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

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 difficulty 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.
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 interpretation. 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.

**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; Mehan & 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 (Ogunk & 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>
<thead>
<tr>
<th>No</th>
<th>Schools List</th>
<th>The Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>School A</td>
<td>98</td>
</tr>
<tr>
<td>2</td>
<td>School B</td>
<td>98</td>
</tr>
<tr>
<td>3</td>
<td>School C</td>
<td>102</td>
</tr>
<tr>
<td>4</td>
<td>School D</td>
<td>97</td>
</tr>
<tr>
<td>5</td>
<td>School E</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td>Total Students</td>
<td>498</td>
</tr>
</tbody>
</table>

Table 1. The Total Participant of the Research

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

<table>
<thead>
<tr>
<th>Test of Homogeneity of Variance</th>
<th>Levene’s Statistic</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The result of students’ final exam</td>
<td>Based on Mean</td>
<td>1.969</td>
<td>4</td>
<td>493</td>
</tr>
<tr>
<td></td>
<td>Based on Median</td>
<td>2.031</td>
<td>4</td>
<td>493</td>
</tr>
<tr>
<td></td>
<td>Based on Median and with adjusted df</td>
<td>2.031</td>
<td>4</td>
<td>478.961</td>
</tr>
<tr>
<td></td>
<td>Based on trimmed mean</td>
<td>2.005</td>
<td>4</td>
<td>493</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Range</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.26-4.00</td>
<td>Very Suitable</td>
</tr>
<tr>
<td>2.51-3.25</td>
<td>Suitable</td>
</tr>
<tr>
<td>1.76-2.50</td>
<td>Less Suitable</td>
</tr>
<tr>
<td>1.00-1.75</td>
<td>Not Suitable</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>The framework of analyzing cause and consequence</th>
<th>Indicators of analyzing cause and consequence</th>
<th>The form of the test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analyzing the cause of a historical event</td>
<td>a. Students can identify the cause of a historical event</td>
<td>Essay</td>
</tr>
<tr>
<td></td>
<td>b. Students can give examples from various fields that cause a historical event in people’s lives.</td>
<td>Essay</td>
</tr>
<tr>
<td></td>
<td>c. Students can recognize and compare several causes in their surrounding environment.</td>
<td>Essay</td>
</tr>
<tr>
<td>Analyzing the consequences of a historical event</td>
<td>a. Students can identify the consequences of a historical event</td>
<td>Essay</td>
</tr>
<tr>
<td></td>
<td>b. Students can give examples of the consequences of a historical event in various people’s lives.</td>
<td>Essay</td>
</tr>
<tr>
<td></td>
<td>c. Students can recognize several influences/consequences, short, medium, and long term, of the events in their surrounding environment.</td>
<td>Essay</td>
</tr>
</tbody>
</table>
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>
<thead>
<tr>
<th>Indicator</th>
<th>Score</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Material</td>
<td>3.5</td>
<td>Very Suitable</td>
</tr>
<tr>
<td>Assessment Instrument</td>
<td>3.75</td>
<td>Very Suitable</td>
</tr>
<tr>
<td>Language</td>
<td>3.75</td>
<td>Very Suitable</td>
</tr>
<tr>
<td>Mean</td>
<td>3.67</td>
<td>Very Suitable</td>
</tr>
</tbody>
</table>

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](image)

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>
<thead>
<tr>
<th>No</th>
<th>The items of Analyzing Cause and Consequence Skills</th>
<th>Validity Value (INFIT Mean of Square)</th>
<th>Difficulty value</th>
<th>Delta 1</th>
<th>Delta 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>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!</td>
<td>1.21</td>
<td>-1.23</td>
<td>-2.82</td>
<td>0.36</td>
</tr>
<tr>
<td>2.</td>
<td>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?</td>
<td>1.08</td>
<td>-0.16</td>
<td>-2.27</td>
<td>1.95</td>
</tr>
<tr>
<td>3.</td>
<td>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!</td>
<td>1.07</td>
<td>-0.62</td>
<td>-2.16</td>
<td>0.93</td>
</tr>
</tbody>
</table>
Table 6. Continued

<table>
<thead>
<tr>
<th>No</th>
<th>The items of Analyzing Cause and Consequence Skills</th>
<th>Validity Value (INFIT Mean of Square)</th>
<th>Difficulty Value</th>
<th>Delta 1</th>
<th>Delta 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.</td>
<td>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!</td>
<td>0.95</td>
<td>-0.06</td>
<td>-1.96</td>
<td>1.83</td>
</tr>
<tr>
<td>5.</td>
<td>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!</td>
<td>0.88</td>
<td>-1.18</td>
<td>-2.44</td>
<td>0.09</td>
</tr>
<tr>
<td>6.</td>
<td>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!</td>
<td>0.91</td>
<td>0.08</td>
<td>0.11</td>
<td>0.24</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Item Analysis Results or Observed Responses</th>
<th>Mean test score</th>
<th>Standard deviation</th>
<th>Internal Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean test score</td>
<td>9.99</td>
<td>3.56</td>
<td>.72</td>
</tr>
<tr>
<td>All on all (N = 150 L = 6 Probability Level= .50)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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
The essay exam is appropriate to address some of the problems and weaknesses in the prior historical thinking instrument. According to Opstand (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 (Seikas 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


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

<table>
<thead>
<tr>
<th>No</th>
<th>Answer Criteria</th>
<th>Score</th>
</tr>
</thead>
</table>
| 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 |