History Assessments of Thinking: Design, Interpretation, and Implementation

Joel Breakstone

Recently there has been heightened interest in new types of assessments to measure student understanding. President Barack Obama (2009) called for "assessments that don't simply measure whether students can fill in a bubble on a test" (para. 21). The debut of the Common Core State Standards accelerated the movement for innovative forms of assessment to gauge students' mastery of higher order skills. High-stakes summative tests have received the lion's share of the funding, but classroom-level measures that provide teachers with useful feedback will be crucial to any new assessment system (Gewertz, 2011; Gordon, 2013). To achieve the goals set forth by the Common Core, teachers will need tools to monitor student understanding and to adjust instruction appropriately, a process known as formative assessment. This pedagogical practice provides teachers with information about student thinking and allows teachers to tailor instruction to ensure student progress. Formative assessment will be challenging in all subject areas, but it poses particular problems in history. The most readily available history tests neither lend themselves to frequent cycles of formative assessment nor yield detailed information about students' historical understanding. Moreover, little research has examined how history teachers use assessment data to inform instruction or how formative assessment transpires in the history classroom.

In an effort to provide teachers with more assessment options, my colleagues and I constructed, piloted, and revised new assessments. *History Assessments of Thinking* (HATs) target both historical content and historical thinking skills. HATs seek to measure disciplinary skills through engagement with primary sources. However, in order for HATs to positively influence teaching and learning, teachers must interpret student responses and enact appropriate curricular revisions, which requires pedagogical content knowledge (PCK) (Shulman, 1986).

This dissertation is comprised of three separate articles that address the design, interpretation, and implementation of HATs. The first article considers the assessment design process and identifies attributes of effective assessments. The second article examines the PCK required to interpret student responses to HATs and to formulate instruction in response. The final article investigates teachers' implementation of HATs in their classrooms. It is the first study to examine how history teachers engage in formative assessment.

Creating New History Assessments

Formative assessment options for history teachers are sparse. Two of the most readily available test item types, multiple-choice questions and document-based questions (DBQs), are poorly suited for formative assessment. Multiple-choice questions provide only darkened bubbles as evidence of student thinking, leaving teachers with little evidence on which to base future instruction (Haney & Scott, 1987; Madaus, Russell, & Higgins, 2009). Moreover, multiple-choice tests have been criticized for an emphasis on recognition and recall of facts rather than higher order aspects of historical thinking (Reich, 2009; Wineburg, 2004). On the other hand, DBQs require students to incorporate a series of documents into an analytic essay. Yet, given the realities of public schools, it is unrealistic to think of DBQs as tools for frequent classroom assessment. Teachers with classrooms crammed with students have to wade through hundreds of pages of student writing to determine next steps for teaching.

We created HATs with classroom assessment in mind. Each item was designed to allow teachers to gauge student understanding quickly with a task more complex than shading bubbles on a Scantron form but less time-consuming and complex than a DBQ. All of the assessments require students to analyze primary sources. The centrality of historical sources is no accident. Studying history without primary sources is like trying to study chemistry without the laboratory. It is only by engaging with these documents that students encounter the raw materials of the discipline. Moreover, without sources, students are left with the mediated narrative of the textbook, which obscures the author's interpretive stance (cf. Crismore, 1984). Instead of a history comprised of competing accounts, students receive a narrative that represents the past as fixed and indisputable.

As part of our development process, we explicitly defined components of historical thinking for the secondary history curriculum (see Appendix A) based upon a review of the existing literature (cf. Holt, 1990; Lévesque, 2008; Seixas & Peck, 2004; Wineburg, 1991). We then designed each HAT to address a particular dimension of the domain of historical thinking. For example, one of our assessments focuses on whether students can attend to the date a document was created. It asks students to decide whether an image of the first Thanksgiving published in 1932 would be a useful resource for historians seeking to understand the relationship between Pilgrim settlers and the Wampanoag in 1621 (see Appendix B). Students must identify in writing the limitations of a source created more than three centuries after the event it depicts. Their brief answers provide evidence for teachers to use to inform instruction.

As we set out to create new tasks, we thought testing companies might provide effective approaches to item construction. Unfortunately, they offered only broad guidelines. For example, the National Assessment of Educational Progress (2007) outlined their development process in detail online, but there was little information about what this process actually looked like in practice. In fact, we found scant guidance from any test developer. Testing outfits rarely revealed the details of item development. Moreover, they neither addressed the unique requirements of formative assessment nor the specific features of the history classroom. Researchers in science education have sought to provide practical advice for the development of science-specific assessments (e.g., Solano-Flores & Shavelson, 1997). Unfortunately, no such parallel exists in history education.

Formative Assessment

Formative assessment entails frequent assessment of student learning to identify gaps in understanding and to modify instruction in response. This data is used in a feedback loop in which teachers repeatedly gather information, revise instruction, and then collect new data about whether the gap has narrowed between students' current level of understanding and learning objectives (Heritage, 2007; Shavelson et al, 2008). This approach to assessment positively influences student learning. In a review of 250 studies, Black and Wiliam (1998) found formative assessment to be a uniquely powerful educational intervention. Unfortunately, the promise of formative assessment has rarely been realized in history classrooms. This is due, in part, to a lack of appropriate materials. History teachers, like teachers in most subject areas, rarely have access to curricular materials with embedded formative assessments (Herman, Osmundson, Ayala, Schneider, & Timms, 2006). Moreover, teachers will find little discipline-specific guidance regarding formative assessment because there is no body of research exploring how history teachers use formative assessments. However, several studies have examined the challenges teachers face when implementing formative assessment in other disciplines (Ayala et al., 2008; Frohbieter, Greenwald, Stecher, and Schwartz, 2011; Shavelson et al., 2008). If formative assessment is to improve teaching and learning in history classes, we must gain a better understanding of how it can be used in real classrooms.

Pedagogical Content Knowledge for Historical Inquiry

Formative assessment requires teachers to accurately interpret student responses and to take appropriate curricular action. This is no mean feat. For example, to use HATs, teachers must (1) understand the question, (2) be familiar with the historical content, (3) evaluate student responses, (4) diagnose student mistakes, (5) devise appropriate forms of remediation, and, finally, (6) implement them. Teachers need a specific type of pedagogical content knowledge to engage in this type of work. Specifically, teacher must possess two of the main components of PCK that Shulman (1986) identified: first, "the understanding of what makes the learning of specific topics easy or difficult: the conceptions and preconceptions that students of different ages and backgrounds bring with them to learning," and, second, "the ways of representing and formulating the subject that make it comprehensible to others" (p. 9). Researchers in other subject areas have been vigorously pursuing the work of mapping the terrain of discipline-specific PCK (Ball, Thames, & Phelps, 2008; Loughran, Mulhall, & Berry, 2004; Van Driel, Verloop, & de Vos, 1998). Studies have considered PCK in history classrooms, with varying degrees of specificity. Scholars have sought to identify aspects of PCK (Monte-Sano, 2011; Monte-Sano & Budano, 2013) and provided portraits of expert PCK in the history classroom (e.g., Bain, 2005; Bain, 2006). Despite this work, our understanding of the PCK teachers need to develop students' historical understanding will benefit from a more detailed examination of how teachers interpret students' historical thinking and how they use that information to inform future instruction.

Article I

The first article of the dissertation considered the principles of effective assessments that emerged from the design process for HATs. As we created HATs, we sought to gather information about their *cognitive validity* (cf. Ayala, Yin, Shavelson, & Vanides, 2002; Linn, Baker, & Dunbar, 1991; Pellegrino, Chudowsky, & Glaser, 2001), the relationship between the constructs targeted by the assessments and the cognitive processes students use to answer them. The validity of any new type of assessment depends upon a deliberate design process. The assessment community has provided guidelines for this endeavor. However, these instructions often belie the messy reality of producing worthy tasks. The article sought to provide something largely absent from other sources: a detailed analysis of how student data informed task revision and a description of the design principles that emerged from the process. Three case studies traced the development of different HATs through expert review, piloting with thousands of students across the United States, and extensive think-aloud interviews. This analysis revealed a series of principles of effective assessment design. Three of the principles are outlined below.

(1) Assessment structure must align with targeted constructs (Pellegrino, Chudowsky, & Glaser, 2001). In principle, this makes sense: ask about the desired construct. If you need directions to the airport, you don't ask someone how to get to the supermarket. By the same token, if teachers need information about students' ability to place a document in context, a question that requires them to write an essay is going to yield a lot of information extraneous to the evaluation of their grasp of contextualization. To disentangle students' writing ability from their historical knowledge we sought to identify assessment structures, like sentence starters, that decoupled compositional fluency from historical thinking.

(2) Pilot data is indispensable. Even the most carefully designed, rigorously reviewed, and theoretically strong prompts rarely worked as expected. Invariably, students' responses yielded clues to improve assessments. Reading through student responses helped to reveal the small tweaks, like the addition of a date, or dramatic restructurings, like entirely new documents, we needed to make. Sometimes it was a student's marginal notes and other times it was patterns

that emerged across dozens of answers that served as signposts toward better assessments. Regardless, data about how prompts functioned in real classrooms were critical for determining whether or not a task would yield accurate information about student thinking.

(3) Assessments must yield information about student thinking. Formative assessments must reveal student thinking. Formative assessment is predicated on an efficient feedback cycle in which teachers gather information about student understanding and revise instruction in response. This process depends upon teachers being able to quickly diagnose student understanding. Tasks that require teachers to slog through long student responses are antithetical to formative assessment.

Our experience suggests that a new generation of history assessments will require additional time and resources for development. For formative assessment to move beyond rhetoric and emerge as a regular part of history instruction, teachers need new tools for formative assessment. They must be carefully designed, rigorously field-tested, revised, and re-piloted. The work described here should inform this effort, but more teachers, researchers, and resources will have to be invested in the process for it to be successful.

Article II

This study examined the nature of the PCK required by teachers to interpret student responses to HATs. To consider influential factors in teachers' ability to make sense of student answers, teachers with differing levels of classroom experience and preparation for teaching with historical documents completed semi-structured interviews. The study addressed the following two questions: (1) What PCK do teachers need to interpret student responses to short, document-based assessments? (2) What factors contribute to teachers' ability to identify student thinking and to formulate appropriate curricular responses?

Eighteen social studies teachers with academic backgrounds in history participated. The teachers were placed into three groups based on their prior classroom teaching experience and their preparation and experience in teaching historical inquiry to students with primary sources. The three groups were: (1) Novice history teachers about to begin a teacher preparation program that emphasized teaching historical inquiry through document analysis; (2) experienced teachers who had taught for at least three years, but who had not received formal preparation for teaching historical inquiry; (3) experienced teachers who had taught for at least three years after graduating from the same teacher education program as the novice teachers and who had incorporated document analysis into their teaching practice.

Teachers answered an initial survey regarding their academic background, pedagogical style, teaching context, and teaching experience (National Assessment of Educational Progress, 2010). During semi-structured interviews, teachers examined sets of six student responses for two different assessments (see Appendix B). Participants examined the HATs, predicted how strong and struggling students would answer, considered sample student responses, looked for patterns in the responses, and ordered the answers based on strength. Next, participants described curricular interventions they would use to address problems in student understanding. Interviews were transcribed verbatim and coded using Dedoose, a web application for qualitative data analysis. Codes were developed based on existing models of historical thinking (cf. Ercikan & Seixas, 2011; Holt, 1990; Lévesque, 2008; Wineburg, 1991**)** and teachers' ability to interpret student responses. Codes addressed the following: teachers' understanding of the historical thinking construct addressed by the HAT; teachers' categorization of student responses; teachers' awareness of content that would challenge students; teachers' proposed curricular interventions. A second rater coded a sample transcript from each of the participant groups (i.e., novice teachers, experienced teachers without preparation, and experienced teachers with preparation). Inter-rater reliability was .91 (Cohen's κ).

Analysis of the interviews revealed that teachers with preparation and experience in teaching inquiry were better able to interpret student responses and suggest curricular interventions to improve student understanding. These teachers understood the challenges students faced in interpreting historical sources and knew specific pedagogical strategies to address student misconceptions. In contrast, neither content knowledge nor general teaching experience provided teachers with a similar understanding of the challenges students faced in interpreting primary sources or strategies for building students' historical understanding.

Although historical content knowledge allowed many participants to understand the HATs and their historical content, these results suggest that there were three main dimensions of PCK for historical inquiry:

(1) *Challenges of historical thinking*. Teachers with preparation and experience in teaching historical inquiry anticipated common student errors, such as when students overlooked the gap in time in the Thanksgiving HAT. These teachers also were able to identify the underlying causes of incorrect answers. For example, teachers with preparation and experience in teaching inquiry understood the challenges students faced when asked to explain how the authors of a 1936 play about the abolitionist John Brown might have been influenced by the historical context of the time (see Appendix B). They explained that even though students wrote down "1936" as the date of the play, they had not considered the significance of the date. One teacher noted,

> I ask this question a lot, “When was this play written?” And for so many students, it’s a quick check. They look for it in the document that I’ve given them, but they don’t think about it . . . students have to think, “What’s going on in that time period?”

This teacher knew that students might simply fill in the answer and not consider the context of the time period. He and the other teachers with PCK for historical inquiry had become familiar with common student errors and their underlying causes.

(2) *Awareness of how students’ historical thinking develops*. This familiarity with students’ developmental trajectory was on display when teachers described additional supports they would give students to interpret the Brown playbill. They realized that for students to complete this type of complex task independently, students first needed to have all of the intermediate steps spelled out. For example, rather than just ask students the playbill’s date, four of the experienced teachers with preparation suggested asking students what else was happening at the time of the play and how it might have motivated the authors. These teachers realized that sourcing was a prerequisite for contextualization. To place the playbill in context, students first had to consider its source. However, these tasks were not necessarily equivalent. As one teacher said, “[Contextualization] is a really hard skill for kids to get.” She knew that analyzing a document through the lens of the time in which it was produced was more challenging for students than noting its date. This awareness of the development of students’ historical thinking aided teachers in formulating appropriate curricular plans.

(3) *Grasp of pedagogical strategies to build students’ understanding*. Teachers with PCK for historical inquiry possessed specific strategies for developing students’ historical thinking. All of the experience teachers with preparation discussed the need to model their analysis of documents for students. They understood how important it was to make their expert reading

strategies visible to their students. This type of "cognitive modeling" (Collins, Brown, & Holum, 1991) is rare in the history classroom. Teachers do not normally think aloud to articulate their historical reading strategies. In contrast, teachers without experience or preparation in teaching historical inquiry struggled to understand how to help students. One teacher, who had a master's degree in history, said, "Maybe this is just something that you have to do over and over and over again, a skill you are constantly working on for the kids to get . . . I don't know." Without experience or explicit instruction in teaching historical inquiry, even teachers with substantial content knowledge and teaching experience struggled to formulate instructional plans. Familiarity with specific strategies for teaching historical thinking represented a key difference between content knowledge and pedagogical content knowledge. Most of the novices and experienced teachers without preparation accurately identified flaws in student answers, but they were unfamiliar with pedagogical strategies to improve students' historical thinking.

This study yielded fruitful information about PCK for historical inquiry. Explicit training and experience teaching students to interpret primary sources were the most important traits in shaping teachers' PCK for historical inquiry. Teachers with these characteristics were able to anticipate student errors and formulate appropriate responses. In contrast, historical content knowledge in isolation seemed to have much less impact on PCK for historical inquiry. Efforts to improve history instruction and to implement new assessments will depend on further mapping of PCK for historical inquiry and consideration of how these new assessments are used in practice. Such information will support the improvement of teacher preparation programs and professional development for experienced educators.

Article III

This design study explored how a group of three experienced high school teachers engaged in formative assessment as they implemented HATs in their classrooms. It addressed two main questions: (1) How do history teachers use new formative assessments in practice? (2) What role did professional collaboration play in the classroom implementation of these assessments? This group of teachers invited me to join them as they met to discuss inquiry-based instruction and assessment during the 2011-2012 school year. These teachers at a large, diverse urban high school in northern California had begun to use the Stanford History Education Group's *Reading Like a Historian* (RLH) curriculum (Reisman, 2012; Wineburg, Martin, & Monte-Sano, 2011) and intended to use HATs as common formative assessments. Such a scenario is exceedingly rare, even more so in history given the lack of available resources, which made this an ideal setting for design research (cf. Kelly, 2004) into how new educational tools, HATs, worked in actual classrooms. Design studies seek "to trace the evolution of learning in complex, messy classrooms and schools, test and build theories of teaching and learning, and produce instructional tools that survive the challenges of everyday practice" (Shavelson, Phillips, Towne, & Feuer, 2003, p. 25). To understand how HATs actually worked in classrooms, it was important to have a setting optimized for success. Consequently, the atypicality of the research site served to facilitate this study.

The data for this study came from a variety of sources. At the beginning of the school year, each teacher completed the same semi-structured, task-based interview as used in Article II. I was also a participated observer at nine monthly meetings in which the teachers discussed HATs and student responses. I audiotaped these meetings focused on examining new HATs, reading student responses to HATs, and discussing how best to use formative assessments in history classrooms. On three occasions I audiotaped individual conversations with all three teachers as they interpreted their students' responses to a common assessment. Teachers noted patterns in student responses and discussed next instructional steps. Subsequently, I videotaped the teachers as they reviewed student responses to the assessments in class. All of these data were transcribed verbatim. Codes based on existing theories of historical thinking and formative assessment were applied in Dedoose. Codes addressed the following: How teachers discussed formative assessments with colleagues; how teachers used formative assessments; and how teachers interpreted student responses. The varied data sources allowed for triangulation of data (cf. Erickson, 1986) and a more complete picture of how history teachers engaged in formative assessment. A second rater coded a sample transcript from each of the data sources (e.g., teacher review of responses, in-class use of HATs, etc.). Inter-rater reliability was .91 (Cohen's κ).

An analysis of these data revealed various barriers to formative assessment in history classrooms. Curricular misalignment, ingrained notions of summative assessment, and the feedback demands of formative assessment all represented obstacles to implementation. Despite these challenges, teachers used HATs to introduce aspects of historical thinking, monitor student understanding, and broach broader historical topics.

At the beginning of the year, all the teachers expressed frustration with assessment. Their unhappiness stemmed, in part, from the fact that assessment and grading were inextricably linked. Assessment meant grades. As a result, grades were a constant presence. Group meetings were initially focused on the generation of scoring rubrics. Similarly, as teachers went through students' responses on their own, determining the appropriate score was often the main focus.

Despite their emphasis on grades, the teachers often deployed assessments in creative and generative ways. Assessments served as introductions to historical thinking, tools to identify gaps in students' understanding, and entry points for deeper historical conversations. For example, at the beginning of the year, the teachers introduced historical thinking with assessments. Teachers used the First Thanksgiving HAT to familiarize students with the need to consider source information during the analysis of historical documents. The teachers modeled how they would answer the prompt and explained that consideration of a document's source information, or "sourcing," is a crucial step in document analysis (cf. Wineburg, 1991). One teacher displayed the assessment on her interactive whiteboard and articulated her thinking as she analyzed the document. She told students that she first considered the attribution information. This became a mantra in her class. Whenever they analyzed a document, the teacher would ask, "What's the first thing we do when we look at a document?" The class would respond in unison, "Source!" In this way, the HAT introduced an aspect of historical thinking and provided an opportunity to establish practices of document analysis.

On the whole, these teachers' work together presented a portrait of the possible. They shifted from merely grading assessments to using them to improve instruction. Instead of right/wrong evaluations of student work, they used HATs to explore with students the ambiguity of historical evidence. Teachers' questions about how to grade evolved into questions about *should* they grade. In the process, they reconsidered their deeply ingrained practice of scoring everything that students submitted. In their classrooms, HATs evolved from assessments *of* learning to assessments *for* learning.

In light of the discouraging research to date, these developments are reasons for optimism. Yet, even here, where conditions were stacked in favor of success, serious challenges remained. The obstacles these teachers encountered will no doubt be magnified if formative assessments are used in typical classrooms. Even for these teachers, the transition to formative assessment was not easy. It required them to reconsider how they worked with students and how they provided feedback. The “summative assessment teaching script” (Ayala et al., 2008, p. 322) that the teachers carried with them strongly influenced how they interacted with new formative assessments. The perceived need to score assessments and enter marks into a gradebook seemed so ingrained in their teaching practice that it might be considered part of the “grammar of schooling” (Tyack & Tobin, 1994). Teachers seemed to get caught up in what might be called a *culture of classification,* in which the first goal, even before understanding the thinking behind students’ responses, was to classify them. Teachers’ rush toward rubrics sometimes impeded reflection on the student thinking that responses revealed. Rather than serve as lenses that allowed them to see student thinking at a greater resolution, HATs, sometimes, became new opportunities to assign numerical grades. A formative approach to assessment also created new demands on teachers. On a practical level, the teachers frequently referred to the difficulty of giving timely feedback on dozens of student responses.

This study provided a glimpse of what formative history assessment could look like in practice. Teachers used HATs to introduce new aspects of historical thinking, to build shared understanding, and to enter into conversations about key aspects of historical evidence. But these results were achieved under special conditions that would be difficult to replicate. And there were still obstacles. Teachers were not accustomed to this type of assessment. Assessments based on historical thinking constructs represented a radical departure from the status quo in history classrooms. For formative assessment to become routine in classrooms beyond a select few, other teachers will need explicit instruction and additional supports. This will require sustained investment in history curricula with embedded assessments and associated professional development for teachers at all stages of their careers. History assessment might then serve as a catalyst for students to meet the rigorous requirements of the Common Core and the 21st century.

References

Ayala, C.C., Shavelson, R.J., Ruiz-Primo, M.A., Brandon, P.R., Yin, Y., Furtak, M.E., Young, D.B., & Tomita, M.K. (2008). From formal embedded assessments to reflective lessons: The development of formative assessment studies. *Applied Measurement in Education, 21*(4), 315-334.

Ayala, C. C., Yin, Y., Shavelson, R. J., & Vanides, J. (2002, April). *Investigating the cognitive validity of science performance assessment with think alouds: Technical aspects*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans: LA.

Bain, R.B. (2005). 'They thought the world was flat?' Applying the principles of *How People Learn* in teaching high school history. In J. Bransford and S. Donovan (Eds.), *How Students Learn History, Mathematics and Science in the Classroom* (pp. 179-214). Washington, D.C.: The National Academies Press.

Bain, R.B. (2006). Rounding up unusual suspects: Facing the authority hidden in the history classroom. *Teachers College Record, 108*(10), 2080-2114.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389-407.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, & Practice, 5*(1), 7-73.

Collins, A., Brown, J. S., & Holum, A. (1991). Cognitive apprenticeship: Making thinking visible. *American Educator*, *6*(11), 38-46.

Common Core State Standards Initiative. (2010). Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects.

Retrieved from http://www.corestandards.org/the-standards

Crismore, A. (1984). The rhetoric of textbooks: Metadiscourse. *Journal of Curriculum Studies, 16*(3), 279-296.

Ercikan, K., & Seixas, P. (2011). Assessment of higher order thinking: The case of historical thinking. In G. Scraw (Ed.), *Assessment of higher order thinking skills* (pp. 245-261). Scottsdale, AZ: Information Age Publishing.

Erickson, F. (1986). Qualitative methods in researchon teaching. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 119-161). New York: Macmillan.

Frohbieter, G., Greenwald, E., Stecher, B., & Schwartz, H. (2011). *Knowing and doing: What teachers learn from formative assessment and how they use the information* (CRESST report 802). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Kelly, A. (2004). Design research in education: Yes, but is it methodological? *The Journal of the Learning Sciences*, *13*(1), 115-128.

Gewertz, C. (2011, February 23). Common-assessment consortia add resources to plans. *Education Week, 30*(21), 8.

Gordon, E. W. (2013). *A public policy statement*. The Gordon Commission on the Future of Assessment in Education. Retrieved from http://gordoncommission.org/rsc/pdfs/gordon_commission_public_policy_report.pdf

Haney, W., & Scott, L. (1987). Talking with children about tests: An exploratory study of test ambiguity. In R. Freedle (Ed.), *Cognitive and linguistic analysis of test performance* (pp. 69-87). Norwood, NJ: Ablex Publishing.

Heritage, M. (2007). Formative assessment: What do teachers need to know? *Phi Delta Kappan,* *89*(10), 140-145.

Herman, J., Osmundson, E., Ayala, C., Schneider, S., & Timms, M. (2006). *The nature and impact of teachers' formative assessment practices* (CSE Technical Report 703). Los Angeles: Center for the Study of Evaluation.

Holt, T. (1990). *Thinking historically: Narrative, imagination, and understanding*. New York: College Entrance Examination Board.

Lévesque, S. (2008). *Thinking historically: Educating students for the twenty-first century*. Toronto: University of Toronto Press.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15.

Loughran, J., Mulhall, P. & Bery, A. (2004). In search of pedagogical content knowledge in science: Developing ways of articulating and documenting professional practice. *Journal of Research in Science Teaching, 41*(4), 370-391.

Madaus, G., Russell, M., & Higgins, J. (2009). *The Paradoxes of High Stakes Testing*. Charlotte: Information Age Publishing.

Monte-Sano, C. (2011). Learning to open up history for students: Preservice teachers' emerging pedagogical content knowledge. *Journal of Teacher Education, 62*(3), 260-272.

Monte-Sano, C. & Budano, C. (2013). Developing and enacting pedagogical content knowledge for teaching history: An exploration of two novice teachers' growth over three years. *Journal of the Learning Sciences, 22*(2), 171-211.

National Assessment of Educational Progress (2007, January 26). NAEP item development process. Retrieved from http://nces.ed.gov/nationsreportcard/contracts/item_dev.aspx

National Assessment of Educational Progress. (2010). Civics, geography, & U.S. history teacher background questionnaire. Retrieved from http://nces.ed.gov/nationsreportcard/bgquest.asp

Obama, B. (2009, March 10). Remarks by the President to the Hispanic Chamber of Commerce on a complete and competitive American Education. The White House: Office of the Press Secretary. Retrieved from http://www.whitehouse.gov/the_press_office/Remarks-of-the-President-to-the-United-States-Hispanic-Chamber-of-Commerce/

Pellegrino, J.W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know.* Washington, DC: National Academy Press.

Reich, G. (2009). Testing historical knowledge: Standards, multiple-choice questions and student reasoning. *Theory and Research in Social Education*, *37*(3), 325-360.

Reisman, A. (2012). Reading like a historian: A document-based history curriculum intervention in urban high schools. *Cognition and Instruction, 33*(1), 86-112.

Seixas, P., & Peck, C. (2004). Teaching historical thinking. In A. Sears & I. Wright (Eds.), *Challenges and prospects for Canadian social studies* (pp. 109-117). Vancouver: Pacific Education Press.

Shavelson, R.J., Phillips, D.C., Towne, L., & Feuer, M.J. (2003). On the science of education design studies. *Educational Researcher, 32*(1), pp. 25-28.

Shavelson, R.J., Young, D.B., Ayala, C.C., Brandon, P., Furtak, E.M., Ruiz-Primo, M.A., Tomita, M., & Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education, 21*(4), 295-314.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational*

Researcher, 15(2), 4-14.

Solano-Flores, G. & Shavelson, S. (1997). Development of performance assessments in science: conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice*, *16*(3), 16-24.

Tyack, D., & Tobin, W. (1994). The "grammar" of schooling: Why has it been so hard to change? *American Educational Research Journal*, *31*(3), 453-479.

Van Driel, J.H., Verloop, N., & De Vos, W. (1998). Developing science teachers' pedagogical content knowledge. *Journal of Research in Science Teaching, 35*(6), 673-695.

Van Hover, S. D. (2006). Teaching history in the old dominion: The impact of Virginia's accountability reform on seven secondary beginning history teachers. In S. G. Grant (Ed.), *Measuring history: Cases of state-level testing across the United States* (pp. 195–219). Greenwich, CT: Information Age Publishing.

Vogler, K. (2008). Comparing the impact of accountability examinations on Mississippi and Tennessee social studies teachers' instructional practices. *Educational Assessment*, *32*(1), 1-32.

Wineburg, S. (1991). On the reading of historical texts: Notes on the breach between school and academy. *American Educational Research Journal*, *28*(3), 495-520.

Wineburg, S. (2004). Crazy for history. *Journal of American History*, *90*(4), 1401–1414.

Wineburg, S., Martin, D., & Monte-Sano, C. (2011). *Reading like a historian: Teaching literacy in middle and high school history classrooms*. New York: Teachers College Press.

Appendix A

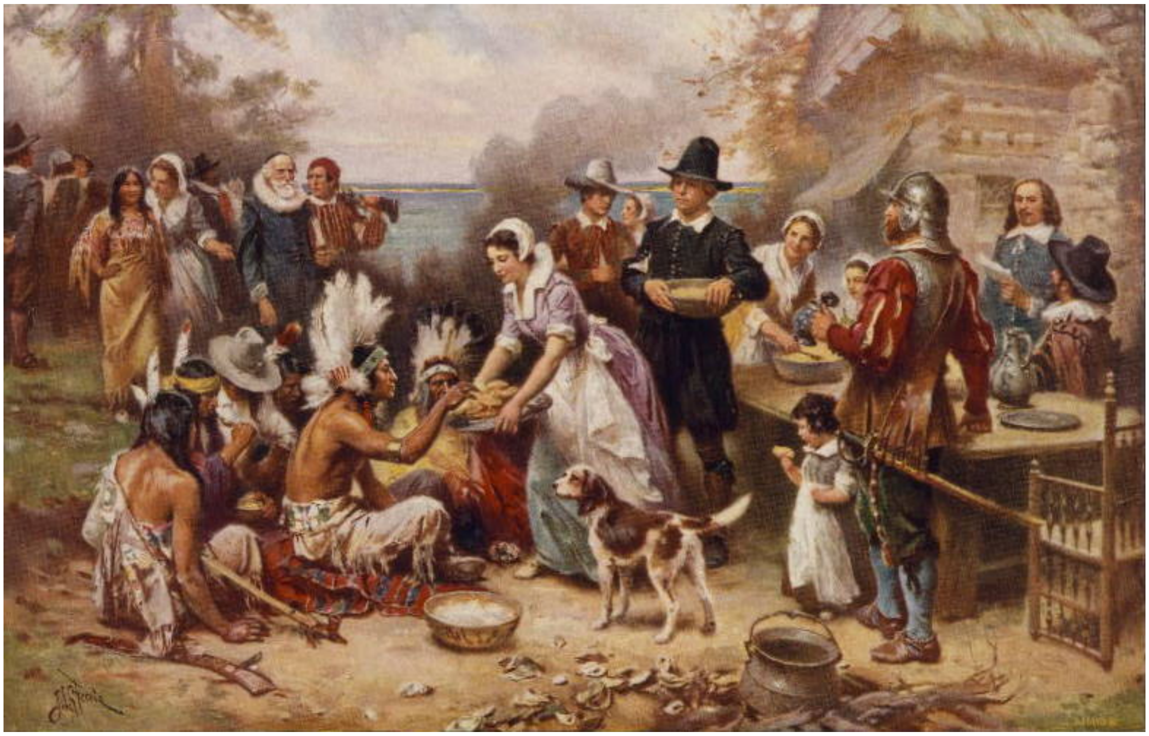
Domain of historical thinking for the secondary history curriculum

Sub-construct	**Aspects**	**Facets**
Historical knowledge	Significance	Consequential, exemplar, and point of view
	Periodization	Grouping, sequence, and location in time
	Narrative	Framework, connections, an point of view
	Historical information	Recall, recognition, and evaluation of fact
Evaluation of evidence	Sourcing	Date, perspective of author, interest/motivation of author, circumstances, credibility of author, genre, and knowledge of missing information
	Corroboration	Comparison, verification, and articulation of need
	Contextualization	Socio-political, biographical, context of entire document, intellectual, environmental/geo-spatial, zeitgeist, and linguistic
Use of evidence/ argumentation	Claims	Legitimate question, generalization, causality, counterfactual, and comparison
	Evidence	Selecting appropriate evidence, sufficient evidence, and evaluating claims
	Coherence	Evidence follows claim, appropriate evidence for claim, and address counter-argument

Appendix B

History Assessments of Thinking

Directions: Use the painting to answer the question below.



Title: "The First Thanksgiving 1621"
By: J.L.G. Ferris
Date published: 1932

Statement:

The painting, "The First Thanksgiving 1621," helps historians understand the relationship between the Wampanoag Indians and the Pilgrim settlers in 1621.

Question:

Do you agree or disagree? (Circle one).

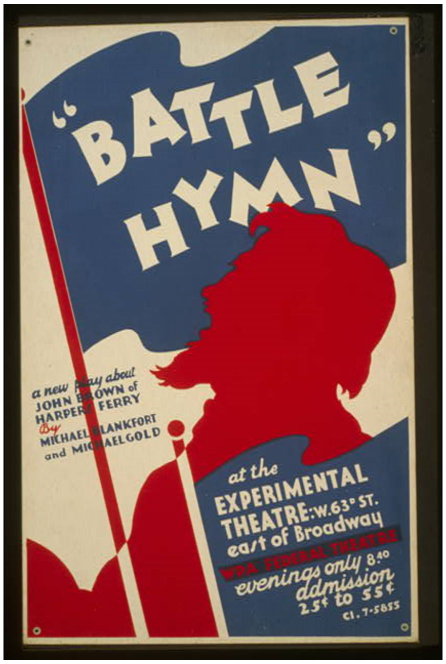
Briefly support your answer:

____

____

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 1936 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?

1. Slaves made up nearly 40% of Virginia's population in 1859.
2. One of the play's authors, Michael Gold, was a member of the Communist Party, which protested against lynching in the1930s.
3. After taking power in 1933, Adolf Hitler enacted racist policies in Germany.
4. After seceding from the Union in 1861, Virginia became the largest state in the Confederacy and the home of its capital, Richmond.

Fact # _______ might help explain why the authors wrote this play because __________________

____

____

Fact # _______ might help explain why the authors wrote this play because __________________

____

____