New York State K-12 Social Studies Toolkit

Assessment White Paper

S. G. Grant
Kathy Swan
John Lee

Social studies and its constituent disciplinary fields (e.g., civics, economics, geography, and history) have faced a rocky road of late. Effectively ignored by the No Child Left Behind legislation, social studies then seemed coopted as a minor part of the Common Core English-Language Arts standards. Over the years, there have been many reasons for the apparent marginalization of social studies—some legitimate, others not—but the outcome was the same: Social studies appeared to be withering on the academic vine.

It is withering no more. Through a range of state and national efforts, social studies is rebounding across the United States. The efforts germane to this paper are the *College, Career, and Civic Life (C3) Framework for State Social Studies Standards,* the *New York State K-12 Social Studies Framework*, and the New York State K-12 Social Studies Toolkit and Professional Development project.

None of these efforts, however, will transform social studies teaching and learning by themselves (Grant, 2003, 2006). Teaching and learning are complex activities and so, while new standards and curriculum can help teachers engage their students, equally important are the assessments that follow. Few observers would argue that assessment alone should drive teaching and learning. But if assessments are not coherent with the standards and curricular directions expressed, then the chance for real and robust reform diminishes (Darling-Hammond & Adamson, 2014).

Assessments have improved over time, especially in the domain of reliability. That is, little doubt now exists that results will vary significantly if students are given repeated versions of the same tasks. The problem is on the validity side of testing: We simply have not consistently developed assessments that inspire confidence that they measure anything of importance (Darling-Hammond & Adamson, 2014; Grant, 2007; Grant & Salinas, 2008; Stiggins, 2014; Supovitz, 2009). Traditional forms of assessment (multiple-choice questions, extended-response tasks) have their purposes and yield a high degree of psychometric reliability. But they reinforce the image that subjects like social studies are but a laundry list of people, places, and events. Moreover, multiple-choice questions, in particular, too often seem designed to trick students rather than accurately assess what they know and can do. Some scholars are trying to redesign multiple-choice questions to better probe students' abilities (Au, 2009; VanSledright, 2013). That effort alone, however, is unlikely to win the trust of those for whom testing matters—students, teachers, parents, policymakers, and the public. Thus the question of how we know what students know, especially with a high degree of confidence, remains problematic.

As part of the Memorandum of Agreement between the New York State Education Department and Binghamton University, Commissioner John King asked that a white paper be developed to "help inform the development of future state assessments in social studies." To honor that request, we offer the following analysis of and proposal for the assessment of social studies in New York State.

The proposal calls for the development of an assessment *program* that incorporates, but extends beyond the current Regents testing program in social studies. More specifically, we argue for a set of performance-based tasks at the elementary and middle school that would be required, but would give local district educators considerable discretion in how these assessment tasks were constructed, scored, and interpreted.

The Challenge of Knowing What Students Know

Odd as it may seem, understanding what children know about history or social studies has proven illusive. Over time, educators and psychometricians have constructed many ways of assessing what children know under the assumption that there is no one best assessment. Offering multiple assessments, however, fails to solve the problem of deciding what children's representations on these assessments mean.

Researchers have identified a number of problems that undercut our ability to know, with a strong degree of confidence, what students know. One problem is that if students are asked different kinds of questions about the same topic, their responses may suggest that they may know very different things (Grant, 2007; Levstik & Barton, 1997; Nuthall & Alton-Lee, 1995; Rogers, & Stevenson, 1988; VanSledright, Kelly, & Meuwissen, 2006). A second problem is that students may appear to know different things over even a relatively short period of time (Rogers & Stevenson, 1988; VanSledright, Kelly, & Meuwissen, 2006). A third problem is the challenge of gaining consensus across evaluators; different scorers may evaluate the same student's response very differently (Baker, 1994). And finally, students' scores on tests, whether classroom or large scale, may be variously interpreted (Cizek, 2009; Horn, 2003, 2006). Taken together, the research literature suggests that both teachers and researchers struggle to answer the question "how do we know what students know?"

The two most common approaches we have to understand what students know are classroom-based and large-scale assessments. As umbrella terms, each of these labels covers a wide terrain. Still, looking at the opportunities and constraints each offers can be instructive.

Classroom-Based vs. Large-Scale Assessments

Assessment is as much a part of the teaching and learning experience as are planning, instruction, resources, and white boards. Teachers teach, students learn; students take tests and teachers grade them. Still, knowing what students know, with any degree of confidence, proves problematic. Classroom assessments garner less criticism than standardized tests and for good reason: Typically, classroom measures are more varied in type and are more directly tied to the taught curriculum than are large-scale assessments. That said, the problems that plague high-stakes tests share sufficient common ground with classroom-based measures such that the very idea of assessment is challenging regardless of the form it

takes. As a field, we have spent considerable time talking about content, teaching, and learning, but we struggle with the question of how to know what students know.

Classroom-based assessment. Classroom-based assessments would seem to offer several benefits as a means of understanding what students know. First, teacher-developed tasks offer more and more varied opportunities for students to show what they know and can do. With occasions to monitor students individually and as a class, through formal and informal tasks, and with a full palette of possibilities, teachers can use classroom-based assessments toward myriad ends. Although multiple approaches to assessing students offer no guarantee of a single and convergent insight, with more assessment opportunities come more occasions to build a more robust understanding of students' capabilities (Cizek, 2009).

A second benefit to classroom-based assessments is the potential for close ties to the taught curriculum (Allen, Ort, & Schmidt, 2009; Brophy & Alleman, 1998). Validity, as we will see, comes in many forms, but the idea that the teacher who is teaching a class is the same one designing an ensuing assessment offers a content coherence that large-scale test designers can only imagine. A third opportunity that classroom-based assessment offers is the ability to chart students' growth over time. Large-scale exams can advance useful snapshots of individual and group-level achievement. Looking across a set of students' performance on classroom-based tasks, however, provides the opportunity to more clearly and coherently gauge students' development (Allen, Ort, & Schmidt, 2009). A final benefit to classroom-based assessments, and an advantage they have over their standardized cousins, is in terms of the speed with which tasks are scored and results are returned to students. Learning involves relearning and one of the big disadvantages of large-scale testing is the general lack of feedback on students' ongoing performance. Grading students' work and returning it with corrective feedback serves as many educative functions as it does evaluative (Black & William, 1998, 2009; Stiggins, 1998; Wineburg, Smith, & Breakstone, 2012).

The benefits of classroom-based assessment are several, but they do not obviate a set of challenges. Those challenges begin with the unevenness in design and presentation that describes the world of classroom-based assessments. The value of teacher-designed evaluations begins to evaporate when one realizes just how wide and uneven the array of assessment designs is (Banks, 2005; Bonner, 2013; Wiggins, 1998). The sheer variety of such tasks underscores two problems that, while typically associated with large-scale testing, yield similar challenges to classroom-based assessments.

The first of these problems is reliability. Generally defined as the dependability of results (Cizak, 2009; Parkes, 2013), reliability is based on the premise that subsequent administrations of the same or very similar tasks will yield the same or very similar results, in large-scale testing situations, reliability is typically achieved by giving the same exam to demographically and academically matched sets of students. If their final scores closely align, then the exam is judged to be reliable (Horn, 2006; Joint Committee on Standards for Educational and Psychological Testing, 2014; Messick, 1989).[1] While generally not a difficulty for large-scale assessments, reliability is a significant problem for classroom assessments (Parkes, 2013). First, classroom teachers rarely engage in the kind of test-retest procedure that psychometricians embrace to build reliability measures, but they often ask

[1] Tests can also be reliable over time if the results of similarly-designed tests produce consistent scores.

students about the same ideas in a variety of ways. Nuthall & Alton-Lee (1995) and Rogers and Stevenson (1988) observe, however, that, when asked about the same topic through different assessment approaches, students can appear to know different things. A second challenge on the reliability front concerns the scoring of students' work. Though many school children have a pet story about a teacher who seemed to score their papers harder than their peers' work, intrarater consistency (i.e., the manner in which a single teacher evaluates her or his students' performances) is less an issue than is inter-rater agreement (Brookhart, 2013; Cizek, 2009; Haladyna, 2009). In a study of how teachers grade DBQ-like responses, Baker (1994) challenges the notion that student performances can be objectively and consistency scored:

Four history teachers, handpicked on the basis of their excellent teaching reputations, shared neither explicit nor implicit sets of criteria to judge the quality of students' history understanding. Judging 85 written explanations, they came to little agreement. (p. 100)

The fact that classroom teachers rarely have their assessments scored by colleagues does not obviate Baker's point: Knowing what students know is as dependent on who is making the judgment as it is on the students' performance itself.[2]

The second challenge to classroom assessments revolves around validity. There are many forms of validity, but the construct generally refers to the agreement between an assessment measure and the behavior it is meant to represent and the "degree to which conclusions based on test information are on target" (Cizek, 2009, p. 68).[3] Though validity is perhaps most associated with high-stakes testing, there is reason to think about it in classrooms as well.

The most common validity concern in classroom assessments surrounds the alignment between the taught and the tested curriculum (Cizek, 2009). Every school child knows the anxiety of seeing a set of questions that look utterly unlike the teaching that preceded them. Other elements of validity are more subtle. For example, identifying lists of prominent people, places, and events is a social studies assessment staple, yet how close a match is there between the results of such as task and engaging in an active civic life? Knowing something about the social world, both past and present, makes sense as an attribute of citizenship, but what something and to what degree? Is it enough to know the dates of the Civil War and that Abraham Lincoln was President at the time? Or must students know that multiple causes are attributed to that event and that the importance of them individually and together is still a topic of considerable debate? The kinds of assessment tasks assigned are a consideration, but validity demands that we make judgments about the fidelity of those tasks to a sense of what matters in and outside of the classroom. We know far less about teachers' and students' experiences around assessment than we do any other phase of classroom activity. That such would be the case is all the more surprising when some observers claim

---

[2] In fact, Baker (1994) found more agreement on students' scores between English and history teachers than she did across the history teachers in the sample.

[3] While most observers view validity and reliability as distinct constructs, Haladyna (2006) argues that reliability functions as a form of validity in that, if high-stakes tests are viewed as unreliable, then confidence in the assessments as valid measures is undermined.

that classroom-based assessments yield far better data about students than do their large-scale peers (Cizek, 2009; Ducker & Perlstein, 2014).

Large-scale assessment. As educators, we have traditionally put aside concerns about the reliability of students' responses on and teachers' scoring of classroom-based assessments. We have long known that not all classroom measures are equally rigorous. And we largely have ignored the fact that, although classroom-based assessments more directly evaluate students' knowledge of the taught curriculum, the question of whether those assessments measure something of value remains (Supovitz, 2009): Passing social studies tests has not always translated into confidence in the value of the courses students take (Epstein, 1994, 2008; Schug, Todd, & Beery, 1984). But until the relatively recent rise in importance of large-scale testing, teachers, students, and the public seemed to have pursued a mutually-agreed upon non-aggression pact regarding classroom-based assessment: The part of "doing school" that includes classroom-based assessment as a way of knowing what students know has gone unchallenged (Bonner, 2013; Cizek, 2009).

For many years, the use of large-scale assessments also went unchallenged (Darling-Hammond & Adamson, 2014; Stiggins, 2014). Large-scale assessments are not new to schooling in general or to social studies in particular; New York students, for example, have taken state-level, Regents history exams for over 100 years. Elevating the anxiety about large-scale testing has been the introduction of high-stakes accountability.

Large-scale and high-stakes tests (we will use these terms interchangeably from here on) have garnered considerable attention, though less in the case of social studies than in literacy and mathematics (Au, 2007; Grant, 2006; Grant & Salinas, 2008). More attention, however, has not produced more confidence in the results. In effect, we have more testing and less certainty than ever (Darling-Hammond & Adamson, 2014; Koretz, 2008; Ravitch, 2011, 2014; Stiggins, 2014; Supovitz, 2009).

Any judgment about the benefits of large-scale assessments, then, begins with the presumption that those benefits will accrue once we have assessments that marshal a high degree of confidence among students, teachers, and the public (Cizek, 2009). Standardized assessments presumably have practical evaluative and sorting functions that help educators make thoughtful decisions about students' school experiences. Increasingly, however, large-scale testing programs exist within a public sphere that weighs the value of those assessments on multiple levels. No longer is it only teachers and students who worry about the amount of time needed for standardized testing, the uncertain match to the taught curriculum, and the overall value to schooling in general. The current calculus around large-scale testing appears to be more doubtful than it has in the past: If anything, the climate around large-scale testing is more challenging than ever (Chatterji, 2014; Klein, 2014; Ravitch, 2014; Stiggins, 2014; Yuan & Le, 2012).

The roll-out of the Common Core State Standards has been accompanied by a general uncertainty about standards and testing across the United States; the roll-out of the attendant standardized tests has produced outcry in states like New York, Florida, Massachusetts, and elsewhere (Aloise, Longhurst, & Platin, 2014; Alvarez, 2014). Since No Child Left Behind, the public's appetite for testing has waned such that some parents are pushing state-level

policymakers to scale back the number of tests they administer (Brody, 2014; Bushaw & Lopez, 2013; Ravitch, 2014; Yuan & Le, 2012).

Although the backlash against Common Core-aligned testing is new, it recalls some of the long-standing critiques of large-scale assessment. A short list of those critiques includes the idea that a) testing offers a patina of objectivity that masks a range of subjective (and too often obfuscating) decisions (Breakstone, Smith, & Wineburg, 2013; Kohn, 2000; Martin, Maldonado, Schneider, & Smith, 2011); b) high-stakes testing promotes a new form of discrimination against poor and minority students (Au, 2009; McNeil, 2000; Ravitch, 2014); c) testing can lead to the narrowing of the curriculum, as in the case of elementary social studies where instructional time has decreased significantly (von Zastrow & Janc, 2004; Fitchett & Heafner, 2010; Wills, 2007); and d) tests inspire a kind of defensiveness in teachers whereby they sacrifice teaching time to test preparation (McNeil, 2000; Ravitch, 2011; Vogler, 2006). Chatterji (2014) summarizes the many problems evident in the use of large-scale tests:

Educators, parents, and local officials reasonably fear that, yet again, tests are serving as blunt policy instruments to drive top-down reforms with inadequate time and resources for designing deeper curriculum and assessments to match, with little or no professional development of teachers and school leaders and in neglect of critical supports that schools need to succeed. (p. 30)

The size and scope of large-scale testing provides psychometricians with the necessary conditions to craft assessments that meet the standard criteria for reliability (Cizek, 2009). The issues of intra- and inter-rater reliability that arise in classroom-based assessments can be magnified on scale evident in a standardized test, but the controls evident around such assessments mitigate much of the possibility for error. Creating assessments that meet specific statistical requirements, however, is not the same thing as creating assessments that inspire confidence that they are measuring something worth measuring.

The Special Problem of Validity

In the context of large-scale assessment, Darling-Hammond (1991) highlights the problem of validity: High-stakes testing "cannot be a constructive lever for reform unless we invest in more educationally useful and valid measures of student learning" (p. 229). She and other observers point to the notion that, despite our best collective efforts, our assessments continue to present validity problems.

Though there are many kinds of validity issues (Horn, 2006; Messick, 1989), three focus on the relationship between the goals for social studies education and the instruments developed to measure them—content, construct, and face validity. The first two are much described and embraced by psychometricians; the third, we argue, may end up being the most important to the other stakeholders in American education—teachers, parents, students, and the public.

Content validity represents the extent to which items on a test sufficiently represent a domain of knowledge or skill such that a set of reasonable inferences can be drawn. Construct validity, by contrast, highlights the extent to which a test is judged to accurately

and adequately measure student knowledge and skill around a particular idea or concept. Put simply—a social studies test would have content validity if it was determined to yield a set of inferences that match the overall goals of a set of social studies standards; it would have construct validity if the questions or tasks were judged to be useful proxies for a particular construct such as the American Revolution. Content and construct validity questions can be answered in a number of ways, but the default approach is to have content and pedagogical experts review draft questions and come to agreement on those that reflect a set of standards.

That process for validating tests seems appropriate, but it masks two problems. First, is the issue of goals. Horn (2006) notes that an endemic concern for those who would assess social studies is the fact that no clear, concise, and generally embraced set of goals exists. Standards documents abound, but they are rarely bounded in ways that provide a coherent sense of guidance to test developers. The generally expressed goal of social studies—preparation of good citizens—is vague at best (VanSledright & Grant, 1994) and it competes for attention with a number of other goals passionately held by members of the field (e.g., transmitting knowledge, improving human relations, transforming societies). If we focus on citizenship education, presumably it means a particular set of knowledge, skills, dispositions, and behaviors. But what particular knowledge and at what depth? What kinds of skills and how are they represented? And can we really develop large-scale test proxies for dispositions and behaviors? With such an ill-defined field, test developers are left with little in the way of content guidance.

The second problem is the issue of face validity. Although generally ignored by psychometricians (Nevo, 1985; Secolsky, 1987), face validity may present the biggest threat to large-scale assessment.

Face validity is defined as "the suitability of the content of a test or item(s) for an intended purpose as perceived by test takers, users, and/or the general public" (Secolsky, 1987, p. 82). What seems to bother many psychometricians about face validity is its seemingly casual nature and the idea that it is a judgment rendered by "technically untrained observers" (Anastasi, 1988, p.144). Clearly, it is important to have subject matter and test design experts involved in the validation of tests, particularly those that hold high stakes for teachers and students. But the new landscape of large-scale assessment means that experts are no longer the only voices in the debate. Instead, validity has become a political issue, one in which politicians and the general public join with teachers, parents, and students as stakeholders (Alvarez, 2014; Macken-Horarik, 2011; Ravitch, 2014). Nevo (1985) argues that, among other things face validity is a key feature of any test that is "intended for a practical use" (p. 288) as it is more likely to inspire cooperation and effort among test takers, reduce dissatisfaction with low scores, and improve public support for testing.

Expert evaluations of content and construct validity are no less important today than they have been in the past. Yet the general concern over too much testing and the particular disquiet over the Common Core-inspired tests has produced a volatile situation with regard to assessment (Ravitch, 2014). For better or worse, non-expert groups have joined the conversation around the nature and use of large-scale tests. Those groups bring many anxieties to bear, but among them is the challenge to the face validity of tests administered

(e.g., Casner-Lotto, 2006; Gorin, 2013. The question remains: How do we as educators know what students know with any degree of confidence?

Promising Developments

Much has been written about doing away with large-scale assessments (e.g., Hagopian & Ravitch, 2014; Jones & Jones, 2003; Nichols & Berliner, 2014). These arguments may have gained more traction of late but, in fact, the current policy debate is less about eliminating such exams than it is about reducing the overall amount of testing and seeking acceptable alternatives (Heitin, 2014). Echoing a national trend, the decision to reduce the amount of social studies testing in New York has already been made: The grades 5 and 8 exams were eliminated in 2010. The decision to revise the grades 10 and 11 Regents exams reflects the idea that large-scale assessments can change. We now turn to possible alternatives to standard test items.

The first development, Weighted Multiple Choice, concerns the manner in which multiple-choice items are constructed and scored. The second, Beyond the Bubble, focuses on the construction of short-answer questions. The third development, the new Advanced Placement exams in history,[4] has implications for multiple-choice, short-answer, and extended-response items. Finally, the Washington State program of Classroom-Based Assessments suggests a different approach to extended-response tasks.

Weighted Multiple Choice

Proponents of Weighted Multiple Choice (WMC) items (e.g., VanSledright, 2013)[5] accept multiple- choice questions as a viable approach to measuring students' social studies knowledge and skills. Rather than testing only for single right answers, however, WMC items are constructed with several possible answers. Those answers are weighted according to defensibility, that is, to how well each answer fits with current historical evidence. The advantage of this approach is that a) it does not penalize students for not knowing a single piece of history (i.e., the single right answer), and b) it offers more opportunities for students to weigh answers against one another as they make their choices. An example with analysis follows:

> Historians argue that the "Boston Massacre" was most likely the result of
>
> A. British soldiers firing into a crowd of angry citizens (0)
> B. A crowd of citizens show taunted and threatened British soldiers (2)
> C. Lack of control by British soldiers who fired on innocent citizens (1)
> D. British soldiers panicked when confronted with potential violence (4)
>
> Explanation:

[4] United States history was the first AP course to be revised; the European and World history courses are slated for change based on the revisions to the US course of study.

[5] VanSledright has been the primary advocate for Weighted Multiple Choice items in social studies. Scholars in other school subjects, however, have made similar arguments (e.g., Lin & Singh, 2013; Modu & Wimmer, 1981).

> Historical evidence from the trial of the British soldiers indicates that "D" is clearly the best response. While "B" has some credibility, it does not address the issues raised the testimony of the soldiers who describe a confusing and terrifying scenario. Response "C," while not without merit, discounts most of the historical evidence, which implies that the crowd was far from acting as "innocent civilians." In addition, it portrays the soldiers as lacking in control or discipline, rather than terrified by an angry, armed mob. Although "A" indicates the crowd was angry, an incident can cause itself. http://www.umbc.edu/che/arch/weighted.php

Haladyna (2004) argues that WMC are challenging to construct and difficult to explain to stakeholders. Still, the idea that multiple-choice items might more accurately portray what students know is enticing. With the redesign of the New York State Regents testing program, opportunities to expand the variety and usefulness of multiple-choice questions seems worth exploring.

Beyond the Bubble

In the space between multiple-choice and extended-response questions, the Beyond the Bubble project highlights the idea that assessment items can be crafted that are as quick to score as multiple choice, but offer more direct avenues to understand what students know about history and historical thinking. Although they are designed for classroom use, the history-based assessments, which focus on the evaluation of evidence, historical knowledge, and historical argumentation, could be adapted for use on standardized assessments.

In an example of the Beyond the Bubble approach, students are shown an oil painting created in 1932 by J. L. G. Ferris entitled, *The First Thanksgiving 1621*. They are then asked to write a short response that answers the question of how useful this source would be to historians trying to understand the relationship between the Wampanoag and the Pilgrims in 1621 (https://beyondthebubble.stanford.edu/assessments/first-thanksgiving).

This particular assessment looks at students' abilities to draw inferences about a historical source. In this example, the key idea is whether students understand the relationship between when an event happened and when the source was created (i.e., the 300 year difference may limit the usefulness of this particular source). Rather than test students' content knowledge only, Beyond the Bubble assessments call on students to apply their content knowledge in ways that go beyond simple recall.

The idea of testing more than students' content knowledge and doing so in efficient ways could be a useful avenue to explore in the redesign of the New York State Regents program.

Advanced Placement Exams

The Advanced Placement (AP) program in history has long combined multiple-choice items with extended-response tasks in the form of Document-Based Questions. Revisions to the multiple-choice items in the AP history exams as well as the inclusion of short answer and the two types of extended-response tasks offer interesting possibilities for rethinking the ways in which revisions to large-scale tests like the Regents program might unfold

(http://media.collegeboard.com/digitalServices/pdf/ap/ap-us-history-course-and-exam-description.pdf).

Rather than a series of discrete, content-focused items, the multiple-choice questions in the AP history exams will be composed in sets of two to five questions designed around a featured primary or secondary source. Students' content knowledge is still assessed, but it is done in the context of items that also examine their historical thinking skills.

For example, an excerpt from the English Navigation Act of 1696 on the use of English ships and crews in carrying imports and exports is followed by three multiple-choice questions which assess students' ability to determine contextualization, causation, and comparison. A second set of three questions are built around two quotations—one from South Carolina governor James Henry Hammond and the second from Frederick Douglass—again with an eye toward testing the ways that students use the sources to apply the skills of contextualization and comparison.

The individual Short Answer questions on the AP history exams seem more traditionally content focused than the Beyond the Bubble examples. Taken together, however, they examine a range of historical thinking skills. The following example tests for the construct of continuity and change over time:

1. Answer a, b, and c.

a) Briefly explain ONE example of how contact between Native Americans and Europeans brought changes to Native American societies in the period 1492 to 1700.
b) Briefly explain a SECOND example of how contact between Native Americans and Europeans brought changes to Native American societies in the same period.
c) Briefly explain ONE example of how Native American societies resisted changes brought by contact with Europeans in the same period.

Students need to have the appropriate content knowledge to answer the questions, but they may draw from a wide range of historical instances in order to craft their responses. Such short-answer questions may or may not have a source (e.g., charts, maps, historians' arguments) as a prompt.

The Document-Based Question (DBQ), a staple of the AP history exams, continues in the latest version. As an extended-response task, the DBQ offers opportunities to look more deeply into students' capacity to demonstrate their historical thinking skills (e.g., periodization, interpretation, synthesis, in addition to those noted above).

The DBQ sets a task for students, presents them with a set of related documents, and asks them to craft an argument or thesis with supporting evidence. An example of a task aimed at evaluating students' ability to address continuity and change over time follows:

Analyze major changes and continuities in the social and economic experiences of African Americans who migrated from the rural South to urban areas in the North in the period 1910–1930.

As with the short-answer questions, students are free to draw on a wide range of content in their responses. So, while the DBQ must be answered with accurate information, it does not privilege a particular set of ideas. At the same time, however, students' responses can be judged on the quality of their application of one or more historical thinking skills.

The Long Essay, an apparent revision of the previous Free-Response question, takes a more directive approach to the task of making and supporting an argument. Again, students are expected to demonstrate one or more historical thinking skills, but the Long Essay asks them to support, modify, or refute an interpretation of an historical issue:

Some historians have argued that the American Revolution was not revolutionary in nature. Support, modify, or refute this interpretation, providing specific evidence to justify your answer.

Unlike the DBQ, which assesses students' ability to craft a coherent thesis statement and then build and support a convincing argument based on the sources supplied and their outside knowledge, the Long Essay focuses more intently on students' capacity to use evidence. Whether they directly support, modify, or refute the historian's interpretation is less the concern than is students' effort to construct coherent support for their thesis statements.

Given its design as full-figured standardized test, the new AP exam offers some intriguing possibilities to consider in the redevelopment of the New York State Regents program. While remaining true to the outward form of multiple-choice, short-answer, and extended-response items, the designers of the new AP test are experimenting with item design that appears to push more deeply into the realm of historical thinking.

Washington State Classroom-Based Assessments

In some ways, the Washington State effort is perhaps the most radical example of a new approach to testing. A state-sponsored test is mandated, but the assessment a) is classroom-based, b) offers teachers some choice as to which assessment they offer their students, and c) and is more authentic than most state assessments (http://www.k12.wa.us/SocialStudies/Assessments/default.aspx).

The Classroom-Based Assessment (CBA) program features assessments that cover all social studies areas—civics, economics, geography, history, and international perspectives—but only the civics CBA is required at the elementary, middle, and high school levels. Two or more CBA options are offered at each level: The elementary options include, for example, "What's the Big Idea?" (history) and "You Decide" (civics). The latter asks students to a) identify a problem and a policy or law that attempts to solve it, b) to explain one way the policy or law attempts to solve the problem, c) to identify individuals and/or groups who participated in the policy or law-making process and explain how they did so, and d) to provide reasons for agreeing or disagreeing with the law or policy by explaining how the law or policy promotes a right or democratic ideal with one or more supporting details. Students who participate in this assessment have the option to write a paper or develop a presentation that "someone outside their classroom could easily understand and review using the rubric"

(http://www.k12.wa.us/SocialStudies/Assessments/Elementary/ElemCivics-WhoseRules-CBA.pdf).

Three features define the Washington State CBAs as distinct. First, the assessment is classroom based. Rather than testing all grade-level students at the same time, the CBA is administered by classroom teachers on a day and time of their choosing. The anxiety that students manifest around high-stakes exams is not eliminated, but it is likely lessened. The second distinctive feature is that the CBA program is rooted in the idea of teacher autonomy. By giving teachers some choice as to which tests their students take, teachers may feel more invested in the validity of the test as well as the import of their students' scores. Finally, the CBA is much closer to the idea of an authentic exam than are most tests. By setting a relatively genuine task (i.e., making and supporting an argument about a current policy), the CBA task creates a more realistic environment in which students can display their social studies knowledge and skills. Taking this tact may not solve the problem of face validity, but those inside and outside of schools are likely to see it as more coherent with real-world contexts.

It is unlikely that the Washington State model could be embraced with regard to the Regents testing program, which is well established and largely embraced. The CBA approach might, however, be a model for creating a broader approach to assessment in New York State. We explore that possibility below.

* * * * *

This brief overview of some of the more promising national developments in assessment practice could prove useful as revision of the Regents testing program continues. Alternative forms of multiple-choice, short-answer, and extended-response items could offer opportunities to probe more deeply students' social studies knowledge and skills. Moreover, if the idea of expanding the scope of social studies assessment in New York State is a possibility, then these developments could suggest options for creating more authentic assessments.

A Proposal for a New York State Assessment Program in Social Studies

The roll-out of the Common Core State Standards has produced a general backlash across the United States; the roll-out of the attendant standardized tests has produced the phenomena of parents pulling their children from testing (Aloise, Longhurst, & Platin, 2014; Alvarez, 2014). Since No Child Left Behind, the public's appetite for testing has waned (Bushaw & Lopez, 2013; Ravitch, 2014; Yuan & Le, 2012) such that parents and others are pushing state-level policymakers to scale back the number of tests they administer.

High-stakes testing is not the death star it is sometimes portrayed to be. The size and scope of large-scale testing provides psychometricians with the necessary conditions to craft assessments that meet the standard criteria for reliability (Cizek, 2009). The issues of intra- and inter-rater reliability that arise in classroom-based assessments can be magnified in a standardized test, but the controls evident around such assessments mitigate much of the

possibility for such errors. In some ways, validity issues, too, can be managed more effectively on large-scale tests (Cizek, 2009) than they can on classroom-based tests.

Yet creating assessments that meet a psychometrician's definition of validity is not the same thing as creating assessments that inspire confidence among stakeholders that they are measuring something worth measuring. Although there may be many ways to rebuild stakeholder confidence in large-scale testing, one avenue may be to break the exclusive connection between *test* and *assessment*. Often used as synonyms, these terms are quite different: Test refers to a single, systematic sample of a person's knowledge and/or skill; assessment refers to the broad approach to and array of evaluative tasks (Cizek, 2009). The former can be included in the latter, but if the latter is defined only by the former, then the possibilities for deeply understanding what students know and can do fade and thus the decline of public support for large-scale evaluation is likely to continue.

Test weariness need not equal assessment weariness. But with large-scale testing increasingly under fire, we argue that the time is now to develop an assessment *program* that incorporates and goes beyond the current Regents testing program in social studies.

Six assumptions undergird our argument:

- Social studies is more than history—all the core disciplines ought to be assessed;
- Social studies is more than content—skills matter too;
- Elementary and middle school social studies need to be assessed—social studies is a K-12 curriculum;
- Standardized social studies testing will continue at grades 10 and 11—large-scale testing seems unlikely to return to grades 5 and 8;
- Social studies offers opportunities for exploring alternatives to large-scale testing—performance-based assessments at the elementary and middle school levels could extend and support the Regents testing program
- Building an *assessment program*, which incorporates both large-scale testing and local performance-based tasks may foster more confidence among the various stakeholders and evaluate students' social studies knowledge and skills more fully.

Based on these assumptions, we propose the construction of a New York State social studies assessment program that has two components—Regents-level testing at grades 10 and 11 and performance-based tasks located at the elementary and middle levels. The performance-based tasks could be required (e.g., for promotion to the next grade, for promotion to the next grade level—elementary to middle and middle to high school, or for high-school graduation), but local districts would have considerable discretion in how the assessment tasks were constructed, scored, and interpreted.

The elementary and middle school performance assessment tasks would feature the following criteria:

- Draw on a range of social studies disciplines and skills;
- Represent one of the array of tasks evident in a NYS Social Studies Toolkit inquiry (e.g., Summative Performance Task, Summative Extension activity, or Taking Informed Action exercise);

- Expressed through one or more modalities (e.g., written, oral, still or moving images);
- Require both a group and an individual effort;
- Administered at grade levels (e.g., 4th and 7th grades) chosen at the local level;
- Scored by teachers, but with the possibility of including parents and community members.

Although such a program of performance-based assessments would be required, districts could opt for waivers of up to three years in order to put the program fully into place.

Conclusion

In this white paper, we have reviewed the current state of research on classroom-based and large-scale assessments, explored the construct of face validity, considered some of the more promising developments in educational assessment, and offered a proposal that could more fully assess the K-12 social studies curriculum and inspire more confidence among various stakeholders in the assessment of students' social studies knowledge and skills.

References

Allen, D., Ort, S., & Schmidt, J. (2009). Supporting classroom assessment practice: Lessons from a small high school. *Theory into Practice*, 48(1), 72-80

Aloise, R., Longhurst, R., & Platin, D. (2014). *Report: NYS PTA survey of opinion on the Common Core, student testing, and advocacy priorities*. New York: New York State Parent Teachers Organization. Available online at: http://nyspta.org/pdf/Advocacy/Report_CCLS_Survey_Jan_2014.pdf.

Alvarez, L. (2014, November 9). States listen as parents give rampant testing an F. *New York Times*. Available online at: http://www.nytimes.com/2014/11/10/us/states-listen-as-parents-give-rampant-testing-an-f.html?_r=0

Anastasi, A. (1988). *Psychological testing*. New York, NY: Macmillan.

Au, W. (2009). *Unequal by design: High-stakes testing and the standardization of inequality*. New York: Routledge.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*(5), 258-267.

Baker, E. (1994). Learning-based assessments of history understanding. *Educational Psychologist, 29*(2), 97-106.

Banks, S. (2005). *Classroom assessment: Issues and practices*. New York: Pearson.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*, 7–74.

Black, P., & William, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation, and Accountability, 21*(1), 5-31

Bonner, S. (2013). Validity in classroom assessment: Purposes, properties, and principles. In J. McMillan (Ed.). *Handbook of research on classroom assessment* (pp. 87-106). Thousand Oaks, CA: Sage.

Breakstone, J., Smith, M., & Wineburg, S. (2013). Beyond the bubble in history/social studies assessments. *Phi Delta Kappan, 95*(5), 53-57.

Brody, L. (2014, April 10). New York State education commissioner pushes ahead. *Wall Street Journal*. Available online at:
http://www.wsj.com/news/articles/SB10001424052702304512504579493913077362226

Brookhart, S. (2013). Grading. In J. McMillan (Ed.). *Handbook of research on classroom assessment* (pp.257-272). Thousand Oaks, CA: Sage.

Brophy, J., & Alleman, J. (1998). Assessment in a social constructivist classroom. *Social Education, 62*(1), 32-34

Bushaw, W., & Lopez, S. (2013). Which way do we go? *Phi Delta Kappan, 95*(1), 9-25. Available online at: http://pdkintl.org/noindex/2013_PDKGallup.pdf.

Casner-Lotto, J. (2006). *Are they really ready to work?: Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce*. New York: Conference Board.

Chatterji, M. (March 12, 2014). Validity counts: Let's mend, not end, educational testing. *Education Week*, *33*(24), 30, 36.

Cizek, G. (2009). Reliability and validity of information about student achievement: Comparing large-scale and classroom testing contexts. *Theory Into Practice, 48*, 63–71.

Darling-Hammond, L. (1991). The implications of testing policy for quality and equality. *Phi Delta Kappan, 73*(3), 220-225.

Darling-Hammond, L., & Adamson, F. (Eds.). (2014). Beyond the bubble test: How performance assessments support 21st century learning. San Francisco: Jossey-Bass.

Ducker, B., & Perlstein, D. (2014). Assessing habits of mind: Teaching to the test at Central Park East Secondary School. *Teachers College Record, 116*(2), 1-33.

Epstein, T. (1994). America revised revisited: Adolescents' attitudes towards a United States history textbook. Social Education, 58(1), 41-44.

Epstein, T. (2008). *Interpreting national history: Race, identity, and pedagogy in classrooms and communities.* New York: Routledge.

Fitchett, P., & Heafner, T. (2010). A national perspective on the effects of high-stakes testing and standardization on elementary social studies marginalization. *Theory and Research in Social Education, 38*(1), 114-130.

Gorin, J. (2013). *Assessment as evidential reasoning.* Princeton, NJ: The Gordon Commission. Available on-line at: http://www.gordoncommission.org/rsc/pdf/gorin_assessment_evidential_reasoning.pdf.

Grant, S. G. (2003). *History lessons: Teaching, learning, and testing in U.S. high school classrooms.* Mahwah, NJ: Lawrence Erlbaum Associates.

Grant, S. G. (Ed.). (2006). *Measuring history: Cases of state-level testing across the United States.* Greenwich, CT: Information Age Publishing.

Grant, S. G. (2007). Understanding what children know about history: Exploring the representation and testing dilemmas. *Social Studies Research and Practice, 2*(2), 196-208. (On-line at: http://www.socstrp.org)

Grant, S. G., & Salinas, C. (2008). Accountability and assessment. In L. Levstik & C. Tyson (Eds.), *Handbook of research in social studies education* (pp. 219-236). Mahwah, NJ: Lawrence Erlbaum Associates.

Hagopian, J., & Ravitch, D. (2014). *More than a test score: The new uprising against high-stakes testing.* Chicago: Haymarket.

Haladyna, T. (2004). *Developing and validating multiple-choice items* (3rd ed.). Mahway, NJ: Lawrence Erlbaum Associates.

Haladyna, T. (2009). The perils of standardized achievement testing. *Educational Horizons, 85*(1), 30-43.

Heitin, L. (2014, October 21). School leaders to trim testing, but keep yearly assessments. *Education Week.* Available online at: http://www.edweek.org/ew/articles/2014/10/22/09testing.h34.html

Horn, C. (2003). High-stakes testing and students: Stopping or perpetuating a cycle of failure. *Theory Into Practice, 42*(1), 30-41.

Horn, C. (2006). The technical realities of measuring history. In S. G. Grant, (Ed.), *Measuring history: Cases of state-level testing across the United States* (pp. 57-74). Greenwich, CT: Information Age Publishing.

Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Jones, G., & Jones, B. (2003). The unintended consequences of high-stakes testing. New York: Rowman & Littlefield.

Klein, A. (2014, October 13). Push to limit federal test mandates gains steam. *Education Week*. Available online at: http://www.edweek.org/ew/articles/2014/10/15/08testing.h34.html.

Kohn, A. (2000). *The case against standardized testing: Raising the scores, ruining the schools*. Portsmouth, NH: Heinemann.

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Boston: Harvard University Press.

Levstik, L., & Barton, K. (1997). *Doing history: Investigating with children in elementary and middle schools*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lin, S., & Singh, C. (2013). Can free-response questions be approximated by multiple-choice equivalents? *American Journal of Physics, 81*, 624-629.

Macken-Horarik, M. (2011). Building a knowledge structure for English: Reflections on the challenges of coherence, cumulative learning, portability and face validity. *Australian Journal of Education, 55*(3), 197-213.

Martin, D., Maldonado, S.I., Schneider, J., & Smith, M. (2011). *A report on the state of history education: State policies and national programs*. National History Education Clearinghouse. Available online at: http://teachinghistory.org/system/files/teachinghistory_special_report_2011.pdf.

McNeil, L. (2000). *Contradictions of school reform: Educational cost of standardized testing*. New York: Routledge.

Messick, S. (1989). Validity. Inn R. L. Linn (Ed.), Educational measurement (3rd edition, pp. 13-103). New York: Macmillan.

Modu, C., & Wimmers, E. (1981). The validity of the Advanced Placement English Language and Composition examination. *College English, 43*(6), 609-620.

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement, 22*(4), 287-293.

Nichols, S., & Berliner, D. (2014). Collateral damage: How high-stakes testing corrupts America's schools. Boston: Harvard Educational Press.

Nuthall, G., & Alton-Lee, A. (1995). Assessing classroom learning: How students use their knowledge and experience to answer classroom achievement test questions in science and social studies. *American Educational Research Journa*l, 32(1), 185-223

Parkes, J. (2013). Reliability in classroom assessment. In J. McMillan (Ed.). *Handbook of research on classroom assessment* (pp. 107-123). Thousand Oaks, CA: Sage.

Ravitch, D. (2011). *The death and life of the great American school system: How testing and choice are undermining education*. New York: Basic Books.

Ravitch, D. (2014). *Reign of error: The hoax of the privatization movement and the danger to America's public schools*. New York: Vintage.

Rogers, V., & Stevenson, C. (1988). How do we know what kids are learning in school? *Educational Leadership, 45*(5), 68-75.

Schug, M., Todd, R., & Beery, R. (1984). Why kids don't like social studies. *Social Education, 47*(5), 382-387.

Secolsky, C. (1987). On the direct measurement of face validity: A comment on Nevo. *Journal of Educational Measurement, 24*(1), 82-83.

Stiggins, R. (1998). *Classroom assessment for student success*. Washington, DC: National Education Association.

Stiggins, R. (2014). *Revolutionize assessment: Empower students, inspire learning*. New York: Corwin.

Supovitz, J. (2009). Can high stakes testing leverage educational improvement? Prospects from the last decade of testing and accountability reform. *The Journal of Educational Change, 10*(2-3), 211-227.

VanSledright, B. (2013). *Assessing historical thinking and understanding: Innovative designs for new standards*. New York: Routledge.

VanSledright, B. A., & Grant, S. G. (1994). Citizenship education and the persistence of classroom dilemmas. *Theory and Research in Social Education, 22*(3), 305-339.

VanSledright, B., Kelly, T., & Meuwissen, K. (2006). Oh, the trouble we have seen: Researching historical thinking and understanding. In K. Barton (Ed.), *Research methods in social studies education* (pp. 207-233). Greenwich, CT: Information Age Publishing.

Vogler, K. (2006). Measuring history through state-level tests: Patterns and themes. In S. G. Grant, (Ed.), *Measuring history: Cases of state-level testing across the United States* (pp. 303-320). Greenwich, CT: Information Age Publishing.

von Zastrow, C., & Janc, H. (2004). *Academic atrophy: The condition of liberal arts in America's schools*. Washington, DC: The Council for Basic Education.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco: Jossey-Bass.

Wills, J. (2007). Putting the squeeze on social studies: Managing teaching dilemmas in subject areas excluded from state testing. *Teachers College Record, 109*(8), 1980-2046.

Wineburg, S., Smith, M., & Breakstone, J. (2012). New directions in assessment: Using Library of Congress sources to assess historical understanding. *Social Education, 76*(6), 290-293

Yuan, K & Le, V. (2012). *Estimating the percentage of students who were exposed to deeper learning on the state achievement tests*. Santa Monica, CA: Rand. Available online at: http://www.hewlett.org/library/grantee-publication/estimating-number-students-who-were-tested-cognitively-demanding-items-through-state-achievement