The Version of Record of this manuscript has been published and is available in *Cognition & Instruction*, published online on September 10, 2018, 

https://doi.org/10.1080/07370008.2018.1499646

History Assessments of Thinking:

A Validity Study

Mark Smith, Joel Breakstone, and Sam Wineburg

Stanford University

Author Note

Mark Smith (corresponding author), Graduate School of Education, Stanford University, 485 Lasuen Mall, Stanford, CA 94305, msmith4@stanford.edu, 214-577-9727.

Joel Breakstone, Graduate School of Education, Stanford University, 485 Lasuen Mall, Stanford, CA 94305, breakstone@stanford.edu, 802-281-2849.

Sam Wineburg, Graduate School of Education, Stanford University, 485 Lasuen Mall, Stanford, CA 94305, wineburg@stanford.edu, 650-725-4411.

This research was supported by the Library of Congress’s Teaching with Primary Sources program and the Bill & Melinda Gates Foundation (Contract No. 19308), Sam Wineburg, principal investigator. Although we acknowledge the generosity of our funders, we alone are responsible for the contents contained herein. We thank Rich Shavelson and Ed Haertel for their guidance in this work, and Jim Pellegrino for encouraging us not to give up on this paper until it reached publication. This research would never have gotten off the ground without the invaluable aid of Rob McEntarffer, Randy Ernst, and Teresa Wanser-Ernest. We thank them all.
Introduction

The dominant mode of assessment in the American testing landscape is just over a century old. The multiple-choice item dates to Frederick J. Kelly’s 1914 doctoral dissertation, *Teachers’ Marks: Their Variability and Standardization*. Concerns about variability in scoring on written assessments led Kelly to devise a new item type to address the need for tests that could be administered efficiently and reliably to a burgeoning population of students in American schools (Davidson, 2011; Kelly, 1914). In 1916, he published the Kansas Silent Test of Reading, which featured an item type that would become familiar to future generations of American students: “Below are given the names of four animals. Draw a line around the name of each animal that is useful on the farm: cow tiger rat wolf” (Kelly, 1916, p. 8).

The widespread use of multiple-choice questions has faced withering criticism in recent decades. Critics contend that multiple-choice items do not adequately reflect achievement in the full range of knowledge, skills, and processes that we hope students will learn in school, especially those that are the most complex (Frederiksen, 1984; Koretz, 2008; Popham, 2003; Shepard, 1992). Shading bubbles, detractors argue, is a poor approximation for being able to write clearly, to think creatively, or to generate novel solutions to problems. These critiques have prompted educators and policy makers to push for new types of assessments (e.g., Gordon Commission, 2013; National Research Council, 2013; Pellegrino, Chudowsky, & Glaser, 2001).

In no subject has the mismatch between multiple-choice items and the goals of reformers been more acute than in history, where testing has often stressed the recall of names and dates. Instead of the traditional approach to history instruction, which has students commit facts to memory, reformers have argued for curricula that engage students in the discipline’s higher-order cognitive processes. Commonly referred to as historical thinking, these processes comprise
the analytic and interpretive skills experts use to make sense of the past, such as the ability to evaluate the credibility of information, weigh competing accounts, and mount historical arguments (cf. Holt, 1990; Lévesque, 2008; Seixas, Morton, Colyer, & Fornazzari, 2012; VanSledright, 2004; Wineburg, 2001).

In recent decades, many scholars in North America and Europe have developed historical thinking curricula and curricular frameworks (cf. Körber, 2011; Rantala, 2012; Reisman, 2012; Samuelsson & Wendell, 2016; Seixas et al., 2012; van Drie & van Boxtel, 2008; von Borries, 1997). Learning standards and curriculum frameworks adopted over the past 25 years in the United States reflect the growing demand for historical thinking curricula. The Common Core State Standards, for example, feature key aspects of historical thinking, such as asking students to “evaluate an author’s premises, claims, and evidence by corroborating or challenging them with other information” (National Governors Association & Council of Chief State School Officers, 2010, p. 61). Myriad state standards also include aspects of historical thinking. The California History-Social Science Content Standards for California Public Schools, for example, require students to be able to “interpret past events and issues within the context in which an event unfolded rather than solely in terms of present day norms and values” (California Department of Education, 1998, p. 41). The Ohio Social Studies Standards ask high school history students to “analyze cause, effect, sequence and correlation in historical events, including multiple causation and long- and short-term causal relations” (Ohio Department of Education, 2010, p. 29).

In addition to new curricula, history education scholars have called for assessments to match these curricular goals (Eliasson, Alvén, Yngvéus, & Rosenlund, 2015; Ercikan & Seixas, 2015; Körber & Meyer-Hamme, 2015; Paxton, 2003; VanSledright, 2014; Waldis, Hodel,
Critics note that most history tests feature lists of discrete multiple-choice items that stress factual recall but tap few of the historical thinking processes central to the discipline. It is almost axiomatic that assessment shapes instruction (Frederiksen, 1984; Koretz, 2008; Madaus, West, Harmon, Lomax, & Viator, 1992; Popham, 2003; Shepard, 1992), and scholars have increasingly argued that the success of history curricular reform depends on the development of aligned assessments (cf. Ercikan & Seixas, 2015).

Our research group undertook a research and development project to address this need. This effort produced tasks known as History Assessments of Thinking (HATs), which are short, constructed-response items designed to reveal thinking on key aspects of historical thinking. We have engaged in a systematic effort to research issues of validity related to the use of HATs. Here we report a study that focused on cognitive validity: an examination of the alignment between the processes elicited by an assessment and the processes the assessment is designed to measure (Pellegrino et al., 2001; Ruiz-Primo, Shavelson, Li, & Schultz, 2001). Specifically, we investigated whether HATs tapped the intended historical thinking constructs. Further, we compared the effects of item format—constructed-response versus multiple-choice—on the cognitive processes elicited by these items.

Problem Statement and Context

Calls for changes in K-12 testing have grown louder as scholars, educators, and policy makers have increasingly concluded that many of the tests used to measure student achievement in the United States are outmoded and inadequate. Critics contend that the preponderance of tests in the United States have been built on the antiquated behaviorist notion that learning is the accumulation of facts and procedures, as well as the idea that assessment of student learning
involves sampling a static body of factual knowledge with a battery of discrete, interchangeable test items (Pellegrino et al., 2001; Resnick & Resnick, 1992; Shepard, 2000). Research in cognitive and socio-cultural psychology in the last half-century has shown that learning is much more complex than the simple accumulation of facts (Anderson, 2004; Chi, Glaser, & Farr, 1988; Ericsson & Smith, 1991; Greeno, Collins, & Resnick, 1996), and schools today are expected to equip students with more than the ability to retrieve facts from long-term memory (cf. National Governors Association & Council of Chief State School Officers, 2010).

In the United States, the Common Core State Standards call for schools to produce graduates able to fully participate in our democracy and economy. This participation requires the ability to solve complex problems, evaluate information from an array of sources, think critically and creatively, and communicate clearly. If schools are to promote such ambitious outcomes, critics contend, new kinds of assessments are needed (Gordon Commission, 2013; National Research Council, 2013; Pellegrino et al., 2001). The need for new metrics is particularly acute in humanistic disciplines like history. History is rooted in complex analytic and interpretative thinking processes (Voss & Wiley, 2006). In the United States, however, the evaluation of student learning in history relies heavily on standardized tests comprised of discrete multiple-choice items. As of 2011, each of the 23 states that mandated standardized testing in history/social studies used multiple-choice items for a significant portion of the exam. Fourteen of those states use only multiple-choice items to measure student achievement in history/social studies (Martin, Maldonado, Schneider, & Smith, 2011). Critics contend that multiple-choice items underrepresent the range of cognitive processes that comprise complex domains like history. In some cases, they may not tap these processes at all (Martinez, 1999; Resnick & Resnick, 1992; Shepard, 1992), and some research suggests that the types of multiple-choice
items currently used to measure history achievement—even those expressly designed to measure historical thinking—tend to measure factual recall more than historical thinking (e.g., Smith, 2017; VanSledright, 2014; Wineburg, 2004).

Reich (2009), for example, used think aloud protocols to investigate the processes used by students to answer multiple-choice questions from the New York Global History and Geography Regents Exams. He found that the items tapped mastery of history content, basic literacy, and test-wisenedness, but there was not clear evidence that the items elicited the cognitive processes identified in the New York state curriculum. Smith (2017) also used think-aloud protocols to examine the cognitive processes that 27 students used to answer four multiple-choice items from the National Assessment of Educational Progress (NAEP) grade 12 U.S. history exam and found that in no instances did the items elicit the specified aspects of “historical analysis and interpretation.”

At the same time, there are few established alternatives to multiple-choice items in the United States for measuring history achievement. Perhaps the best-known alternative is the document-based question (DBQ), a complex writing task originally developed in the 1970s by the College Board’s AP program. Today, the DBQ is featured on the AP United States History, European History, and World History exams, as well as New York’s Global History and Geography Regents Examinations. The current AP DBQ allows students 55 minutes to read six or seven historical documents and write an expository essay in response to a broad historical question. The DBQ on the 2017 AP United States History exam, for example, asked students, “Evaluate the extent of change in ideas about American independence from 1763 to 1783” (College Board, 2017).
Some contend that the DBQ is an effective measure of higher-order disciplinary skills (e.g., Gradwell, 2006). But little empirical research exists to support this conclusion (Grant, 2006), and some scholars have called the authenticity of the DBQ into question. Grant, Gradwell, and Cimbricz (2004) conducted an analysis of the New York Regents DBQ and concluded that although the DBQ had the patina of authenticity, it did not actually require students to engage in authentic selection and analysis of historical evidence when constructing an argument. Young and Leinhardt (1998) analyzed student responses to DBQs across a school year and found that students rarely engaged in sophisticated analysis of the historical documents provided. Instead, they tended to raid the documents for facts and quotes to fill the lines of an essay.

Even if the DBQ does require students to engage in aspects of historical thinking, the format has other limitations. Broad essays like the DBQ require the orchestration of so many skills that it is difficult to draw inferences about proficiency in any particular aspect of the domain. If students struggle, is it because they lack the ability to evaluate historical documents? Or are they missing key elements of knowledge? Or is it because of weaknesses in compositional fluency? It can be difficult to untangle these factors when so many cognitive processes are implicated at once. Indeed, research suggests that general composition skills are a significant determining factor in the scoring of DBQ essays (Brelan, Danos, Kahn, Kubota, & Bonner, 1994), and it is not clear how much a student’s essay reflects proficiency in historical thinking or the ability to quickly sort through a set of documents and compose coherent prose under challenging time constraints. Further, there are technical concerns about the DBQ. Extended DBQs and free-responses essays, like those featured on AP history tests, suffer from low estimates of reliability (Wainer, 2011), a concern that weakens the validity of inferences about student achievement.
In recent years, many voices have recognized the need for a wider range of historical thinking assessments and have called for new tools for measuring higher order cognitive processes (e.g., Bain & Shreiner, 2005; Eliasson et al., 2015; Ercikan & Seixas, 2015; Paxton, 2003; VanSledright, 2014; Waldis et al., 2015). In response, efforts have emerged to develop new historical thinking assessments. The College Board, for example, has revised its AP history curriculum frameworks to include historical thinking constructs and has developed new multiple-choice and short constructed-response items to measure these constructs (College Board, 2014; Robelen, 2011). Seixas, Gibson, and Ercikan (2015) have also developed new types of multiple-choice and short-answer questions intended to measure historical thinking among high school students. Although various development efforts are underway, few studies have examined the effectiveness of new measures of historical thinking. This state of affairs provided further impetus for the present work.

**Defining the Measurement Domain**

We sought to develop new assessments that differed from traditional multiple-choice and DBQ formats. Following the principles of *evidence-centered design* (ECD), the first step in our assessment development process was to specify the target domain (cf. Mislevy, Almond, & Lukas, 2004, Mislevy & Haertel, 2006; Mislevy & Riconscente, 2006). Defining the target domain at the outset of the project allowed us to articulate the specific processes we aimed to measure. Moreover, it allowed us to better evaluate whether student responses reflected proficiency in the target domain when we piloted items with students.

To begin, our target domain was *historical thinking* writ large. We surveyed the entire field and developed an outline of the processes central to the discipline. As with many complex domains, there is no singular, universally-accepted definition of historical thinking. Rather, the
broad notion of historical thinking reflects overlapping theoretical work done by various scholars to identify and define the knowledge, skills, and processes that comprise the domain (e.g., Andrews & Burke, 2007; Mandell, 2008; National Center for History in the Schools, n.d.; Seixas & Peck, 2004; VanSledright, 2004). We drew upon the scholarship of the history education community (e.g., Bain, 2006; Booth, 1993; Dickinson & Lee, 1984; Holt, 1990; Körber, 2011; Lévesque, 2008; Levstik & Barton, 1996; Seixas & Peck, 2004; Shemilt, 1983; van Drie & van Boxtel, 2008; VanSledright, 2004; Wineburg, 1991, 1998, 2001), historians (e.g., Appleby, Hunt, & Jacob, 1994; Carr, 1961; Evans, 2000; Fischer, 1970), and philosophers of history (e.g., Collingwood, 1946/1993; Danto, 1965; Mink, 1987; Rüsen, 1987) to define the central aspects of historical thinking. After surveying this expansive landscape, the target domain was narrowed to focus on the aspects of cognition that we saw as essential to learning history in secondary classrooms (ages 11 years to 18 years) in the United States. In short, we accounted for the state of historical thinking broadly and then determined which disciplinary processes were key for secondary students to master.

The resulting domain, which we call *Historical Thinking for the Secondary Classroom*, comprises three broad pillars:

I. *Historical Knowledge* includes reasoning processes that require students to apply knowledge of the past, like the ability to evaluate historical significance, to explain how events are connected across time, and to see patterns and themes across time.

II. *Evaluation of Evidence* comprises the expert processes employed by historians when making sense of historical evidence, including the ability to consider the source of information when evaluating reliability, to reason about evidence as a product of its time, and to corroborate claims and evidence.
III. *Use of Evidence/Argumentation* encompasses the discipline-specific skills needed to use historical evidence to make appropriate claims and to clearly communicate sound historical arguments. Included are skills like the ability to engage in causal reasoning, to make appropriate generalizations, and to reason counterfactually.

These three pillars provided an overarching structure to the domain, but we felt that a more specific delineation of the domain was necessary to promote clarity and precision when creating assessments and evaluating validity. Each of the three pillars was further defined using the following structure:

1) **Pillar** – one of the three central components of the domain (e.g., *Evaluation of Evidence*)
   
   a) **Constructs** – the primary processes of each pillar (e.g., Evaluation of Evidence → *Sourcing, Contextualization, Corroboration*)
   
   i) **Facets** – the specific processes of each construct that secondary students are expected to master (e.g., Evaluation of Evidence → Corroboration → *Verification, Comparison, Articulation of Need*)

Our aim in defining the domain was a departure from most attempts to define historical thinking to date, which have tended to focus on broader theoretical aspects of historical thinking like “continuity and change” and “cause and consequences” (cf. Seixas et al., 2012). We took a more assessment-specific approach. Our goal was to define a domain that would allow us to carefully specify the facets of historical thinking that assessments would target and to better evaluate whether they were tapping these processes.

Our measurement domain reflected American education reform goals and prevailing notions of historical thinking in the United States. Although we tailored the measurement domain for American classrooms, many of its constructs appear in international scholarship. The
“evaluation of evidence” pillar, for example, included constructs identified in historical thinking models from Canada (e.g., Lévesque, 2008; Seixas et al., 2013), the Netherlands (e.g., van Drie & van Boxtel, 2007), and Germany (e.g., Körber, 2011). However, the domain does not incorporate aspects of these models that have been less influential in the United States. For example, we did not include some aspects of “historical consciousness” in the FUER model, like “the ability to (re-)shape one’s self in relation to the outside world and the past,” or, “the ability to (re-)shape the own conceptions of what can be done, achieved, hoped for in the present and future . . .” (Körber, 2011, p. 150).

**Item Development**

Inspired by principles of *design thinking* (cf. Brown, 2009; Martin, 2009; Stanford University Institute of Design Thinking, n.d.), we set out to generate new historical thinking tasks. The idea for HATs evolved over four months of brainstorming – or “ideation” in design thinking terms (Brown, 2009; Stanford University Institute of Design Thinking, n.d.). We believed that presenting students with one or two historical documents and asking targeted constructed-response questions might hold advantages over restricted-response items such as multiple-choice and true/false. Prompting students to explain their reasoning might provide more evidence of student thinking than a shaded bubble and, therefore, yield better information from which to draw inferences about students’ historical thinking.

To develop tasks, we engaged in iterative cycles of prototyping and testing. Group members would generate ideas for items and present them for discussion and analysis. Promising items went through cycles of piloting and revision. Each item was piloted with hundreds of secondary history students across the United States and evidence from student responses was
used to revise assessments for clarity and effectiveness (cf. Breakstone, 2014). Items that seemed effective in piloting were selected for further validity testing.

**Sample HATs**

Figure 1 provides an example of a HAT that targets the construct of *sourcing* (Brante & Strømsø, 2017; Wineburg, 1991), the process of evaluating the reliability of historical evidence based on the time, place, and circumstances of its creation. In nearly every study of historical reading to date, sourcing is one of the touchstones that distinguish expert from novice practice (e.g., De La Paz et al., 2014; Leinhardt & Young, 1996; Monte-Sano & De La Paz, 2012; Mosborg, 2002; Nokes, Dole, & Hacker, 2007; Reisman, 2012; Rouet, Favart, Britt, & Perfetti, 1997; Shanahan, Shanahan, & Misischia, 2011; Shreiner, 2014; Wineburg, 1991, 1998). In particular, this HAT targets a facet of sourcing we called *date of creation*, which is an evaluation of the relationship between when a source was created and the date of the event, person, or place it depicts. To assess this facet, the HAT presented students with a 20th century painting that portrays the First Thanksgiving in 1621. It then asked them to explain whether they agree or disagree that the painting would be useful to a historian as evidence of what happened between the Pilgrims and the Wampanoag in 1621.
Piloting suggested that students who were proficient in sourcing recognized the three-century gap in time as problematic and explained how the gap might limit the usefulness of the painting as evidence for events in 1621. Pilot data also showed that the assessment revealed common misconceptions in student thinking. One misconception follows the adage of "a picture is worth a thousand words." According to this reasoning, the painting is useful because it shows an image of what happened, or as one student reasoned, "The painting shows the two sides getting along and befriending one another. A picture allows readers to infer their relations were okay." Another common misconception involved students evaluating the veracity of the evidence
based on how well it matched the understanding they brought to the task. For example, one student answered, “I agree [it would be useful to historians], because the Native Americans and the white settlers are eating and sharing with each other.” For this student, the provenance of the evidence was unimportant compared to whether it matched a prior understanding of the past.

Figure 2 provides an example of a second HAT, in this case one designed to measure students’ ability to engage in *contextualization*, the interpretation of a document in light of the historical context in which it was created (Wineburg, 1991). Students are presented with a playbill for a 1930s production about the American abolitionist John Brown’s raid at Harper’s Ferry. Students are then asked to identify the year in which the play was written. This question about the date, which no student missed in piloting, was an intentional scaffold to help orient the students in time. The assessment then presented students with four facts and asks which *two* might help explain why the authors wrote the play in 1936. All of the facts are true, but only two are germane to the socio-political context of the 1930s and relevant for understanding the authors’ motivations in writing the play. The other two relate to the time of John Brown’s raid on Harpers Ferry but are less useful for understanding the context of the 1930s. Students are then asked to choose two facts and explain in a sentence or two why they are relevant to understanding why the authors wrote the play. Piloting indicated that proficient students were able to select the two facts from the 1930s and provide a plausible explanation for how these facts could have affected the authors’ motivations. Less proficient students missed the idea that historical evidence is a product of the temporal context in which it was created. Instead, they gravitated to facts about the time in which the play was set.
Validity

Responsible test development requires systematic evaluation of validity. Today, measurement theorists adhere to an argument-based approach to evaluating validity (AERA, APA, & NCME, 2014; Cronbach, 1988; Kane, 2006, 2013). Rather than a property of a test, validity is a logical argument about the soundness and plausibility of the interpretations of test results (Kane, 2006, 2013). As Pellegrino, DiBello, and Goldman (2016) note, “An assessment is
not valid or invalid per se . . . but rather that validity has to be judged relative to the intended interpretive use of the results” (p. 16). HATs were designed to support inferences about student proficiency in historical thinking, so a sound validity argument about the use of HATs to draw inferences about student historical thinking requires evidence that these inferences are sound.

Here, we report the results of a study designed to gather evidence central to a validity argument about the use of HATs as measures of historical thinking. More specifically, this study yielded evidence of cognitive validity, which focuses on the alignment between the processes an assessment was designed to measure and the processes it actually elicits with test-takers (Pellegrino et al., 2001; Ruiz-Primo, Shavelson, Li, & Shultz, 2001). A validity argument for HATs rests, in part, on the assumption that these items prompt students to engage in historical thinking and that student responses reflect proficiency in this domain. If not, then an argument for the use of HATs as measures of historical thinking would be undermined.

**Research Questions**

This study used think-aloud protocols to explore two research questions related to cognitive validity:

1. Do HATs elicit the constructs of historical thinking they were designed to measure?

2. How do the cognitive processes elicited by HATs compare with those elicited by stem-equivalent multiple-choice items?

The first question addressed a primary consideration for cognitive validity: Do students engage in the intended cognitive processes when completing the assessment? The second question also focused on evidence relevant to cognitive validity. It considered the effect of item format on the kinds of cognitive processes elicited. Do HATs elicit different processes than stem-equivalent multiple-choice items? Could multiple-choice items be used to measure the same processes? No
published research has explored the processes elicited by stem-equivalent history assessments. Little is known about whether items that share the same prompt and stimuli but have different response structures evoke different thinking processes. We hoped that answers to the second question would yield evidence about the effects of item format on the processes elicited among students.

**Research Methods**

This study employed *concurrent* and *retrospective* think-aloud protocols to collect evidence about the thinking processes elicited by both HATs and their stem-equivalent multiple-choice parallels (Ericsson & Simon, 1993; Taylor & Dionne, 2000). Decades of research suggests that think-aloud protocols are an effective tool for revealing the cognitive processes that individuals use to solve problems in a variety of domains (e.g., Ericsson & Simon, 1993; Kuusela & Pallab, 2000; Leighton, 2004), including the complex cognitive processes experts and novices use when working through history problems (e.g., Shanahan & Shanahan, 2008; Wineburg, 1991, 1998).

Think-alouds are also a recommended method for eliciting evidence about whether test items tap the intended knowledge, processes, and skills. They are also a widely-accepted tool for evaluating the validity of test score interpretations (AERA et al., 2014; Kane, 2006, 2013). Further, think-aloud protocols have been used extensively to reveal the processes and strategies students use to solve standardized test items (e.g., Farr, Pritchard, & Smitten, 1990; Hamilton, Nussbaum, & Snow, 1997; Katz, Bennett, & Berger, 2000) and have been embraced by history education scholars as useful method for examining the cognitive demands of assessments designed to measure historical thinking (e.g., Ercikan, Seixas, Lyons-Thomas & Gibson, 2015; Kaliski, Smith & Huff, 2015). Thus, there is reason to believe that think-aloud interviews could
yield evidence about whether HATs and their stem-equivalent multiple-choice items elicit the targeted aspects of historical thinking.

**Materials**

Materials included eight constructed-response HATs and eight stem-equivalent multiple-choice items. Each of the HATs selected had survived the iterative item development process that we used and had shown promise during piloting. We created eight parallel multiple-choice items based on established design conventions for writing multiple-choice items (cf. Haladyna, 2004; Osterlind, 1998). Each multiple-choice parallel used the same exact documents as the parallel HAT and was designed to measure the same construct. To the greatest degree possible, the exact same wording was used in the multiple-choice question stem as the parallel HAT question (see Figure 3 for the stem-equivalent multiple-choice items for the Thanksgiving and John Brown Playbill HATs).

Answer choices for the multiple-choice items were written to reflect the types of thinking that we observed when piloting HATs with over 7,400 students across the country in the first two years of the project. The correct answer choice demonstrated proficient reasoning in the targeted aspect of historical thinking, and incorrect choices included the most common mistakes or misconceptions students made when answering constructed-response HATs in piloting. Thus, the multiple-choice items presented students with prompts that were highly similar (or identical) to their HAT counterparts, but rather than generate a response to the prompt, students selected an answer from among the most common types of student responses.
Figure 3. HATs and stem-equivalent multiple-choice questions.

Thanksgiving HAT

Directions: Use the painting to answer the question below.

Title: "The First Thanksgiving 1621"
By: J.G. Farris
Date Painted: 1992

Question: The painting, "The First Thanksgiving 1621," helps historians understand the relationship between the Wampanoag Indians and the Pilgrim settlers in 1621. Do you agree or disagree? (Circle one).

Question 1: Briefly support your answer.


Thanksgiving multiple-choice question

Directions: Use the painting to answer the question below.

Title: "The First Thanksgiving 1621"
By: J.G. Farris
Date Painted: 1992

Question: Do you agree or disagree with the following statement? "The painting, The First Thanksgiving 1621, helps historians understand the relationship between the Wampanoag Indians and the Pilgrim settlers in 1621."

- Agree, because it is impossible to know what the first Thanksgiving was like without seeing a picture of it.
- Agree, because the painting shows the Puritans and Wampanoag getting along peacefully in the spirit of Thanksgiving.
- Disagree, because the painting was made over 300 years after the first Thanksgiving.
- Disagree, because it shows the relationship between the Puritans and Wampanoag as happier than it really was.

John Brown HAT

Directions: Use the background information, your knowledge of history, and the poster to answer the questions below.

Background information: This is a poster for a play written in 1956 that celebrates the abolitionist John Brown, who tried to start a slave revolt in Harpers Ferry, Virginia in 1859.

Question 1: When was the play written?

Question 2: Which two of the facts below might help explain why the authors wrote this play?

- Slaves made up nearly 40% of Virginia’s population in 1859.
- One of the play’s authors, Michael Gold, was a member of the Communist Party, which protested against lynching in the 1930s.
- After taking power in 1933, Adolf Hitler enacted racist policies in Germany.
- After seceding from the Union in 1861, Virginia became the largest state in the Confederacy and the home of its capital, Richmond.

John Brown multiple-choice question

Directions: Use the background information, your knowledge of history, and the poster to answer the question below.

Background information: This is a poster for a play written in 1956 that celebrates the abolitionist John Brown, who tried to start a slave revolt in Harpers Ferry, Virginia in 1859.

Question: Which of the facts below help explain why the authors wrote this play in 1956?

- Fact 1: Slaves made up nearly 40% of Virginia’s population in 1859.
- Fact 2: One of the play’s authors, Michael Gold, was a member of the Communist Party, which protested against lynching in the 1930s.
- Fact 3: After taking power in 1933, Adolf Hitler enacted racist policies in Germany.
- Fact 4: After seceding from the Union in 1861, Virginia became the largest state in the Confederacy and the home of its capital, Richmond.

- Both fact 1 and fact 3.
- Both fact 2 and fact 3.
- Fact 1, Fact 2, Fact 4.
- Only fact 4.
- None of the facts.
Participants

We collaborated with an urban public school district in the Midwest to recruit participants who had completed a one-year AP United States History (APUSH) course and had scored a three or better (on a five-point scale) on the national APUSH exam. According to the College Board, a score of three or better indicates that a student is capable of doing work in an introductory college class in the subject (College Board, 2014). These selection parameters ensured that students had the opportunity to learn the material covered by the tasks. All items selected for the study drew upon content covered by the APUSH curriculum. The high bar we set for selection also reduced the likelihood that the literacy demands of the items would impose barriers to students’ cognitive processes. The district identified and recruited 26 participants across three high schools who matched our criteria. Twenty-five were seniors (the final year of high school) and one participant was a junior (the penultimate year of high school).

Data

Data included transcribed verbal reports from the concurrent and retrospective think-aloud interviews (Ericsson & Simon, 1993; Taylor & Dionne, 2000). During the concurrent portion of the interviews, students were asked to think aloud without interruption while solving the items. After students completed each item, they were asked retrospective questions to probe and clarify their thinking. Concurrent reports are thought to yield more accurate data than retrospective reports about the processes individuals use to solve problems (Ericsson & Simon, 1993; Nisbett & Wilson, 1977; Pressley & Afflerbach, 1995; Taylor & Dionne, 2000; White, 1980). Thus, the analyses presented here focus primarily on concurrent interviews; retrospective data were used only to clarify inconsistencies.

Participants were randomly assigned to one of two groups. Students in each group responded to a different set of items. Students in Group 1 completed protocol Form A,
comprising four constructed-response HATs and four non-equivalent multiple-choice items. Students in Group 2 completed Form B, which included four different HATs and four different multiple-choice items. In other words, the multiple-choice items on Form A were stem-equivalent to HATs from Form B. This crossed design allowed us to compare the processes elicited by each HAT and its stem-equivalent multiple-choice question.

**Analysis**

Student interviews were audio recorded and transcribed verbatim. Two analyses were conducted with the interview transcripts.

**Analysis 1: Cognitive Processes Elicited.** The first analysis was designed to reveal whether both HATs and their stem-equivalent multiple-choice items elicited the intended aspects of historical thinking. To do this, verbal protocols were iteratively coded to discern the cognitive processes students used to answer each item. (See Appendix for codes.) Next, each of the responses was coded for whether the processes elicited were the intended constructs of historical thinking. If a student’s answer reflected proficiency in the targeted construct—or engaged in a process that demonstrated a lack of understanding about the targeted construct—then the response was coded as *construct-relevant*. Thus, if an item was designed to gauge student ability to *source*, then the item was coded as *construct-relevant* if the student engaged in sourcing or engaged in a process that revealed a fundamental misunderstanding about sourcing.

Figure 4 provides an example of a coded transcript from a concurrent report in response to the Thanksgiving HAT. The iterative coding process identified two aspects of sourcing that this student used to arrive at an answer: genre and date. The student reasoned about both the limitations of the medium of information—a painting—and questioned the reliability of the evidence based on the gap between the date of the event and the date of the painting’s creation.
Because the student engaged in sourcing, the transcript as a whole was coded as *construct-relevant*. In cases where students’ thought processes were irrelevant to the targeted construct, the responses were coded as *construct-irrelevant*. Thus, when a student engaged in a test-taking strategy like process of elimination, the response was coded as *construct-irrelevant* (see Figure 4 for a transcript coded as construct-irrelevant).

*Figure 4. Construct-relevant and construct-irrelevant coding for Analysis 1.*

<table>
<thead>
<tr>
<th>Protocol transcript for Thanksgiving Assessment</th>
<th>Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>The painting</em> The First Thanksgiving 1621 is a useful resource for historians who wish to understand the relationship between the Wampanoag Indians and the Pilgrim settlers in 1621. Do you agree or disagree? Briefly support your answer.</td>
<td>Construct-relevant; historical thinking; Sourcing – genre</td>
</tr>
<tr>
<td>I disagree because you know paintings can be false and inaccurate representations. They can be biased toward an opinion.</td>
<td></td>
</tr>
<tr>
<td>I would be especially concerned with the fact that the painting was painted in 1932 when it is so far from the date of 1621. I would be a skeptic.</td>
<td>Construct-relevant; historical thinking; Sourcing – date</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protocol transcript for John Brown Assessment</th>
<th>Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Fact 1] Slaves made up nearly 40% of Virginia’s population in 1859.</td>
<td>Construct irrelevant; goodness of fit;</td>
</tr>
<tr>
<td>Fact 1 could help explain why the authors wrote this play, because it was the only fact that dealt with slavery directly. [Student turns the page over and sees there are spaces for two answers.]</td>
<td>Student evaluates the relative veracity or plausibility of an answer vis-à-vis other possible answer choices</td>
</tr>
<tr>
<td>Oh, there are two facts. Then Fact 1 is not right because it stands alone. It's the only one that does <em>that</em>. If there’s two facts, I think it will be the two modern facts that have to do with when it was written and not when [the raid on Harpers Ferry] occurred. So, I’m going to change my answer to Fact 2 and 3.</td>
<td></td>
</tr>
</tbody>
</table>

To review, our approach to coding yielded evidence relevant to the two research questions. It allowed for an exploration of whether the items tapped the intended constructs as well as whether there were differences in the processes due to item format. (In cases where
students engaged in both construct-relevant and construct-irrelevant reasoning in the same response, the item was double coded. Thus, it was possible for a single item to include multiple codes.

A second rater coded a sample of 48 of the 208 transcripts (both HATs and multiple-choice). Estimates of inter-rater agreement were high for coding on cognitive processes (Cohen’s Kappa = .91) as well as whether these processes matched the intended construct (Cohen’s Kappa = .92).

**Analysis 2: Answer alignment.** At the heart of cognitive validity is an examination of whether a student’s written answer reflects proficiency in the intended cognitive processes. Under normal test conditions, written answers are the only evidence of student thinking available to assess achievement. So, if the aim of the assessment is to measure cognitive processes, then it is crucial that written responses reflect the thinking students used to arrive at the answer. The second analysis was designed to examine this alignment.

Students’ think-aloud transcripts were first coded at the item level for the primary cognitive process used to arrive at an answer. Second, students’ written responses were coded for the degree of alignment between the think-aloud transcript and the written answer. Each transcribed protocol/written answer pair was assigned one of three codes:

1. **Process Match:** The written response reflected a student’s thinking.
2. **Full Mismatch:** The written response and think-aloud were misaligned. This occurred when a student’s written answer was correct, but the think-aloud showed considerable misunderstanding; or, conversely, when a written answer reflected a lack of proficiency, but the think-aloud transcript suggested otherwise.
3. **Partial Mismatch**: Written response reflected misunderstanding, but the think-aloud provided evidence of partial understanding; or, conversely, written response reflected understanding, but the think-aloud provided evidence of some misunderstanding.

Codes were tallied for the protocols so that alignment between online thinking and written response could be compared by item type. A random number generator was used to select 48 of the 208 responses (numbered in the order that they were recorded). A second rater coded the sample of 48 responses using the three codes, with high levels of reliability (Cohen’s *Kappa* = .98 for the HATs and Cohen’s *Kappa* = .87 for the multiple-choice items).

Figure 5 displays a coded transcript for the Thanksgiving prompt. The concurrent verbal report revealed that sourcing was the primary process that the student used to arrive at an answer. Thus, the transcript was coded as “Historical Thinking – Sourcing – Date” (see Appendix for description of codes). The student written response (shown in Figure 5) reflected the historical thinking processes revealed in the concurrent report, so the student written response was coded as a “Process Match.”
Results

Analysis 1: Do HATs and stem-equivalent multiple-choice items tap the intended constructs?

Finding 1: HATs tapped the intended historical thinking constructs.

The first analysis revealed that HATs were effective in eliciting the targeted facets of historical thinking. Protocol analysis revealed that HATs elicited relevant processes in all student responses (104/104 cases). Student reasoning about the Thanksgiving task, for example, illustrated how a HAT tapped *date of creation*, a facet of *sourcing*. Four of the 13 AP students in the sample demonstrated proficiency in this process when answering the Thanksgiving HAT. A response from a student named Jennifer provided an example of how the task elicited historical thinking (see Figure 6). Jennifer noticed the gap in time and questioned the usefulness of a painting made roughly three centuries years after the event occurred. She explained how the document was more likely the product of the 20th century than a reflection of the 17th century.
Further, she understood how the source might be limited as evidence because the author may not have been in a position to know what happened three centuries earlier.

*Figure 6.* Student think aloud responses to Thanksgiving HAT and stem-equivalent multiple-choice item.

<table>
<thead>
<tr>
<th></th>
<th>HAT</th>
<th>Stem-equivalent multiple-choice item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td><strong>Jennifer:</strong> Do I agree or disagree [that the source would be useful evidence]? It was painted a long time after this event was supposed to happen, so I don’t know. I just don’t feel that it’s too relevant. I mean it is three hundred years later. They probably just painted what they thought from <em>that</em> time. So, no, it is unlikely that this is a primary source because any information from the picture will have probably come from what the artist knows about the First Thanksgiving at <em>that</em> time.</td>
<td><strong>Becky:</strong> I’m going to disagree [that it is useful evidence] because the painting was painted more than three hundred years after the First Thanksgiving. So, it’s just an impression of what happened. The person wasn’t actually standing there watching the dog trying to steal corn or the settlers mingling with the Indians. So, let’s see the answers ... I’m going to go with C, disagree just because there’s not enough information for telling if that’s an impression or not.</td>
</tr>
<tr>
<td>Incorrect</td>
<td><strong>William:</strong> <em>Do you agree or disagree?</em> Um, I’m actually not sure. I haven’t really gone into that before, but I wouldn’t ... Probably for this scene, I would agree. I guess that I remember that they did interact with them a little ... quite a bit in the beginning – a little more peacefully in the beginning.</td>
<td><strong>Emily:</strong> <em>Answer D Disagree, because it shows the relationship between the Puritans and the Wampanoag as happier than it really was.</em> Well, I’m going to cross out A just because there’s better reasons that are listed; and, I’m also going to cross out B because they really didn’t get along peacefully. And, yeah, the picture shows it, but that’s not really true. So, I’m going to go with D because it shows them getting along and walking together and feeding each other, and that’s not really what happened.</td>
</tr>
</tbody>
</table>

Responses also showed that HATs were effective in evoking relevant misconceptions about historical evidence. A verbal report from William (see Figure 6 above) revealed a cognitive process exhibited by nine of the thirteen participants in response to the Thanksgiving HAT. Like eight other students, William focused on content of the picture while ignoring the accompanying bibliographic information. Whereas Jennifer saw the painting as a product of its
time, William ignored its provenance – how and when the object came into being. Instead, he evaluated the evidence based on whether it matched what he already believed about the past, a version of historical confirmation bias. The picture matched his understanding of Thanksgiving, so it was deemed useful. Although Jennifer and William engaged in very different forms of reasoning, both answers provided evidence of their ability to reason about historical evidence. Jennifer’s response reflected proficiency in the construct, while William’s revealed a fundamental misconception about evidence.

Finding 2: Stem-equivalent multiple-choice items tapped historical thinking processes, but they prompted students to engage in more instances of construct-irrelevant reasoning than did HATs.

Multiple-choice items also elicited relevant historical thinking processes. Out of 104 total student responses to multiple-choice items, 102 (98%) revealed proficiency (or lack thereof) in the targeted construct. (The two responses that did not provide information about student proficiency were classified as “inconclusive” because student responses did not provide enough data to be coded.) A response from Becky showed how the Thanksgiving multiple-choice item (see Figure 6) elicited construct-relevant reasoning. Before reading the answer choices, Becky noted the date of the painting. The data suggest that Becky saw the problem of the gap in time and understood how that might limit the painting’s value as evidence of the First Thanksgiving. She then went to the answer choices and found the one that matched the answer that she had already generated when evaluating the source.

At the same time, multiple-choice items also prompted students to engage in more instances of construct-irrelevant reasoning than did the corresponding stem-equivalent HATs. Although multiple-choice items elicited aspects of historical thinking at high rates, they also prompted students to engage in construct-irrelevant processes in 75% of the responses (78 of
HATs, on the other hand, elicited construct-irrelevant reasoning in only seven percent of the responses (8 of 104). Table 1 shows the instances of construct-irrelevant reasoning by item and format. This table shows that each HAT elicited fewer instances of construct irrelevant reasoning than its multiple-choice counterpart. Further, each multiple-choice item elicited irrelevant processes for at least six of the 13 student responses. No instances of irrelevant processes were observed for six of the HATs. Two HATs did, however, elicit irrelevant processes. (These instances will be discussed below.)

<table>
<thead>
<tr>
<th></th>
<th>Thanksgiving</th>
<th>NAACP</th>
<th>Philippines</th>
<th>Slave Quarters</th>
<th>Edison</th>
<th>Fredericksburg</th>
<th>Paterson</th>
<th>John Brown</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multiple-choice</strong></td>
<td>11</td>
<td>11</td>
<td>9</td>
<td>12</td>
<td>8</td>
<td>13</td>
<td>7</td>
<td>7</td>
<td>78</td>
</tr>
<tr>
<td><strong>HAT</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td><strong>Difference a</strong></td>
<td>11</td>
<td>11</td>
<td>9</td>
<td>12</td>
<td>8</td>
<td>13</td>
<td>4</td>
<td>2</td>
<td>70</td>
</tr>
</tbody>
</table>

Two irrelevant processes were prevalent in the responses to multiple-choice items. One was a process of “best fit” reasoning in which students weighed answer choices against each other to find the one that best fit the demands of the question. This was a process observed in 70% (73 of 104) of the responses. Stephen provided an example of “best fit” reasoning (see Figure 3 for the Thanksgiving multiple-choice item):
So, it's going to be [answer] A or B. I don’t think it will be C or D. [Answer] A is *impossible to know what the First Thanksgiving was like without seeing a picture of it.*

No, you can also have text and stuff. [B] *Because the painting shows Puritans and Wampanoag getting along.* Yeah, that could be it. It shows both of them and it shows what they’re doing. And they got along. So, yeah, it's going to help understand the relationship. [But] I’ll read C and D to make sure. [C] *the painting was made over three hundred years...* that’s not relevant because they still know what's going on there. And [answer] D *describes how it shows the relationship in the period happier than it really was.* That could be true also. That actually might be best. It’s either B or D. I don’t know.

But I guess according to the painting they’re having a good time. So, I’ll go with B.

The multiple-choice item was effective in eliciting relevant reasoning about the importance of the gap between when the source was painted and the event depicted. But for Stephen, the gap was “not relevant.” Like William’s response (in Figure 6), the main consideration for Stephen was whether the source’s content matched his understanding of the past. Stephen believed that the First Thanksgiving was an amiable affair, so the painting must be sound evidence because it showed that “they’re having a good time.” However, Stephen also engaged in a process of weighing answer choices against each other to determine which best fit the demands of the question. After reading the prompt, Stephen immediately began evaluating answer choices serially to see which fit. As he descended the list, he alighted on answer B, but he continued down the list to see if there was a better choice. When he encountered choice D, which he thought “could also be true,” he weighed the two answers against each other before ultimately settling on choice B. This type of “best fit” reasoning may be an effective heuristic for solving multiple-choice items but it is foreign to the discipline of history.
Multiple-choice items prompted a second construct-irrelevant process. “Elimination” was observed in 73% of the student responses (76 of the 104 cases). This process involved narrowing the problem space by ruling out answer choices and focusing attention on more plausible choices. Stephen’s response exemplifies an “elimination” response. Stephen first tried to narrow the answers by focusing disproportionately on answers A and B. As he read the other choices, he also considered C and D. In the end, he eliminated choices A and C, and focused his attention on only answers B and D. Between the two, he decided that B was the better choice.

The types of irrelevant test-taking strategies seen in response to multiple-choice items were rarely observed in response to HATs. There were exceptions, however. Two HATs, John Brown Playbill (see Figure 2) and Work in Paterson (see Figure 7), elicited construct-irrelevant reasoning similar to that observed with multiple-choice items. These two HATs shared a common structure. Both presented students with a historical document and four facts. Students were asked to determine which two of the four facts could be relevant for evaluating the reliability of the document and then required to compose a short explanation of how each fact would help them evaluate the document.
Both HATs were designed to see if students could use contextual information to reason about historical evidence. Although the two items elicited construct-relevant thinking in all 26 responses, they also prompted 8 of 26 students to engage in irrelevant processes when selecting two facts. Anthony said this in response to the John Brown Playbill HAT:

[Fact 1] *Slaves made up nearly 40% of Virginia’s population in 1859.*

Fact 1 could help explain why the authors wrote this play, because it was the only fact that dealt with slavery directly.
[Anthony turns the page over and sees there are spaces for two answers.]

Oh, there are *two* facts. Then Fact 1 is not right because it stands alone. It's the only one that does that. If there’s two facts, I think it will be the two modern facts that have to do with when it was written and not when [the raid on Harpers Ferry] occurred. So, I’m going to change my answer to Fact 2 and 3.

Anthony incorrectly believed that Fact 1—that slaves made up 40% of Virginia’s population in 1859—was the most helpful for understanding why the authors wrote the play in 1936. It was only after he realized that he had to choose *two* facts that he went in search of an alternative. In the retrospective interview, Anthony elaborated on why he changed his answer:

I forgot that there were two facts. When I got to the back, I had picked that one fact and it was just a stand-alone fact. A lot of times three facts are similar and then the one fact that is stand-alone usually is right. Then when I realized it's two facts, I changed.

Anthony’s response was similar to what we observed with multiple-choice items. Like multiple-choice items, this item format asked students to choose from a bounded list. As a result, the item format elicited the types of construct-irrelevant processes often prompted by multiple-choice items. As was the case with multiple-choice items, inferences about student ability are undermined when the structure of an item systematically elicits high rates of construct-irrelevant processes.

In addition to construct-irrelevant processes, HATs and multiple-choice items differed in the extent to which students generated their own evaluation of the historical evidence. In all but two cases, HATs prompted students to engage processes similar to what Guilford (1959) called *divergent production*. With HATs, students generated their own evaluations of the historical
documents. In the case of the Thanksgiving HAT, for example, students independently formulated an explanation for why the painting might (or might not) be useful. Multiple-choice items, on the other hand, were less likely to prompt independent reasoning. Instead, many students immediately evaluated answers in an attempt to converge on a single correct answer. Less than a third of students answering multiple-choice items (31 of 104) generated an answer before examining the four answer choices. Thus, instead of formulating their own evaluations of the evidence during think-aloud interviews, they picked one from a bounded list.

**Analysis 2: Do written answers reflect proficiency in historical thinking?**

*Finding 1: Written responses to HATs reflected student proficiency in historical thinking.*

A high proportion of written responses to HATs reflected student proficiency in the aspects of historical thinking the items were designed to measure. Of the 104 written responses, 99 (95%) matched the cognitive processes that students used to arrive at the answer. In only five instances were answers out of sync with student proficiency. Figure 8 provides a visual representation of the coding results for each HAT. (Visit sheg.stanford.edu/history-assessments to view the HATs referenced in the figure.) The chart shows that there were no instances of misalignment between proficiency and student answers for five of the eight HATs. Two HATs—Fredericksburg and Paterson—had one instance of partial misalignment in which the student revealed a higher level of proficiency in the think-aloud than was reflected in the written answer. In both cases, students provided abbreviated or incomplete written responses that reflected some proficiency but did not fully reflect their thinking. The Thank You Edison HAT yielded three instances in which items were coded as “full mismatch.” The written responses led to completely incorrect inferences about student proficiency in the targeted construct. (These instances will be discussed further in Finding 2 below.)
Finding 2: Students’ written responses to multiple-choice items did not reflect student thinking as closely as did written answers to HATs.

The second analysis showed that written answers to HATs better reflected student proficiency in historical thinking than did stem-equivalent multiple-choice items. Figure 8 provides an overview of the coding results for each multiple-choice item. Out of the eight multiple-choice items, only the Edison multiple-choice item had a rate of alignment as high as its associated HAT. For all other items, HATs outperformed their multiple-choice counterparts.

In some instances, student answers to multiple-choice items completely misrepresented their thinking. Twelve of the 104 student responses were coded as “full mismatch.” In these cases, the multiple-choice answer would lead to a completely incorrect inference about the nature of student thinking. Stephanie’s response provides an example. In response to the John Brown Playbill multiple-choice item (see Figure 3), Stephanie reasoned:
I’m kind of thinking almost all four facts can show why the authors wrote the play.

Fact 1 is saying that slaves were nearly half of all the people in Virginia. This is a fairly large population.

And then [Fact 2], one of the play’s authors was a member of the Communist Party, which protested against lynching. So he may not have been for integration, but chances are he probably had some thoughts of equal rights.

The same goes for Fact 3 [about Hitler enacting racist policies after taking power in 1933], you are kind of seeing the racism take hold and you’re getting some thoughts in your mind about basic human rights. Everybody should be treated fairly. This doesn’t always happen, but it should.

Fact 4 fits.

I would probably say – if I had to – that the best two would be Facts 2 and 3, showing that these are people, too, and that they deserve to be treated fairly. Facts 2 and 3 kind of are looking towards the more humanistic side and not the purely statistical side.

For Stephanie, all the choices were relevant. She landed on the right answer without showing proficiency in the targeted process of contextualization. Instead, she chose the correct answer because she favored “humanistic” explanations over “statistical” ones. Her think-aloud response gave no indication that she understood how the temporal context could affect the motivation of the authors. Nonetheless, in this case, Stephanie’s answer would give rise to the inference that she had indeed understood this key aspect of contextualization.

On three occasions a HAT written response completely misrepresented student thinking. Each of these three instances was in response to the Thank You Edison HAT. The item provided students with a letter from a woman in rural Kansas thanking Thomas Edison for the electrical
inventions that had improved her life dramatically. The assessment presented students with four correct facts and asked them to select the two that could help them determine whether this woman was “typical” of women in the 1920s. Two of facts showed that she was, in fact, *not* typical. The other two were irrelevant for determining typicality. Ten of the students in the protocol understood that the question allowed them to conclude that she was atypical. Three of 13, however, misunderstood the prompt. They thought that instead of asking which facts would help determine *if* Lathrop was typical, it was asking them which facts would help determine that Lathrop *was* typical. The verbal reports for these three students revealed that they understood how the two correct facts showed she was atypical of women at the time, but their written answers did not reflect this understanding. Because the multiple-choice version of the Thank You Edison assessment was stem-equivalent, the same confusion was observed in responses to both formats. The ambiguity in wording also led two students in the wrong direction on the multiple-choice item. Thus, the multiple-choice version of the item did not offer significant advantages over the HAT in terms of item clarity or alignment between proficiency and students’ written answers.

An even more common alignment issue arose from the differences between the two item formats. The multiple-choice items only allowed for binary inferences: right or wrong. This limited the accuracy of the inferences drawn from students’ answers. At the item level, a correct answer can only support the inference that the student was proficient in the targeted historical thinking process; an incorrect answer, that they were not. But in many instances, students’ thinking was somewhere in between. The verbal protocols of 31 of the 104 student responses to multiple-choice items showed shades of gray for which either/or inferences would distort the texture of student thinking.
James’s response to the Thanksgiving multiple-choice question exemplifies these issues:

So, choice A, *Agree, because it is impossible to know what the first Thanksgiving was like without seeing a picture of it*. I don’t think that’s true because there could be written accounts of it.

[Answer] B was *Agree, because the painting shows the Puritans and Wampanoag getting along peacefully in the spirit of Thanksgiving*. I think that’s true because you can see the Indians and setters they were joining together and eating Thanksgiving dinner together. And [answer] C was *Disagree, because the painting was made over 300 years after the first Thanksgiving*. That could actually also be true because it was so long after the first Thanksgiving that the facts could be confused. The painting might not exactly depict what exactly went on at the First Thanksgiving.

And so [answer] D, *Disagree, because it shows the relationship between the Puritans and Wampanoag as happier than it really was*. I don’t think that’s true because the Puritan settlers had a good relationship with the Indians there. So, I think it's B. Definitely not A or D.

The verbal report revealed that James engaged in the type of confirmation bias common among responses to the Thanksgiving item. He believed the source was good evidence because it lined up with what he believed about the past; this reasoning led him to select the incorrect answer. But James also revealed sound historical thinking when he first noticed the painting’s date and puzzled about it. He understood that the gap in time could limit the painting’s usefulness as evidence. James’s understanding was masked by his incorrect answer. This lack of alignment between James’ thinking and the inferences that would be drawn from his response highlights an inherent weakness of multiple-choice items: their failure to support inferences about gradations
in student understanding. The constructed-response format of the HAT, where students must explain their thinking in a short written response, provided more information about student thinking, which better reflected variations in student thinking.

**Discussion**

History Assessments of Thinking (HATs) were designed to meet the need for assessments of complex disciplinary processes in history. This study yielded evidence that informs a validity argument, with a particular focus on cognitive validity. We examined whether HATs tap the targeted historical thinking processes among a group of secondary history students. The results suggest that they do. Each HAT elicited thinking that reflected student proficiency in the targeted process. Further, students’ written answers closely resemble the thinking processes used to arrive at them. This alignment indicates students’ short written responses can be used to draw sound inferences about student thinking on key facets of historical thinking and supports an argument for their use.

The think-aloud protocols also revealed instances in which HATs elicited construct-irrelevant reasoning, as well as several instances of misalignment between student answers and student thinking. Two of the HATs elicited construct-irrelevant processes, and three written answers to a single HAT were misaligned with student thinking. These instances appear to have resulted from identifiable structural flaws in these particular items and were not problems pervasive among HATs more generally. These findings underscore the utility of think-aloud protocols in the assessment development process. Not only were the verbal protocols effective in revealing flaws, the data provided concrete and focused information that we used to subsequently revise the wording and structure of the items in question.
The data also suggested that stem-equivalent multiple-choice versions of HATs elicited historical thinking processes in some responses. This finding represents a departure from prior scholarship in the field. History education scholars have raised questions about the efficacy of traditional multiple-choice items for measuring complex disciplinary reasoning in history (e.g., VanSledright, 2014; Wineburg, 2004). Similarly, extant research has found that multiple-choice items currently used to measure achievement in history often do not elicit higher-order historical thinking processes—even when they are expressly designed to do so (cf. Reich, 2009; Smith 2017).

Why did the multiple-choice items in this investigation tap historical thinking processes in some responses? One possibility lies in the process we used to create them. Rather than start from scratch with vague notions about how we might measure historical thinking, we adapted tasks that directly targeted aspects of the measurement domain. The stem of each multiple-choice item closely matched the question posed by the parallel HAT and these prompts had already elicited evidence of historical thinking in extensive piloting. We then developed answer choices that reflected the types of responses that we saw from students when piloting HATs, a practice widely recommended by measurement scholars (e.g., Gierl, Bulut, Guo, & Zhang, 2017; Haladyna & Downing, 1989; Haladyna & Rodriguez, 2013; Rodriguez, 2011). The answer choices to the First Thanksgiving multiple-choice item (see Figure 3), for example, reflected the range of written answers we saw when the task was given to 793 students across the United States. The correct answer reflected sophisticated reasoning about the problematic gap in time between the painting and the event. The incorrect choices reflected common student misconceptions, like evaluating the painting based on whether the scene matched their understanding of the event and believing that a painting is a direct window to the past. So, unlike
traditional multiple-choice questions, which commonly include irrelevant “distractors” alongside the correct answer, our multiple-choice questions featured incorrect choices that reflected common lines of reasoning about historical sources that we observed in extensive piloting of HATs. Presenting common misconceptions alongside the correct answer prompted students to evaluate different ways of thinking about historical evidence, and in the process, engage in aspects of historical thinking.

Although stem-equivalent multiple-choice versions of HATs elicited aspects of historical thinking, there were differences between HATs and their multiple-choice counterparts in the types of thinking they elicited. The multiple-choice items were more likely to elicit construct-irrelevant processes than HATs. Each multiple-choice item was more likely to elicit processes of elimination and “best fit” reasoning than its HAT counterpart, and the fact that irrelevant processes were elicited by three-quarters of the multiple-choice items weakens the case for their use as measures of complex processes.

Multiple-choice items also prompted students to generate answers less frequently than HATs. Students came up with their own answers in only 30% of responses to multiple-choice items, compared to 98% of responses to HATs. Rather than reasoning independently about the trustworthiness and relevance of historical evidence, a hallmark of historical thinking, the multiple-choice questions prompted students to rely on the answer choices provided when evaluating the documents. Historical thinking requires the evaluation of evidence without the support of a bounded list of answer choices—and without the knowledge that one choice is correct. If we hope to draw inferences about student proficiency in historical thinking, we need assessments that require students to generate answers. Even if multiple-choice items prompt students to engage in some aspects of historical thinking, they may not approximate the full
complexity of historical thinking. History education scholars have called for history instruction that teaches students complex processes, like the ability to use evidence from a variety of perspectives to answer historical questions and to challenge historical claims (e.g., Körber, 2011, Seixas et al., 2013; Wineburg, 2001). It seems unlikely that sound conclusions about student proficiency in these ways of disciplinary reasoning can be drawn from multiple-choice items that do not require students to independently generate responses.

Student responses to multiple-choice items also provided less evidence of student thinking than HATs, and in 30% of responses, the student responses to the multiple-choice item masked gradations in student proficiency in aspects of historical thinking. This could limit their usefulness as tools for formative assessment, a practice widely recommended by education scholars (Black & Wiliam, 1998; Gordon Commission, 2013; Pellegrino et al., 2001; Stiggins, 2002). Effective formative assessment requires tools that provide classroom teachers with specific, actionable information about student needs. The results of this study provide reason to believe that HATs, which provided specific information about student thinking, may be better options for formative assessment than their multiple-choice counterparts.

**Limitations**

A limitation of the present work is the small sample size, which limits the generalizability of the findings. We generated a strategic sample that was not representative of all secondary students. To reduce the likelihood that the literacy demands of HATs would exceed students’ ability, we sampled students who had passed a university-level AP US history exam with high literacy demands. It is likely that the processes we elicited do not fully represent the types of cognitive processes that would be observed in a broader population of high school students. Research with a larger and more representative sample would help to identify threats to validity.
that were not uncovered with this sample. In particular, research with students with a broader range of literacy skills would inform our understanding of how HATs function. It would further bolster our understanding of the role of item format in measuring historical thinking.

We tailored HATs for secondary history classrooms in the United States. We did not seek to assess some aspects of historical thinking identified by international scholars. As a result, the findings of this study do not indicate whether HATs are effective for measuring these other aspects of historical thinking. For example, our findings tell us little about whether HAT-like items would be effective in measuring students’ ability to use information from the past to orient themselves in the present and future, something central to the FUER model of historical thinking proposed by German scholars (Körber, 2011; Körber & Meyer-Hamme, 2015). Additional research is needed to investigate whether HAT-like constructed-response items or their multiple-choice counterparts could be devised to tap constructs other than those included in our definition of historical thinking.

**Conclusion**

We are under no illusion that HATs can cure all that ails history assessment. Significant research is needed to find the best ways to measure historical thinking. We think our design process could help to inform future efforts to construct new assessments. Following best practices in assessment design, we clearly defined the domain, constructed tasks that targeted specific aspects of it, piloted assessments extensively with students, revised tasks, and re-piloted as necessary. Multiple-choice questions were then constructed based on the thinking revealed in student responses. Most importantly, we sought out evidence of cognitive validity by conducting think-aloud interviews with students. Regardless of the type of assessment researchers seek to
create, this design process may further spur the development of assessments that better tap aspects of historical thinking.

Item format will likely factor prominently in ongoing deliberations about how best to measure historical thinking. Multiple-choice items offer advantages in terms of cost and efficiency. Compared to constructed-response items, multiple-choice are reliable and inexpensive to score. However, validity is also a primary concern in decisions about item format. Both HATs and their stem-equivalent multiple-choice items elicited historical thinking, but the two item formats were not interchangeable. Constructed-response HATs elicited fewer instances of construct-irrelevant reasoning, and responses to HATs reflected student thinking better than their restricted-response counterparts. The use of multiple-choice versions of HATs is defensible only if test developers can account for the kinds of weaknesses that this study revealed. Multiple-choice items alone may not be able to provide a complete picture of student proficiency in the full range of historical thinking processes, and test developers may need to utilize a variety of formats when creating assessments of historical thinking.
References


https://doi.org/10.1080/00933104.2014.965860


https://doi.org/10.1037/h0046827


https://doi.org/10.1207/s15324818ame0201_4


https://doi.org/10.1207/s15324818ame1002_5


https://doi.org/10.1111/j.1745-3992.2006.00075.x


https://doi.org/10.1177/1086296X12450445


https://doi.org/10.1207/S1532690XCI2003_2


https://doi.org/10.1007/s10648-007-9056-1


Appendix

Codes that emerged from the first phase of Analysis 1

**Codes Derived from Definition of the Domain and Protocol Analysis**

**Historical Knowledge**

**Historical Knowledge – Periodization – Sequence:** Student places events, people, or movements in temporal order

**Historical Knowledge – Periodization – Location in Time:** Student shows general awareness of when events occurred or people lived

**Historical Knowledge – Narrative – Framework:** Student places historical information and evidence into the framework of a broad narrative of the past

**Historical Knowledge – Narrative – Connections:** Student draws connections between events and actors in the past

**Historical Knowledge – Narrative – Point of View:** Student explains how a perspective is subjective based on perspective

**Evaluation of Evidence**

**Evaluation of Evidence – Sourcing – Date of Creation:** Student evaluates historical evidence in light of the relationship in time between an event and evidence about the event

**Evaluation of Evidence – Sourcing – Perspective of Author:** Student evaluates the effect that the position of an author (place, affiliation) has on the historical evidence provided in a source

**Evaluation of Evidence – Sourcing – Interest/Motivation of Author:** Student reasons about the effect that personal motivation or interest of an author has on the evidence provided

**Evaluation of Evidence – Sourcing – Circumstances:** Student reasons about how the circumstances under which a document was created affect the information as evidence

**Evaluation of Evidence – Sourcing – Credibility of the Author:** Student reasons about how information about the character of the author affects interpretation of evidence provided by the author

**Evaluation of Evidence – Sourcing – Genre:** Student reasons about the effects of the type of source on the credibility of the evidence

**Evaluation of Evidence – Sourcing – Knowledge of Missing Info:** Student discerns what additional information would be helpful to evaluate the source

**Evaluation of Evidence – Contextualization – Socio-political context:** Student reasons about how socio-political environmental factors affect a document as evidence

**Evaluation of Evidence – Contextualization – Biographical Context:** Student reasons about how the biographical timing of a document affects the document as evidence

**Evaluation of Evidence – Corroboration – Comparison:** Student recognizes differences and similarities in various sources
Evaluation of Evidence – Corroboration – Verification: Student evaluates information in light of historical evidence from other sources.

Use of Evidence

Use of Evidence – Selecting Appropriate Evidence: Student identifies appropriate evidence for a claim and/or shows the ability to determine the relative strengths and weaknesses of different pieces of evidence

Use of Evidence – Sufficient Evidence: Student assesses whether evidence is sufficient to anchor a historical claim

Use of Evidence – Evaluating Claims: Student uses appropriate evidence to determine the quality of a historical claim

Construct Irrelevant Codes that Emerged from Protocol Analysis

Factual Recall/Recognition: Student engages in factual recall to arrive at an answer or recognizes factual information as correct

Process of Elimination: Student narrows problem space by eliminating answers that are not believed to be correct

Goodness of Fit: Student evaluates the relative veracity or plausibility of an answer vis-à-vis other possible answer choices

Guessing: Student randomly guesses an answer