, provide arguments, analyze data, reason, and conclude (Kaipa, 2020). Essay tests provide opportunities for students to improve higher-order thinking skills compared to multiple-choice tests. Measuring students' skills in explaining an event, presented systematically and chronologically based on the analysis results of historical evidence during history learning,

was conducted using an essay test (Darling-Hammond, 2017). The results indicate that the essay test is suitable for measuring higher-order thinking skills such as explaining, analyzing, critical and historical thinking skills because students are challenged to respond ideas presented in sentences or narratives series.

Methodology

Research Design

This research was conducted using Gall et al. (2003) developmental research framework. The instrument development procedure is divided into two stages including: (i) needs analysis of instrument development of analyzing cause and consequence skills through library research, and (ii) the development of the instrument. A draft of the analysis cause and consequence skill assessment instrument was created in the first and development stages, including expert validation, instrument testing, validity, reliability testing, and test difficulty level testing. Exam's validation was conducted by experts, including two materials, evaluation and language experts. This research engaged experts with historical science backgrounds, education, languages, and field of evaluation. Historical study and education experts evaluated the assessment instrument's content and build. Language experts assessed aspects of the designed instrument's language, while evaluation experts assessed the instrument's form analyzing cause and effect skills. This involved 150 students chosen using a proportionate sample approach. The Quest data processing program version 2.0 was used to examine the validity, reliability, and item difficulty level of the data. This determined the instrument's feasibility and validity in assessing the analyzing cause and consequence skill. The triangulation of methods analysis or mixed-method analysis verified the result of the development.

Research Participant

The research participants (498 students) came from 5 senior high schools, as shown in Table 1:

Table 1. The Total Participant of the Research

No	Schools List	The Number of Students
1	School A	98
2	School B	98
3	School C	102
4	School D	97
5	School E	103
	Total Students	498

Students' scores from each school in the final exam showed in the homogeneity of variance test.

Table 2. The Result of the Homogeneity of Variance Test

Test of Homogeneity of Variance					
		Levene's Statistic	df1	df2	Sig.
The result of students' final exam	Based on Mean	1.969	4	493	.098
	Based on Median	2.031	4	493	.089
	Based on Median and with adjusted df	2.031	4	478.961	.089
	Based on trimmed mean	2.005	4	493	.093

Table 2 showed the value of homogeneity mean was 0.098 larger than 0.05, making the data homogenous.

The 30 participants were chosen based on the proportional sampling technique. Therefore, this research involved 150 students and 6 experts in the developmental process.

Research Instrument, Data Collection, and Data Analysis

The data were collected through a validation form, test, and scoring rubric. The developed instrument was validated by 6 experts consisting of 2 assessment, content, and language experts. This research involved experts from historical science backgrounds, historical education, languages, and the evaluation field. Historical study and education experts were tasked with evaluating the assessment instrument's content and build. Language experts assessed aspects of the designed instrument's language, while evaluation experts assessed the instrument's form to analyze cause and effect skills.

The feasibility of the developed instrument was measured using the criteria used by Sutimin et al. (2018), as shown in Table 3.

Table 3. The Feasibility Criteria of the Developed Instrument

Range	Criteria
3.26-4.00	Very Suitable
2.51-3.25	Suitable
1.76-2.50	Less Suitable
1.00-1.75	Not Suitable

This research used the triangulation of data analysis of the quantitative and qualitative approaches. The qualitative analysis was used to strengthen the developed instrument, and an open questionnaire was used to collect responses from experts' judgment towards the developed instrument quality. These responses were analyzed to revise and improve the instrument quality in the developmental process. Contrastingly, the quantitative analysis aimed to measure the validity and reliability of the developed instrument using the Rasch Model through the Quest Program. The difficulty level of the developed instrument was obtained. The Rasch I-PL model with a polyatomic scale was used to assess the test item. This research used a 0-2 or 3-category scale using the following categorization:

Category 1: If students do not provide a correct answer

Category 2: If students can get 1 correct answer from the specified criteria,

Category 3: If student can get 2 or more correct answers from the specified criteria.

This research used the Rasch Model because of its ability to analyze the validity and reliability of the developed instrument (Alhadabi & Aldhafri, 2021). Moreover, the Rasch Model also fulfills the five principles model of assessment (Rasch, 1977; Wright & Masters, 1982) that covers the instrument ability to give linear scale with constant interval, predict the missing data, give an accurate estimation, detect the inaccuracy of the model, and its replicability of the measurement.

The Rasch Model was used to analyze the fit items based on INFIT Means of Square (INFIT MNSQ) and the standard deviation or the average of INFIT Means of INFIT t. The items are valid if the values of INFIT MNSQ are between 0.77 to 1.30 (Adam & Khoo, 1996). The reliability of the developed instrument was analyzed based on the value of Internal Consistency. The developed instrument difficulty range was between ±2 or -2 to +2 where -2 = very easy and +2= very difficult (Hambleton & Swaminathan, 2013).

Results

The Framework of the Analyzing Cause and Consequence Skill

The theoretical framework of the analyzing cause and consequence skill follows Seixas' (2006) historical thinking model. Seixas (2006) stated that analyzing cause and consequence is important in triggering the changes in history. The causality of historical events is differently understood by historians based on ideology, perspective, and approaches. Humans can cause historical changes interconnected with other aspects of human life. Seixas explained that students could achieve four skills from understanding the frameworks, including identifying: (i) human intended action and interaction, (ii) many historical events factors, (iii) the ripple effects of historical factors, and (iv) be able to explain the cause and consequence from historical events. An instrument consisting of students' four skills to analyze the cause and consequence of historical events was designed based on the results.

Table 4. The Draft of Analyzing Cause and Consequence Skill Instrument

The framework of analyzing cause and consequence	Indicators of analyzing cause and consequence	The form of the test
Analyzing the cause of a historical event	a. Students can identify the cause of a historical event	Essay
	b. Students can give examples from various fields that cause a historical event in people's lives.	Essay
	c. Students can recognize and compare several causes in their surrounding environment.	Essay
Analyzing the consequences of a historical event	a. Students can identify the consequences of a historical event	Essay
	b. Students can give examples of the consequences of a historical event in various people's lives.	Essay
	c. Students can recognize several influences/consequences, short, medium, and long term, of the events in their surrounding environment.	Essay

The Result of Developmental Processes

The Expert Validation

The experts tested the feasibility of the developed instrument using 2 materials, 2 assessments, and linguists' experts, as shown in Table 5.

Table 5. Feasibility Test Results by Experts

Indicator	Score	Interpretation
Material	3.5	Very Suitable
Assessment Instrument	3.75	Very Suitable
Language	3.75	Very Suitable
Mean	3.67	Very Suitable

Based on the expert validity, the developed instrument was suitable for assessing students' skills in analyzing historical events causality. Generally, the developed instrument represented the theory and conceptual framework of historical thinking in historical events causality. However, language revision was necessary for the developed instrument with the level of students' comprehension of the instrument's academic and scientific terminologies. Revision of language improved the quality of the developed instrument.

The Validity, Reliability, and Difficulty of the Instrument

The analysis of item validity and reliability was conducted through the Rasch Model analysis using the Quest Program. The validity indicated that the items of the question in the developed instrument were valid. The item results validity can be seen in figure 1.

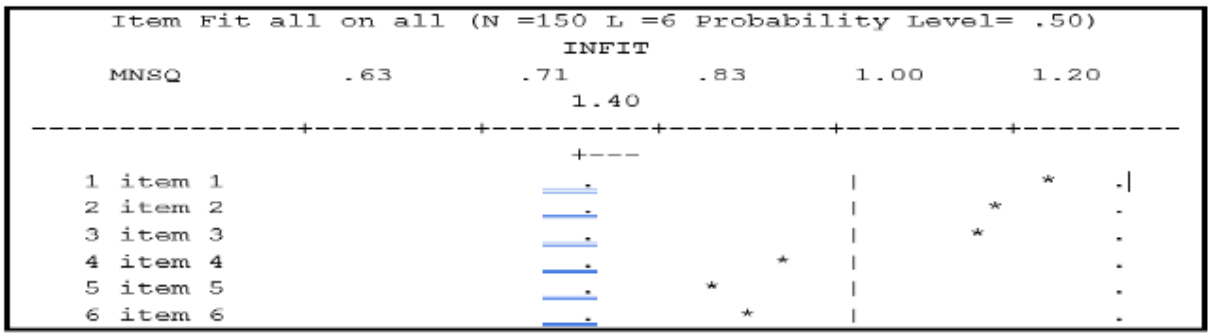

Figure 1. Item Validity of the Developed Instrument

Figure 1 shows that items validity was proved by the item fit value located between 0.77 to 1.30. Table 6 presents detailed information on the item validity.

Table 6. The Validity and Level of Difficulty of the Developed Items

No	The items of Analyzing Cause and Consequence Skills	Validity Value (INFIT Mean of Square)	Difficulty value	Delta	
				1	2
1.	Many things can cause a historical event, for example, the influence of the ideas of historical figures. What other identification can be the cause of a historical event? Write at least two answers!	1.21	-1.23	-2.82	0.36
2.	A historical event can occur because it is caused by several factors such as social and cultural, economic, political, education, natural environmental, and religious factors. Choose at least two factors that cause a historical event and give an example of an event that occurred?	1.08	-0.16	-2.27	1.95
3.	If you are asked to observe the cause of an event around you, explain the relationship between each cause you notice and occurrences that occur in your daily life? Write at least two answers!	1.07	-0.62	-2.16	0.93

Table 6. Continued

No	The items of Analyzing Cause and Consequence Skills	Validity Value (INFIT Mean of Square)	Difficulty value	Delta	
				1	2
4.	One of the consequences that a historical event can cause is progress or regress in various people's lives. Identify other things that might occur as a result of a historical event! Write at least two answers!	0.95	-0.06	-1.96	1.83
5.	A historical event can have an impact on various fields in people's lives. The effect/impact can be seen, for example, in the social and cultural, economic, political, education, and religious fields. Write at least two examples of the impact caused by various fields of people's lives due to a historical event!	0.88	-1.18	-2.44	0.09
6.	If you are asked to observe the effect of an event around you, explain how the influence you saw has affected the lives of others around you in the short, medium, and long term? Write at least two answers!	0.91	0.08	0.11	0.24

The difficulty level analysis indicated that the developed model items were between level -2 to +2. Therefore, no items of questions in the developed instrument were dropped.

The developed instrument reliability analysis showed that the instrument items were reliable using an internal consistency value (0.72). Table 7 presents detailed information for item reliability.

Table 7. The Reliability of the Developed Instrument

Item Analysis Results or Observed Responses	
Mean test score	9.99
Standard deviation	3.56
Internal Consistency	.72

All on all (N = 150 L = 6 Probability Level= .50)

The results showed that the developed instrument consisted of valid and reliable items. Therefore, the developed instrument of the analyzing cause and consequence skills is used to assess students' skills in analyzing the causality of historical events.

Discussion

The results indicated that the developed analyzing cause and consequence skill instrument was used to assess students' skills in analyzing the causality of historical events. The developed instrument diminished the weaknesses of One Hour and multiple-choice tests. This instrument provides linear scales with the same intervals, predict missing data, provide more precise estimates, and detect inaccurate models and replicable assessment (Rasch, 1977; Wright & Masters, 1982).

The developed assessment instrument such those stated by Smith (2017), Ningsih et al. (2019), and Seixas et al. (2015) offered an instrument for testing historical thinking skills. However, some developed instruments differ from one another. For example, Smith (2017) developed a historical thinking instrument comprising multiple-choice questions. This differs from the historical thinking skills evaluation instrument of this research, which was an essay test. Additionally, Ningsih et al.'s historical thinking skills evaluation tool comprised two instruments, namely essay exams and interview sheets. The only difference between Ningsih et al. (2019) and this research during essay exams is the element of measuring historical thinking. This research focused on assessing features of analyzing the causes of historical events, whereas Ningsih et al. (2019) focused on measuring components of utilizing primary source evidence. The Seixas et al. (2015) designed a One-Hour Test comprising short answers and a multiple-choice test. These exams differed from this research, particularly the essay test. The One-Hour Test focused on employing primary source information and recognizing the moral aspect of historical interpretation measurement. The elements examined are also distinct from the instruments of this research, which focused on analyzing the causes of historical events.

Instruments created to overcome a lack of thinking skills were compiled previously by Seixas et al. (2015), Smith (2017), and (Ningsih et al., 2019). Seixas et al. (2015) and Smith (2017) short answer and multiple-choice exams lacked instruments that evaluated students' skills at a low level. These tests helped students to memorize factual knowledge and not deeper reasoning (Gusev et al., 2017; Leber et al., 2018; Mingo et al., 2018; Ristov et al., 2015). Historical thinking skills helped students to reason profoundly about historical events in establishing a full historical narrative. Multiple-choice exams can diagnose students' knowledge, unlike multiple-choice tests that only ask pupils to select the correct

answer from a set of options. This instrument was designed in the form of an essay test used to overcome the weaknesses of the multiple-choice test (Karaoğlan Yılmaz et al., 2020). This type of test prevents students from acquiring higher levels of understanding (Chan & Kennedy, 2002; Opstad, 2021). Ningsih et al. (2019) portfolio evaluation tool takes longer for teachers to assess student potential making it expensive.

The essay exam is appropriate to address some of the problems and weaknesses in the prior historical thinking instrument. According to Opstad (2021), essay assessments are more effective than multiple-choice tests in fostering pupils' higher-order thinking skills. Essay tests were used in determining a student's potential. Essay assessments result in higher-order thinking abilities such as analytical, evaluative, and reflective thinking. In essay assessment, students use their own words to examine historical problems and demonstrate how they organize related concepts to construct historical narratives through relevant reading. This is different from multiple-choice exams, which only involve a part of students' thinking ability to choose the correct answer from a list.

Moreover, it teaches students to be independent in solving historical problems. According to Meyer (2006), students who were evaluated using the essay exam did better than those who used other types of exams, including short answers and multiple choice. Students preparing for essay test assessments learned more than those who studied for short answer and multiple-choice tests.

Theoretically, the essay test assessed students' higher-order thinking skills because it involved analytical, critical thinking, and communication skills showing a deep learning approach (Indah et al., 2020; Kriswantoro et al., 2021). Students answer questions by presenting ideas and information relevant to the test given in essay assessments (Ogunka & Iweka, 2021). They submitted their answers in descriptions or narratives that allowed them to integrate knowledge, develop arguments, analyze data, reason, and conclude. Additionally, in essay tests, students need to use their higher thinking skills such as analysis, evaluation, and creativity. This test evaluated higher-order thinking skills, including analytical, critical, historical thinking, and reasoning.

Practically, the developed instrument was paired with other historical thinking skills, such as One Hour (Seixas et al., 2015), Multiple-choice (Smith, 2017), or portfolio tests (Ningsih et al., 2019). The developed instrument assessed students' skills in analyzing the causality of historical events. In contrast, other instruments such as One Hour and Portfolio tests assessed other aspects of historical thinking including, students' skill in using primary source evidence and understanding the moral aspect of historical interpretation. Multiple-form tests were used as an alternative scheme of assessing historical thinking skills. For instance, teachers simultaneously used the One Hour Test and The Instrument for analyzing cause and consequence skills to achieve more accurate information regarding students' level in using their historical thinking skills. This meant that teachers had alternative test essay tests besides the multiple-choice or portfolio tests in assessing students' historical thinking skills. The developed instrument supported teachers to assess students' historical thinking skills during learning.

The results showed that the developed instrument of analyzing cause and consequence skill was used with other instruments of historical thinking skills. The developed instrument focused on assessing students' skills in analyzing the causality of historical events. The other instruments focused on assessing other aspects of historical thinking, such as students' skill in using primary source evidence and understanding the moral aspect of historical interpretation. Therefore, developed instruments supported teachers to get accurate information about students' historical thinking skills levels.

Conclusion

The results showed that students' ability to analyze the causality of historical events was divided into 6 aspects, specifically analyzing the cause and consequences of historical events. The developmental process, which consisted of validity, reliability, and difficulty test analysis, showed that the instrument was valid, reliable, and suitable for assessing students' historical thinking skills in analyzing the causality of historical events. History teachers use the developed instrument to assess students' historical thinking skills. It gives more accurate information about students' historical thinking skills than other instruments such as One Hour and a multiple-choice test.

Recommendations

The developed instrument had some constraints despite the many advantages including, time-consuming. The essay test consumed time to assess students' skills in analyzing the causality of historical events. Moreover, essay test struggles to fulfill their objectivity in measuring students' skills (Fuller, 2017). Therefore, a developed instrument for analyzing cause and consequence should be advanced for future research to minimize the weakness of the developed instrument. It is recommended to pair the developed instrument with other forms of historical thinking skills assessment to improve the accuracy of students' historical thinking skills in learning history. For future research, a historical thinking skills instrument should be developed with other forms of assessment focusing on historical thinking skills, such as building historical significance, identifying continuity, and taking perspectives to complement existing instruments for assessing students and obtaining valid results.

Limitation

This research was limited to assessing one of the 6 aspects of historical thinking abilities. Therefore, there were poor representative results to explain historical thinking skills adequately. Furthermore, the trial subjects were limited to 150 students from 5 different schools, and this research cannot be used to represent a wider number of analytical abilities.

Authorship Contribution Statement

Ofianto: conceptualizing and designing, gathering and evaluating data, and preparing manuscripts. Aman: Aman is in charge of providing technical support, supervising, and approving the final product. Ningsih: Data collection and analysis, paper compilation, and data interpretation. Abidin: Data collection and analysis, as well as manuscript revision.

References

Adam, R., & Khoo, S. T. (1996). *Quest: Interactive item analysis program*. The Australian Council for Educational Research.

Adom, D., Mensah, J. A., & Dake, D. A. (2020). Test, measurement, and evaluation: understanding and use of the concepts in education. *International Journal of Evaluation and Research in Education*, *9*(1), 109-119. https://doi.org/10.11591/ijere.v9i1.20457

Alcoe, A. (2015). Post hoc ergo propter hoc? using causation diagrams to empower sixth-form students in their historical thinking about cause and effect. *Teaching History*, (161), 6-24. https://bit.ly/3rqCCzY

Alhadabi, A., & Aldhafri, S. (2021). A Rasch model analysis of the psychometric properties of student-teacher relationship scale among middle school students. *European Journal of Educational Research*, *10*(2), 957-973. https://doi.org/10.12973/eu-jer.10.2.957

Alifah, M., Pargito, P., & Adha, M. M. (2020). The development of test instruments based on HOTS (higher-order thinking skills) using Edmodo. *IOSR Journal of Research & Method in Education*, *10*(6), 42-46. https://bit.ly/3Ef9cZr

Amzaleg, M., & Masry-Herzallah, A. (2021). Cultural dimensions and skills in the 21st century: The Israeli education system as a case study. *Pedagogy, Culture & Society*. Advance online publication. https://doi.org/10.1080/14681366.2021.1873170

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives*. Longman. http://eduq.info/xmlui/handle/11515/18345

Araya, A. M. (2020). Pensamiento crítico para la vida ciudadana en educación primaria: combinando narrativa y herramientas de pensamiento [Critical thinking for civic life in elementary education: combining storytelling and thinking tools]. *Revista Educación*, *44*(2), 23-43. https://doi.org/10.15517/revedu.v44i2.39699

Barton, K. C. (2011). History: from learning narratives to thinking historically. In W. B. Russell III (Ed.), *Contemporary social studies: An essential reader* (pp. 119–138). Information Age Publishing.

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, *25*(6), 551-575. https://doi.org/10.1080/0969594X.2018.1441807

Boadu, G., & Donnelly, D. J. (2020). Toward historical understanding: leveraging cognitive psychology for progression in school history. *The Social Studies*, *111*(2), 61-73. https://doi.org/10.1080/00377996.2019.1659748

Cantor, A. D., Eslick, A. N., Marsh, E. J., Bjork, R. A., & Bjork, E. L. (2015). Multiple-choice tests stabilize access to marginal knowledge. *Memory & Cognition*, *43*(2), 193-205. https://doi.org/10.3758/s13421-014-0462-6

Carroll, J. E. (2019). Epistemic explanations for divergent evolution in discourses regarding students' extended historical writing in England. *Journal of Curriculum Studies*, *51*(1), 100-120. https://doi.org/10.1080/00220272.2018.1499805

Chan, N., & Kennedy, P. E. (2002). Are multiple-choice exams easier for economics students? A comparison of multiple-choice and "equivalent" constructed-response exam questions. *Southern Economic Journal*, *68*(4), 957-971. https://doi.org/10.1002/j.2325-8012.2002.tb00469.x

Darling-Hammond, L. (2017). *Developing and measuring higher-order skills: Models for state performance assessment systems.* Learning Policy Institute. https://files.eric.ed.gov/fulltext/ED606777.pdf

Duquette, C. (2015). Relating historical consciousness to historical thinking through assessment. In K. Ercikan, & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 51-63). Routledge. https://doi.org/10.4324/9781315779539

Fuller, K. (2017). Beyond reflection: Using ePortfolios for formative assessment to improve student engagement in non-majors introductory science. *The American Biology Teacher*, *79*(6), 442-449. https://doi.org/10.1525/abt.2017.79.6.442

Gall, M. D., Gall, J. P., & Borg, W. R. (2003). *Educational research: An introduction* (7th ed.). Pearson Education, Inc.

Gestsdóttir, S. M., van Boxtel, C., & van Drie, J. (2018). Teaching historical thinking and reasoning: Construction of an observation instrument. *British Educational Research Journal*, *44*(6), 960-981. https://doi.org/10.1002/berj.3471

Gibson, L., & Peck, C. L. (2020). More than a methods course: Teaching preservice teachers to think historically. In C. W. Berg & T. M. Christou (Eds.), *The palgrave handbook of history and social studies education*. Palgrave Macmillan. https://doi.org/10.1007/978-3-030-37210-1_10

Gómez, C. J., Solé, G., Miralles, P., & Sánchez, R. (2020). Analysis of cognitive skills in history textbook (Spain-England-Portugal). *Frontiers in Psychology*, *11*, 1-11. https://doi.org/10.3389/fpsyg.2020.521115

Gusev, M., Kostoska, M., & Ristov, S. (2017, April). A new e-Testing platform with grading strategy on essays. In C. Patrikakis, & S. Schreiter (Eds.), *2017 IEEE Global Engineering Education Conference* (EDUCON) (pp. 676-6783). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/EDUCON.2017.7942919

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.

Hughes, C., Adlam, A., Happé, F., Jackson, J., Taylor, A., & Caspi, A. (2000). Good test-retest reliability for standard and advanced false-belief tasks across a wide range of abilities. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *41*(4), 483-490. https://doi.org/10.1111/1469-7610.00633

Indah, R. N., Rohmah, G. N., & Afifuddin, M. (2020). What i know best: Assessing Indonesian student's structural knowledge through student-generated topics. *European Journal of Educational Research*, *9*(2), 629-638. https://doi.org/10.12973/eu-jer.9.2.629

Kaipa, R. M. (2020). Multiple choice questions and essay questions in curriculum. *Journal of Applied Research in Higher Education*, *13*(1), 16-32. https://doi.org/10.1108/JARHE-01-2020-0011

Karaoğlan Yılmaz, F. G., Üstün, A. B., & Yilmaz, R. (2020). Investigation of pre-service teachers' opinions on advantages and disadvantages of online formative assessment: An example of online multiple-choice exam. *Journal of Teacher Education and Lifelong Learning*, *2*(1), 1-8. https://dergipark.org.tr/tr/pub/tell/issue/52517/718396

Keleşzade, G., Güneyli, A., & Özkul, A. E. (2018). Sosyal yapılandırmacı öğrenmeyi ve tarihsel düşünme becerilerini geliştirmeyi temel alan tarih öğretiminin etkililiği [Effectiveness of history teaching based on social constructivist learning and development of historical thinking skills]. *Education & Science/ Egitim ve Bilim*, *43*(195), 167-191. https://doi.org/10.15390/EB.2018.7479

Kesuma, A. T. (2020). The effects of MANSA historical board game toward students' creativity and learning outcomes on historical subjects. *European Journal of Educational Research*, *9*(4), 1689-1700. https://doi.org/10.12973/eu-jer.9.4.1689

Kriswantoro, Kartowagiran, B., & Rohaeti, E. (2021). A critical thinking assessment model integrated with science process skills on chemistry for senior high school. *European Journal of Educational Research*, *10*(1), 285-298. https://doi.org/10.12973/eu-jer.10.1.285

Laksana, K. (2020). Promoting historical thinking for pre-service social studies teachers: A case study from Thailand. *International Journal of Curriculum and Instruction*, *12*(2), 12-24. https://bit.ly/3G5kS1b

Leber, J., Renkl, A., Nückles, M., & Wäschle, K. (2018). When the type of assessment counteracts teaching for understanding. *Learning: Research and Practice*, *4*(2), 161-179. https://doi.org/10.1080/23735082.2017.1285422

Lombardi, M. M. (2008, January 8). *Making the grade: the role of assessment in authentic learning*. EDUCAUSE Learning Initiative. https://library.educause.edu/-/media/files/library/2008/1/eli3019-pdf

Maba, W. (2017). Teacher's perception on the implementation of the assessment process in 2013 curriculum. *International Journal of Social Sciences and Humanities*, *1*(2), 1-9. https://doi.org/10.21744/ijssh.v1i2.26

Mehany, M. S. H. M., & Gebken, R. (2021). Assessing the importance and cognition level of access student learning outcomes: Industry, educator, and student perceptions. *International Journal of Construction Education and Research*, *17*(4), 333-351. https://doi.org/10.1080/15578771.2020.1777487

Meyer, R. E. (2006). Review essay: Visiting relatives: Current developments in the new sociology of knowledge. *Organization*, *13*(5), 725-738. https://doi.org/10.1177/1350508406067011

Mingo, M. A., Chang, H. H., & Williams, R. L. (2018). Undergraduate students' preferences for constructed versus multiple-choice assessment of learning. *Innovative Higher Education*, *43*(2), 143-152. https://doi.org/10.1007/s10755-017-9414-y

Monte-Sano, C. (2010). Disciplinary literacy in history: An exploration of the historical nature of adolescents' writing. *The Journal of the Learning Sciences*, *19*(4), 539-568. https://doi.org/10.1080/10508406.2010.481014

Ningsih, T. Z., Sariyatun, & Sutimin, L. A. (2019). Development of portfolio assessment to measure student's skill of using primary source evidence. *The New Educational Review*, *52*(2), 101-113. https://doi.org/10.15804/tner.19.56.2.08

Ofianto, O., & Suhartono, S. (2016). An assessment model of historical thinking skills by means of the Rasch model. *Research and Evaluation in Education*, *1*(1), 73-83. https://doi.org/10.21831/reid.v1i1.4899

Ogunka, R. I., & Iweka, F. O. E. (2021). Application of generalizability theory in estimating dependability of public examination essay questions in English language in rivers state. *International Journal of Innovative Education Research, 9*(2), 105-114. https://bit.ly/3DqXWYN

Opstad, L. (2021). Can we identify students who have success in macroeconomics depending on exam format by comparing multiple-choice test and constructed-response test? *International Journal of Education Economics and Development*, *12*(4), 311-328. https://doi.org/10.1504/IJEED.2021.118415

Osadebe, P. U., & Nwabeze, C. P. (2018). Construction and validation of physics aptitude test as an assessment tool for senior secondary school students. *International Journal of Assessment Tools in Education*, *5*(3), 461-473. https://doi.org/10.21449/ijate.442406

Ramadhan, S., Mardapi, D., Prasetyo, Z. K., & Utomo, H. B. (2019). The development of an instrument to measure the higher-order thinking skill in physics. *European Journal of Educational Research*, *8*(3), 743-751. https://doi.org/10.12973/eu-jer.8.3.743

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*(1)*,* 58-93. https://doi.org/10.1163/24689300-01401006

Ristov, S., Gusev, M., & Armenski, G. (2015). Massive development of E-testing questions. *International Journal of Emerging Technologies in Learning*, *10*(4), 46-53 https://doi.org/10.3991/ijet.v10i4.4688

Schoemaker, P. J. (2020). How historical analysis can enrich scenario planning. *Futures & Foresight Science*, *2*(3-4), e35. https://doi.org/10.1002/ffo2.35

Seixas, P. (2006). *Benchmarks of historical thinking: A framework for assessment in Canada.* The Center for the Study of Historical Consciousness. https://bit.ly/3rAwrt8

Seixas, P. (2017). A model of historical thinking. *Educational Philosophy and Theory, 49*(6), 593-605. https://doi.org/10.1080/00131857.2015.1101363

Seixas, P., Gibson, L., & Ercikan, K. (2015). A design process for assessing historical thinking: The case of a One-Hour Test. In K. Ercikan & P. Seixas (Eds), *New Directions in Assessing Historical Thinking* (pp.102-116). Routledge. https://doi.org/10.4324/9781315779539

Seixas, P., & Morton, T. (2013). *The big six historical thinking concepts*. Nelson Education. https://bit.ly/3FX9lB8

Seixas, P., & Peck, C. (2004). Teaching historical thinking. In A. Sears & I. Wright (Eds.), *Challenges and prospects for Canadian social studies* (pp. 109-117). Pacific Educatonal Press. https://www.judithcomfort.ca/files/seixas-and-peck_2004-1.pdf

Slepkov, A. D., & Godfrey, A. T. (2019). Partial credit in answer-until-correct multiple-choice tests deployed in a classroom setting. *Applied Measurement in Education*, *32*(2), 138-150. https://doi.org/10.1080/08957347.2019.1577249

Smith, M. D. (2017). New multiple-choice measure of historical thinking: An investigation of cognitive validity. *Journal Theory and Research in Social Education, 46*(1), 1-34. https://doi.org/10.1080/00933104.2017.1351412

Soifah, U., & Pratolo, B. W. (2020). Teachers' belief, implementation, and challenges in portfolio assessment in writing. *Journal of Critical Reviews*, *7*(9), 986-990. http://www.jcreview.com/fulltext/197-1591271178.pdf

Suhaini, M., Ahmad, A., & Mohd Bohari, N. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, *10*(3), 1529-1540. https://doi.org/10.12973/eu-jer.10.3.1529

Sutimin, L. A., Joebagio, H., Sariyatun, M., & Abidin, N. F. (2018). The development of a deconstructive learning history model to promote the higher-order thinking skills of university students. *The New Educational Review*, *51*(1), 19-29. https://doi.org/10.15804/tner.2018.51.1.01

Woodcock, J. (2011). Causal explanation. In D. Ian (Ed.), *Debates in history teaching* (pp. 124-136). Routledge.

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

Zhai, X., & Li, M. (2021). Validating a partial-credit scoring approach for multiple-choice science items: an application of fundamental ideas in science. *International Journal of Science Education*, *43*(10), 1-27. https://doi.org/10.1080/09500693.2021.1923856

Appendix

Essay Test Instrument

1. Many things can make a historical event, for example, the influence of the ideas of historical figures. What other identification can cause a historical event? Write at least two answers!
2. A historical event can occur because it is caused by several factors such as social, cultural, economic, political, education, natural environmental, and religious factors. Choose at least two factors that cause a historical event and give an example of an event that occurred?
3. If you are asked to observe the cause of an event around you, explain the relationship between each cause you notice and occurrences in your daily life. Write at least two answers!
4. One of the consequences that a historical event can cause is progress or regress in various people's lives. Identify other things that might happen as a result of a historical event! Write at least two answers!
5. A historical event can have an impact on various fields in people's lives. The effect/ impact can be seen, for example, in the social and cultural, economic, political, education, and religious fields. Write at least two examples of the impact caused by various fields of people's lives due to a historical event!
6. If you are asked to observe the effect of an event around you, explain how the influence you saw has affected the lives of others around you in the short, medium, and long term? Write at least two answers!

Scoring Rubric

No	Answer Criteria	Score
1	Students identify at least two causes for a historical occurrence.	Score 0: If pupils do not describe the causes of historical events
		Score 1: If pupils can only name one cause for historical events
		Score 2: If pupils can only explain two or more causes for historical events
2	Students should select at least two factors that cause a historical event and give examples of historical occurrences caused by the elements they have chosen.	Score 0: If pupils are unable to provide instances of the causes of historical events in various fields
		Score 1: if pupils can only give one example of a historical event's cause in several domains
		Score 2: if pupils can only give two or more examples of historical event causes in diverse domains
3	If you are asked to observe the cause of an event around you, explain the comparison between each cause historical event you notice and occurrences that occur in your daily life.? Write at least two answers!	Score 0: if pupils are unable to describe the connection between the causes of a present occurrence and previous historical events
		Score 1: if pupils can only explain one link between the causes of a present occurrence and previous historical events
		Score 2: if pupils can describe two or more linkages between the causes of a present occurrence and previously occurring historical events
4	Students present two or more historical consequences.	Score 0: If pupils are unable to describe the effect of a historical event
		Score 1: if pupils can only explain one historical event's effect
		Score 2: if students can only explain two or more historical event impacts
5	Students provide at least 2 examples of the impact of a historical event in various fields.	Score 0: If pupils are unable to provide instances of the influence of a historical event on various aspects of life
		Score 1: If pupils can only give one example of a historical event's influence on numerous spheres of life
		Score 2: if pupils can identify two or more instances of a historical event's influence on diverse spheres of life
6	Students describe the short, medium, and long-term effects of the observed historical events on the surrounding life.	Score 0: if pupils are unable to explain the short, medium, and long-term effects of an observable historical event on the surrounding lives
		Skor 1: If pupils can only describe one effect of a historical event on surrounding life in the short, medium, and long term
		Skor 2: If pupils can only describe two or more effects of a historical event on surrounding life in the short, medium, and long term