Stanford University makes this peer-reviewed final draft available under a Creative Commons Attribution-Noncommercial License. The published version is available from the publisher, subscribing libraries, and the author.

This is a prepublication version of the article. For the final printed version see
Wineburg, S., Smith, M., & Breakstone, J. (2012). New directions in assessment: Using Library of Congress sources to assess historical understanding. *Social Education, 76*(6), 288-291.

New Directions in Assessment:

Using *Library of Congress* Sources to Assess Historical Understanding

(~2,500 words)

Sam Wineburg, Mark Smith, and Joel Breakstone

Stanford University

September 18, 2012

wineburg@stanford.edu

650-725-4411

Stanford University makes this peer-reviewed final draft available under a Creative Commons Attribution-Noncommercial License. The published version is available from the publisher, subscribing libraries, and the author.

Educators could learn a lot about assessment from their cars' navigation systems. Sure, the constant barking of directions can sometimes prompt us to throw these contraptions out the window. But often, we begrudgingly admit, our GPS provides us with quick and accurate feedback. When we miss a turn, it even gives us with a roadmap to get back on track.

In education, we would be well served if our assessments -- and the feedback offered on their basis -- approached the precision of our cars' navigation systems. Accurate information about student understanding is crucial if we are to adapt instruction to meet students' needs, especially in light of the new Common Core State Standards. According to the Common Core, students in history/social studies are expected to analyze primary and secondary sources, cite textual evidence to support arguments, consider the influence of an author's perspective, corroborate multiple sources, and develop written arguments. In short, students need to learn how to think.

Current tests are hardly up to the task. Consider the most commonly used gauge**,** the multiple-choice test. What kind of information does it provide about students' thinking?

Walt Haney from Boston College and Laurie Scott from the Huron Institute used alarmingly simple means to explore this question. They gave items from nationally-normed tests to a group of 12-year-olds but asked kids to explain out loud how they chose their answers. One question depicted three different plants -- a cactus in a pot, another potted houseplant, and a head of cabbage. It read: "Which plant requires the least amount of water?" Most students picked the cactus, the "keyed" answer. But one girl insisted on the cabbage. Using irrefutable logic, she argued that the two potted plants needed watering. But, in her mind, the cabbage already rested comfortably in the fridge's crisper and needed water only to "clean it" (1987, p. 33).

Stanford University makes this peer-reviewed final draft available under a Creative Commons Attribution-Noncommercial License. The published version is available from the publisher, subscribing libraries, and the author.

Our cabbage picker's shrewd reasoning illustrates the perils of relying exclusively on multiple-choice questions to assess what students know. Teachers who bother to listen to students after giving a multiple-choice test know the sinking feeling of hearing them reveal how they got the wrong answer for the right reason or how they managed to get the right answer for the wrong reason.

Blackened circles on a Scantron can mean any number of things: that students have guessed correctly, that they have been swayed by a "distracter" that can be justified with novel reasoning, that they have arrived at the right answer by taking a faulty route, or something else entirely. Without learning more about what students actually think, it's hard to know what feedback to give them.

Multiple-choice tests have another downside. They encourage memorization in an era when students should be learning to interpret and analyze. Sometimes the result is so absurd it's hard to know if to laugh or cry. Consider this standard from California's 1998 History/Social Science Framework that asks students "to analyze how change happens at different rates at different times … and affects not only technology and politics but also values and beliefs" (California State Department of Education, 1998, p. 177; see also Fogo 2011; Wineburg, 2005). An item on the state exam that purportedly measured this standard asked students to pick which of four groups "had the greatest increase in membership due to increasing immigration from Latin American countries": Catholics, Jews, Muslims, or Protestants (California State Department of Education, 2009, p. 13). So much for analysis.

At the other end of the spectrum sits what many consider to be the gold standard of history testing, the Advanced Placement Program's document-based question. Widely known by its acronym, the "DBQ" asks students to read ten to twelve documents and use them to compose

Stanford University makes this peer-reviewed final draft available under a Creative Commons Attribution-Noncommercial License. The published version is available from the publisher, subscribing libraries, and the author.

an analytic essay. To its credit, the DBQ parallels many of the tasks students face in college. However, as an assessment tool, it's hard to know what the DBQ actually measures. Is it students' ability to engage in historical thinking and arrive at a defensible thesis? Their ability to sort through and organize disparate documents? Or their ability to express themselves in writing while wiping beads of sweat from their brows under timed conditions?[1]

In one of the few studies to examine how students approached the DBQ, Katherine McCarthy Young and Gaea Leinhardt (1998) found that students often "raided" documents for appropriate quotes but failed to analyze them as historical sources. Yet, even if students deftly analyzed documents and presented their thoughts in essay form, we'd still be left with the harsh realities of teaching. If public school teachers were to give weekly DBQs and provide students with the comments needed to hone their thinking and compositional skill, they'd have to cut their nightly sleep to two hours and give up their weekends. DBQs are not designed to provide quick, diagnostic feedback. And, lest we forget: students first need to learn how to analyze *one* document before they can effectively analyze twelve. Where are the assessments for measuring *that*?

History Assessments of Thinking

Research has shown that formative assessment is a key ingredient in raising student achievement (Shavelson, et al., 2008). The goal of formative assessment is not to grade students, but to pinpoint where they are having trouble and then to take appropriate instructional action. In a meta-analysis of 250 studies, Black and Wiliam (1998) found that formative assessments had a more significant effect on student achievement than practically any other classroom innovation. However, effective formative assessments must provide insight into student thinking, something

Stanford University makes this peer-reviewed final draft available under a Creative Commons Attribution-Noncommercial License. The published version is available from the publisher, subscribing libraries, and the author.

multiple-choice tests don't do very well, as well as allowing for quick evaluation, the Achilles' heel of DBQs. Designing such assessments was our challenge and our opportunity.

With support from the Library of Congress Teaching with Primary Sources Program, we set out to marshal the resources of the digital revolution to create new history assessments. How could we use the Library's vast collection of documents, photos, paintings, speeches, radio broadcasts and film clips to help teachers track students' growth as thinkers?

In partnership with the San Francisco (California) Unified School District and the Lincoln (Nebraska) Public Schools, our research group has spent nearly three years constructing, piloting, and revising assessments that provide social studies teachers with new options. We call our items *History Assessments of Thinking*, or *HATs*.

HATs fill the void between the recall of discrete facts on a multiple-choice question and the complex orchestration of skills required by a document-based essay. We have piloted dozens of HATs with thousands of students in California, Nebraska, Florida, Iowa, Wisconsin, and as far away as Singapore, while also conducting validity studies in which we ask students to tell us what they are thinking as they complete our assessments.[2] The results have been promising. Evidence suggests that our items do indeed tap important aspects of historical thinking. Just as promising is feedback from teachers, who report that HATs give them the kind of information they need to make adjustments in their teaching.

Each HAT prompts students to answer questions about historical sources and to justify their reasoning in two to three sentences. Most HATs can be completed in ten minutes, some in less than five. Even in a classroom with 37 students, a teacher can get a quick sense of what students do and don't know by skimming a batch of responses.

Stanford University makes this peer-reviewed final draft available under a Creative Commons Attribution-Noncommercial License. The published version is available from the publisher, subscribing libraries, and the author.

Consider an assessment that students can finish in five minutes. This question taps a core historical understanding: how and when a document was created must be considered when judging its value as evidence. Students are presented with a painting of the first Thanksgiving from 1932, and asked to decide whether it would be useful to historians who want to understand the relationship between settlers and the Wampanoag in 1621.[3] A 311-year gap separates illustration from event. Yet many students ignore this information entirely. Rather than considering the effect of three intervening centuries, ample time for distortions, myths, and legends to seep into collective memory (cf. Siskind, 1992), students often fixate on the painting's details, skipping over its attribution.

In reviewing hundreds of responses, we've identified common patterns in student reasoning. Many take the painting at face value: if their understanding of Thanksgiving accords with the illustration, the source is deemed useful. In a twist on the saying "A picture is worth a thousand words," one middle school student wrote: "You can see how they are interacting with each other. Without any picture, you couldn't really see how Wampanoag Indians and the Pilgrims acted." Similarly, a 12th grader wrote, "The First Thanksgiving 1621 shows how the Pilgrim settlers are sharing their food with the Wampanoag Indians. The sharing of food with the Indians shows that no matter the social class or the color of skin, sharing is possible and should be prevalent today."

Other students looked at the source more critically but ignored the attribution just the same. One wrote, "As soon as the settlers arrived, there was mass curiosity which turned into violence and hatred. There was never such a 'party' between the two peoples. They couldn't even understand each other." This student has brought prior knowledge to his evaluation. However, while clearly critical, this response does not, in our opinion, constitute *critical*

Stanford University makes this peer-reviewed final draft available under a Creative Commons Attribution-Noncommercial License. The published version is available from the publisher, subscribing libraries, and the author.

thinking. Like the other student, this one also engages in "matching," comparing the image to his prior beliefs about the event, and making factual errors in the process (What happened to Squanto and his role as interpreter? What about William Bradford's 1639 account of a festive meal?).

Although many students struggled, others composed responses that sparkled with historical nuance. Consider this response by an 11th grader: "This painting was drawn 311 years after the actual event happened. There is no evidence of historical accuracy, as we do not know if the artist did research before painting this, or if he just drew what is a stereotypical Pilgrim and Indian painting." Other answers invoked the need to compare the painting with diaries from the time period. One response even suggested that such a strategy would be hampered by the dearth of native sources.

To our surprise, we have not seen a consistent developmental trend. What matter most, it seems, is whether students have been taught the skills of documentary evaluation, not their age. The following response by a 6th grader, while awkward in syntax, displays a more sophisticated understanding than many responses from 11th and 12th graders: "The painting is not a showing of how the Pilgrims and Indians reacted to each other. The painting was made in 1932. J.L.G. Ferris would not know how they reacted to each other in 1621. The Indians and the Pilgrims could have fought. J.L.G. Ferris has no proof this is true."

In contrast to the blackened circles of multiple-choice forms, students' short written responses provide teachers with rich information. Strong answers indicate that students have grasped this dimension of evaluating historical evidence. Less developed responses also point teachers in specific directions. In both cases, teachers have a clearer sense of where to go to improve their students' thinking.

Stanford University makes this peer-reviewed final draft available under a Creative Commons Attribution-Noncommercial License. The published version is available from the publisher, subscribing libraries, and the author.

Assembling HATs

Our assessments are designed to measure historical understanding from multiple vantage points. Assessments like the Thanksgiving exercise ask students to evaluate the reliability of historical evidence. Others focus on whether students can use evidence to mount a historical argument. Still others require students to evaluate the historical significance of particular images, such as the one that shows the raised fists of John Carlos and Tommie Smith at the Mexico City Olympics. Rather than ask students merely to identify the picture, our assessment asks them to connect the gesture to the social upheavals that rocked America during the 1960s. There are also HATs that require students to put events into context. One exercise presents students with Dorothea Lange's iconic "Migrant Mother" photo, and asks how her employment by the Resettlement Administration, with its goal of drumming up support for FDR's programs, might affect their evaluation of the image.

In a website created to disseminate HATs (beyondthebubble.stanford.edu), we include sample student responses and an easy-to-use, three-level rubric. We've kept it simple because teachers are busy people. If HATs are to catch on, our tools have to be efficient and user-friendly.

The Future of History Testing

Bemoaning not only the state of history testing but assessment in general, the psychometrician Robert Mislevy noted, "It is only a slight exaggeration to describe the test theory that dominates educational measurement today as the application of 20th century statistics to 19th century psychology" (1993, p. 19). To be sure, our HATs don't solve all of the problems of modern testing. But our hope, at least with respect to social studies, is that HATs will give

Stanford University makes this peer-reviewed final draft available under a Creative Commons Attribution-Noncommercial License. The published version is available from the publisher, subscribing libraries, and the author.

teachers new tools to nurture the development of historical understanding. With the adoption of the Common Core State Standards and efforts to create new tests, we hope that HATs might spur efforts to go beyond discrete multiple-choice tests, on one hand, and full-blown DBQs, on the other. At present, these two options virtually exhaust the range of history testing even though countless other options fall between these two poles. Only a stubborn resistance to change prevents us from finding them.

References

Black, P. & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), *139-144.*

California State Department of Education. (1998). *History-social science content standards for California public schools*. Sacramento, CA: California State Department of Education.

California Department of Education. (2009). *California Standards Test: Released test questions U.S. History.* Sacramento, CA: California State Department of Education.

Fogo, B. (2011). Making and measuring the California history standards. *Phil Delta Kappan, 92*(8), 62-67.

Mislevy, R. (1993). Foundations of a new test theory. In N. Frederiksen, R. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19-31). Hillsdale, NJ: Erlbaum.

Reisman, A. (2011). The "Document-Based Lesson:" Bringing disciplinary inquiry into high school history classrooms with adolescent struggling readers. *Journal of Curriculum Studies.* DOI:10.1080/00220272.2011.591436.

Reisman, A. (2012). "Reading like a historian:" A document-based history curriculum

Stanford University makes this peer-reviewed final draft available under a Creative Commons Attribution-Noncommercial License. The published version is available from the publisher, subscribing libraries, and the author.

intervention in urban high schools. *Cognition and Instruction, 30*(1), 86-112.

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.

Siskind, J. (1992). The invention of Thanksgiving: A ritual of American nationality. *Critique of Anthropology, 12*(2), 167–191.

Wainer, H. (2011). *Uneducated guesses: Using evidence to uncover misguided education policies.* Princeton: Princeton University Press.

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 classrooms*. New York: Teachers College Press.

Young, K. M., & Leinhardt, G. (1998). Writing from primary documents: A way of knowing in history. *Written Communication, 15*(1) 25-68.

[1]Even trained scorers have great difficulty agreeing how to score the DBQ. After decades of administering the Advanced Placement "Free Response" questions, reliability estimates rarely exceed .5. See Wainer, 2011.

[2] Our work is also generously supported by the Bill and Melinda Gates Foundation. However, no endorsement of the views expressed here should be inferred from this support. More information about the "Reading Like a Historian"

Stanford University makes this peer-reviewed final draft available under a Creative Commons Attribution-Noncommercial License. The published version is available from the publisher, subscribing libraries, and the author.

approach can be found in Wineburg, Martin, and Monte-Sano (2011) and at http://sheg.stanford.edu. For effectiveness studies in urban schools, see Reisman (2011, 2012).

[3] http://www.loc.gov/pictures/item/2001699850/

New Directions in Assessment: Using Library of Congress Sources to Assess Historical Understanding. Wineburg, Sam; Smith, Mark; Breakstone, Joel. Social Education , v76 n6 p290-293 Nov-Dec 2012. Research has shown that formative assessment is a key ingredient in raising student achievement. The goal of formative assessment is not to grade students, but to pinpoint where they are having trouble and then to take appropriate instructional action.Â National Council for the Social Studies. 8555 Sixteenth Street #500, Silver Spring, MD 20910. Tel: 800-683-0812; Tel: 301-588-1800; Fax: 301-588-2049; e-mail: membership@ncss.org; Web site: http://www.socialstudies.org. Background: Assessment of the change in tumour burden is an important feature of the clinical evaluation of cancer therapeutics: both tumour shrinkage (objective response) and disease progression are useful endpoints in clinical trials.Â HIGHLIGHTS OF REVISED RECIST 1.1: Major changes include: Number of lesions to be assessed: based on evidence from numerous trial databases merged into a data warehouse for analysis purposes, the number of lesions required to assess tumour burden for response determination has been reduced from a maximum of 10 to a maximum of five total (and from five to two per organÂ Finally, a section on detection of new lesions, including the interpretation of FDG-PET scan assessment is included. Sam Wineburg's 43 research works with 2,345 citations and 10,151 reads, including: Studentsâ€™ Civic Online Reasoning: A National Portrait.Â Learners who use the Internet must be able to assess the credibility and trustworthiness of sources and information (McGrew et al., 2018; Wineburg et al., 2018) , they have to balance new information against their prior knowledge and any beliefs they may hold (van Strien et al., 2014;Alexander, 2017, 2018), and they must recognize how a given text or media format. Library of Congress, the de facto national library of the United States and the largest library in the world. The Library of Congress serves members, committees, and staff of the U.S. Congress, other government agencies, libraries throughout the country and the world, and the scholars who use its resources.Â Its collection was growing at a rate of about two million items per year; it reached more than 155 million items in 2012. The Library of Congress serves members, committees, and staff of the U.S. Congress, other government agencies, libraries throughout the country and the world, and the scholars, researchers, artists, and scientists who use its resources.