

Identifying the Markers of History: Teachers and Researchers

Describe the Assessment of Historical Knowledge and Understanding

S. G. Grant
University at Buffalo

Jill M. Gradwell
Buffalo State College

Abstract

Although standardized testing of K-12 student knowledge and understanding garners considerable attention, few observers profess satisfaction with the assessments in place. In this exploratory paper, we report on the data gathered from an open-ended email survey of small, convenience samples of teachers and researchers. Although no clear consensus about alternative assessments of students' historical knowing and understanding emerged, we argue that the potential for a consensus exists. Any emergent consensus, however, must be negotiated with several issues in mind.

Introduction

While proponents and critics continue to argue about the advantages and disadvantages of state-level testing, they generally do so on a rhetorical rather than an empirical level (Cimbricz, 2002). Those arguments will likely continue because standardized state exams in most school subjects are unlikely to disappear. The possible exception is history/social studies,¹ given the uncertain spot that that subject holds in the No Child Left Behind legislation (Gaudelli, 2006). In fact, some states (e.g., Maine, Kansas) have abandoned their social studies tests to concentrate on reading and mathematics. At the same time, however, states that have not had social studies exams (e.g., Tennessee, West Virginia) have added them to their assessment portfolio. The total number of states with social studies tests has remained relatively constant at 22-24 for the last

seven years, but there has been a fair amount of movement into and out of the social studies testing arena (Grant, 2006b).

Although standardized testing of K-12 student knowledge and understanding garners considerable attention (Grant, 2006b), few observers profess satisfaction with the assessments in place. Questions about the validity and usefulness of high-stakes testing, in particular, exemplify many of the issues raised about assessment (Craig, 2004; Fuller & Johnson, 2004; Grant, 2006a; Grant & Salinas, in press; Haladyna, Nolen, & Haas, 1991; Linn, 2003). The limits of standardized test questions, the lack of student choice on tasks, the problem of single measures, and the price of the stakes attached to students' scores are all red flags to those who worry about the consequences of state-level testing programs.

Identifying the problems with state-level tests is easy. Doing so, however, has left both testing proponents and critics unsatisfied. To that end, we propose reframing the issue of testing vs. no testing to an issue about how we all—teachers, researchers, parents, and the public—may feel more confident about the markers used to assess students' historical knowledge and understanding. The principal research question that drives this study is as follows: Assuming that state-level testing is going to stay in place, what other markers might better measure students' historical knowing and understanding?

This question is not new: Presumably, researchers and teachers have debated informally the question of how to assess K-12 students' historical knowledge and understanding for some time (Kornhaber, 2004). The perception of a cultural divide between teachers and researchers has been long held (Leming, 1989; VanSledright & Grant, 1991). Beginning in the mid-1970s, Shaver (1977; 1981) argued that the diverse views these groups held about social studies education had the potential to undermine the profession as a whole. Later, Leming (1989; 1994) asserted that researchers and teachers hold fundamentally different views about the purpose of social studies education and the most meaningful types of instructional activities and approaches. VanSledright and Grant (1991) argued that the potential for a "conversational community" among social studies teachers and researchers has yet to materialize.² There may be some consensus between these two groups on alternative markers of students' historical knowing and understanding, but we do not know what it is.

In this exploratory paper, we report on the data gathered from an open-ended email survey of small, convenience samples of teachers and researchers.³ We argue that, although no clear consensus about alternative assessments of students' historical knowing and understanding emerged, the *potential* for a consensus exists. Any emergent consensus, however, must be negotiated with several issues in mind.

Methodology

In this exploratory study, 10 researchers and 10 teachers were invited to participate in an open-ended email survey. The total number of actual participants was 17 (8 teachers and 9 researchers). The history education researcher group was drawn from scholars working in eight different states and one Canadian province who conduct empirical studies around teaching, learning, and assessing history. Of those residing in the US, two were from the Midwest, two from the Northeast, three from the South, and one from the West. Of the represented states, six have history/social studies state exams with only one not carrying high-stakes consequences for students. The Canadian researcher works in a province that has a high school graduation history exam.

The history teacher sample was drawn locally from the groups we have worked with on a US Department of Education Teaching Traditional American History grant-funded, professional development project in New York state where there are high-stakes social studies exams in grades 10 and 11. The experienced teachers work in a large urban school district that has academic, neighborhood, and specialty-area schools. The students of these particular teachers varied in academic ability and ethnicity. Based on the New York State School Report Card for 2004-2005 year, 2.8 % of the students were American Indian, Alaskan, Asian, or Pacific Islander; 57.6% were Black; 13.6% were Hispanic, and 26% were White.

Data Collection

After agreeing to participate in the project, the participants were sent an open-ended, email survey. The demographic questions (e.g., undergraduate and graduate majors, years of teaching and courses taught) varied somewhat by group, but the principal questions were identical:

1. Assuming that state-level history/social studies testing is going to stay in place, what other assessment markers might better measure students' historical knowing and understanding?
2. What advantages and disadvantages do each of these markers offer?
3. How might such markers need modification for different kinds of students?
4. For each assessment listed, what criteria would differentiate a good representation vs. a poor one?
5. What psychometric problems (e.g., validity, reliability), if any, would you anticipate for any or all of the markers you have listed?

Data Analysis

With the small number of participants and the open-ended nature of the email survey, we employed open-coding procedures to analyze the data (Bogdan & Biklen, 1982). In our initial review, we read through the data several times, labeling ideas reflective of each question and identifying them with either teachers or researchers. Because of the perceived cultural divide between teachers and researchers, in our first level of analysis, we looked to see what prevailing patterns emerged in the responses within and across the participant groups. For the first question, in which we asked the respondents to identify alternative assessment tasks, we identified and coded for six categories of assessments: (a) traditional vs. non-traditional tasks, (b) school vs. real world applications, (c) higher vs. lower-level learning, (d) knowing vs. doing history, (e) high vs. low-level technology needs, and (f) individual vs. group tasks. For the second question, which asked about advantages and disadvantages of each assessment task offered, we identified instances of student choice, deeper insights into student thinking, and fostering student thinking as advantages, with time and teachers' content and pedagogical knowledge as disadvantages. With question three, which asked about necessary modifications, we identified assessment procedures, teacher support, tasks, and materials as useful analytic labels. The idea of being able to differentiate good from poor student representations (question #4) divided across high and low quality indicators and the use of rubrics. Finally, the psychometric concerns teachers and researchers identified broke down into three areas: technical issues, validity, and reliability.

With all the data coded, we completed a series of frequency counts for the data associated with each question in order to see what patterns might surface. For example, we conducted overall counts of coded items such as the number of references to validity as a psychometric problem, the number of references to traditional vs. non-traditional tasks as alternative markers, and the number of references to student choice as an advantage to alternative assessment markers. With those frequency counts in hand and represented in a data display (Miles & Huberman, 1994), we identified patterns across the data and between the two sets of respondents, keeping in mind the argument of a perceived divide between teachers and researchers. We then analyzed the patterns for trends and themes.

Taken together, our analysis suggests that teachers and researchers hold considerably coherent views on most of the questions we posed, thus we argue that a consensus position on assessment is possible.

A Potential Consensus on the Markers of History

Although there are important issues that must be dealt with before any real consensus position on alternatives to state-level testing might be identified, our analysis supports the idea that there are far more points of agreement between teachers and researchers than some observers (Leming, Ellington, & Porter-Magee, 2003; Schug, 2003) believe. Of the five areas we probed, each suggested considerable agreement between the two groups.

In the sections that follow, we present our findings related to the kinds of alternative markers offered, the modifications felt to be necessary, the ability to judge good vs. poor student representations, the potential for psychometric problems, and the advantages and disadvantages of alternative assessments.

An Emergent Consensus on a Range of Assessment Alternatives

Given the open-ended nature of the survey, the absence of clear consensus on a single set of alternative assessments either within each respondent group or across them is no surprise. There were some similarities within each group, but those similarities were no stronger than the similarities across the two groups. On the whole, teachers and researchers preferred real-world, higher-order, non-traditional, low-tech, and doing kinds of assessments.

Overall, the 17 participants (8 teachers, 9 researchers) offered 21 distinct assessment suggestions with each group naming essentially the same number (see Table 1). Almost half of the items ($n = 10$) were cited by both groups (e.g., projects, term papers, portfolios, oral presentations). Several were offered by one group or the other. For example, one or more teachers nominated oral histories, worksheets, and policy reports as useful assessment alternatives, while one or more researchers nominated performances, discussions, and critiques.

Table 1***The Range of Teacher and Researcher Assessment Suggestions***

Tasks	No. of Teacher Suggestions/Task	No. of Researcher Suggestions/Task	Total No. of Suggestions/Task
1. Project	4	6	10
2. Essay	4	3	7
3. Performance	0	6	6
4. Portfolio	2	3	5
5. Non-traditional, short answer questions ⁴	3	1	4
6. Term Paper	2	1	3
7. Discussion	0	3	3
8. Non-traditional M-C questions ⁵	0	3	3
9. Oral presentation	2	1	3
10. Extended Response	0	2	2
11. Multiple-choice questions	1	1	2
12. Critique	0	2	2
13. Historical fiction, song, poem	1	1	2
14. DBQ	1	1	2
15. Poster	1	1	2
16. Test	1	0	1
17. Oral history	1	0	1
18. Worksheets	1	0	1
19. Model/diorama	1	0	1
20. Policy report	1	0	1
21. PowerPoint	1	0	1
TOTALS	27	35	62

The breadth of tasks suggested is not surprising: Even a surface reading of the classroom assessment literature supports the conclusion that there are many ways to assess students' knowledge and understanding. More surprising are two additional observations: (1) the clustering of tasks in that over half of the total suggestions (32 of 62) involve only five assessments, and (2) the considerable agreement between the two groups: Both teachers and researchers suggested projects (10), essays (7), portfolios (5), and non-traditional kinds of short answer questions (4), while only researchers (6) suggested performances. Of course, one could focus on the total number of responses and conclude that the gulf between teachers and researchers is wide, but doing so, in our view, gives undue weight to single suggestions.

Of the five analytic criteria we applied to the data, teachers and researchers expressed substantial agreement on each. A comparison of the aggregate preferences of both groups demonstrated a preference for real-world over school-based tasks, higher-order thinking over lower-level thinking tasks, non-traditional over traditional tasks, "doing" history over "knowing" history tasks, and low-level technology over high-level technology tasks.

Real-world tasks favored over school-based. The criteria of real world vs. school tasks centers on the context in which the assessment is typically associated and the range of knowledge and skills necessary. School-based tasks are those generally found only within

classroom contexts and which call for a limited knowledge and/or skill set. Examples include tests, term papers, and worksheets. Real-world tasks tend to require that students use multiple knowledge and/or skills to complete a task that has an authentic or real world dimension (Newmann, Bryk, & Nagaoka, 2001). Examples include discussions, reports, and oral presentations. Four tasks could be classified as either school-based or real-world. For example, although social studies teachers rely heavily on essay assignments, historians also write in this genre.

Both teachers and researchers offered slightly more real-world than school-based tasks (15 real-world tasks; 11 school-based tasks).⁶

Table 2

Real-World v. School-Based Assessment Tasks

Task	School-Based (S) and/or Real-World (R) Tasks	Teacher (T) and/or Researcher (R) Suggestions
1. Project	R	T/R
2. Essay	S/R	T/R
3. Performance	R	R
4. Portfolio	R	T/R
5. Non-traditional, short answer questions	S	T/R
6. Term Paper	S	T/R
7. Discussion	R	R
8. Non-traditional M-C questions	S	R
9. Oral presentation	R	T/R
10. Extended Response	S/R	R
11. Multiple-choice questions	S	T/R
12. Critique	R	R
13. Historical fiction, song, poem	R	T/R
14. DBQ	S/R	T/R
15. Poster	R	R
16. Test	S	T
17. Oral history	R	T
18. Worksheets	S	T
19. Model/diorama	R	T
20. Policy report	R	T
21. PowerPoint	S/R	T
TOTALS	10 School-Based; 15 Real-World Tasks	Teachers—10 Real-World and 7 School-Based Tasks; Researchers—11 Real-World and 7 School-Based Tasks

Looking at the between group data, the ratio of school vs. real-world assessments is essentially the same for both groups: In the aggregate, researchers offered 11 real-world and 7 school-based tasks; teachers offered 10 real-world tasks and 7 school-based tasks. Although there is some variation in their specific suggestions, both groups favored real-world over school-based assignments and did so with essentially the same frequency.

Higher-level tasks favored over lower-level. In the case of higher v. lower-level thinking tasks, teachers' and researchers' preferences again mirror one another. On this measure, however, a very strong trend supports the use of higher-level tasks. We define the difference between higher- and lower-level thinking tasks as one of intellectual complexity. Higher-level assessments call for the analysis, synthesis, and evaluation of ideas, data, and documents (e.g., oral histories, historical fiction, essays), while lower-level assessments typically demand only the recall or recognition of ideas or simple comparisons (e.g., multiple-choice questions, worksheets). Given the wide ranging possibilities, we categorized two tasks—tests and essays—as both higher and lower level.

By a roughly 5-1 margin, teachers and researchers favored higher-order measures over lower-level ones.

Table 3

Higher-Level v. Lower-Level Assessment Tasks

Task	Higher-Level Tasks (H) and/or Lower-Level (L) Tasks	Teacher (T) and/or Researcher (R) Suggestions
1. Project	H	T/R
2. Essay	H/L	T/R
3. Performance	H	R
4. Portfolio	H	T/R
5. Non-traditional, short answer questions	H	T/R
6. Term Paper	H	T/R
7. Discussion	H	R
8. Non-traditional M-C questions	H	R
9. Oral presentation	H	T/R
10. Extended Response	H	R
11. Multiple-choice questions	L	T/R
12. Critique	H	R
13. Historical fiction, song, poem	H	T/R
14. DBQ	H	T/R
15. Poster	H	R
16. Test	L/H	T
17. Oral history	H	T
18. Worksheets	L	T
19. Model/diorama	H	T
20. Policy report	H	T
21. PowerPoint	H	T
TOTALS	19 Higher-Level Tasks; 4 Lower-Level Tasks	Teachers—13 Higher-Order and 4 Lower-Level Tasks; Researchers—14 Higher-Level and 2 Lower-level Tasks

Both teachers and researchers gave clear preference to higher-level tasks as alternatives to current standardized exam formats. Moreover, they did so in virtually identical ways in terms of the aggregate frequencies.

Non-traditional tasks favored over traditional. Overall, teachers and researchers clearly preferred non-traditional tasks over those that are more pedestrian. Of the 21 tasks offered, we classified 16 as non-traditional (see Table 4).

Traditional tasks are those staples of classroom assessment. Examples include term papers, worksheets, and tests. Non-traditional tasks, then, are those that are either relatively rare (e.g., policy reports) or are rarely used as assessments of student performance (e.g., discussions). Unsure how respondents were using the term “extended response,” we identified it as both a traditional and a non-traditional task.

Table 4

Traditional v. Non-Traditional Assessment Tasks

Task	Traditional (T) v. Non-Traditional (NT) Tasks	Teacher (T) and/or Researcher (R) Suggestions
1. Project	NT	T/R
2. Essay	T	T/R
3. Performance	NT	R
4. Portfolio	NT	T/R
5. Non-traditional, short answer questions	NT	T/R
6. Term Paper	T	T/R
7. Discussion	NT	R
8. Non-traditional M-C questions	NT	R
9. Oral presentation	NT	T/R
10. Extended Response	T/NT	R
11. Multiple-choice questions	T	T/R
12. Critique	NT	R
13. Historical fiction, song, poem	NT	T/R
14. DBQ	NT	T/R
15. Poster	NT	R
16. Test	T	T
17. Oral history	NT	T
18. Worksheets	T	T
19. Model/diorama	NT	T
20. Policy report	NT	T
21. PowerPoint	NT	T
TOTALS	6 Traditional Tasks; 16 Non-Traditional Tasks	Teachers—5 Traditional and 10 Non-Traditional; Researchers—4 Traditional and 12 Non-Traditional Tasks

The clear preference for non-traditional over traditional assessment tasks emerged across both teacher and researcher responses and in relatively equal frequencies.

Doing history favored over knowing history. The fourth criteria on which teachers and researchers largely seem to agree is the distinction between knowing and doing history. Knowing history refers to the largely passive student stance of taking in facts and interpretations and replicating them on an assessment. A classic example of knowing history is a worksheet. Doing history, by contrast, reflects a more active stance as students engage in the habits and practices of

historians. Examples of doing history include performances and oral histories. The line between knowing and doing history can blur, however. Multiple-choice test questions fit within the knowledge category, but the kinds of non-traditional, multiple choice respondents describe seem more in line with the characteristics of doing history. More complex still are essays and tests, which *can* emphasize pedantic goals just as easily as they can more ambitious objectives.

Teachers' and researchers' responses, by nearly a 5-1 margin, suggested tasks that reflect a doing rather than knowing history approach.

Table 5

Knowing v. Doing History Assessment Tasks

Task	Knowing (K) and/or Doing (D) Tasks	Teacher (T) and/or Researcher (R) Suggestions
1. Project	D	T/R
2. Essay	K/D	T/R
3. Performance	D	R
4. Portfolio	D	T/R
5. Non-traditional, short answer questions	D	T/R
6. Term Paper	D	T/R
7. Discussion	D	R
8. Non-traditional M-C questions	D	R
9. Oral presentation	D	T/R
10. Extended Response	D	R
11. Multiple-choice questions	K	T/R
12. Critique	D	R
13. Historical fiction, song, poem	D	T/R
14. DBQ	D	T/R
15. Poster	D	R
16. Test	K/D	T
17. Oral history	D	T
18. Worksheets	K	T
19. Model/diorama	D	T
20. Policy report	D	T
21. PowerPoint	D	T
TOTALS	4 Knowing Tasks; 19 Doing Tasks	Teachers—4 Knowing and 13 Doing Tasks; Researchers—2 Knowing and 14 Doing Tasks

The results from our analysis of the knowing v. doing criteria largely mirrors that of the higher-level thinking v. lower-level thinking criteria (see Table 3): Teachers and researchers heavily favored the more ambitious approach and did so in essentially the same proportions. They might disagree about the value of particular assessments, but they clearly preferred the same kinds of tasks.

Low-tech tasks favored over high-tech. Although teachers and researchers expressed considerable agreement in their choices of alternative assessments, a surprise surfaced: Both groups favored low tech over high tech assessments.

Table 6

High-Tech v. Low-Tech Assessment Tasks

Task	High-Tech (H) v. Low-Tech (L) Tasks	Teacher (T) and/or Researcher (R) Suggestions
1. Project	H/L	T/R
2. Essay	L	T/R
3. Performance	L	R
4. Portfolio	H/L	T/R
5. Non-traditional, short answer questions	L	T/R
6. Term Paper	L	T/R
7. Discussion	L	R
8. Non-traditional M-C questions	L	R
9. Oral presentation	H/L	T/R
10. Extended Response	L	R
11. Multiple-choice questions	L	T/R
12. Critique	L	R
13. Historical fiction, song, poem	H/L	T/R
14. DBQ	L	T/R
15. Poster	L	R
16. Test	L	T
17. Oral history	H	T
18. Worksheets	L	T
19. Model/diorama	H/L	T
20. Policy report	L	T
21. PowerPoint	H	T
TOTALS	7 High-Tech Tasks; 19 Low-Tech Tasks	Teachers—7 High-Tech and 13 Low-Tech Tasks; Researchers—4 High Tech and 15 Low-Tech Tasks

As noted above, we find surprising the preference teachers and researchers gave to low-tech over high-tech assessments. We had no particular hunch as to which group would favor high-tech tasks, but, given the emphasis on technology today, we would have predicted that at least one group would. True, almost twice as many teachers offered high-tech assessments than did researchers, but the small size of our sample undercuts our confidence in that finding.

Potential consensus, potential pitfall. Looking across these data, we see two trends. First, the types of alternative assessment tasks teachers and researchers suggest demonstrate far more consensus than some have predicted (Leming, 1989; Shaver, 1977). Second, a potential pitfall looms: the slipperiness of assessment language.

The assumed gulf between history teachers and researchers fails to unfold in the data for this study. The stereotypes of practical-minded teachers and lofty-minded professors may characterize some in their respective professions, but the sample drawn for this study suggests otherwise. Rather than affirm hard and immutable stances on student assessment, the two groups demonstrate considerable agreement, especially on those tasks most often mentioned.

The patterns revealed in our analysis of teacher and researcher preferences for alternative assessments are interesting, for they suggest that there may be far more common ground than

many assume. But at least one hurdle looms large: language. The slippery nature of language is widely recognized (Austin, 1975; Searle, 1989). “The meaning of words,” Bernbaum (1962) notes, “varies with the context in which they are used, and also with the contexts in which speaker and listener have previously experienced the words” (p. 39). Despite the open-ended nature of our survey, we wonder whether the respondents are using assessment language in consistent ways.

A single example should suffice: “Projects” was the most commonly-offered suggestion, yet we are struck by the different ways that respondents use this term. “Senior project,” “individual or personal project,” and “inquiry-based projects” are just some of the identifiers that teachers and researchers employed. Moreover, at times, participants offered virtually the same description but called it two different things. For example, a task that requires the development of a question, use of historical resources, and an evidence-based interpretation was labeled a “local history project” by one participant and a “performance assessment” by another. Such examples suggest that teachers and researchers may be interested in the same set of ideas, but slippery language labels may undercut a consensus.⁷

General Agreement about Modifications to Alternative Assessments

Widespread agreement surfaced across both teacher and researcher groups around the idea that all students can and should be assessed with alternative measures. Moreover, concerns expressed about the difficulties possible for the least advantaged students (i.e., minority, ESL, special education) especially on open-ended written tasks were also widely shared. Specifically, each group focused on modification of assessments in the areas of the assessment procedures, teacher support, tasks, and materials.

Although there was consistency between the groups, two observed differences emerged: Only teachers suggested that assessments be modified in ways allowing for students to help one another (e.g., through peer support); and only researchers mentioned modifying the materials necessary for students to complete the alternative assessment.

Agreement on assessment procedures, support, tasks, and materials. Assessment procedures refer to the ways in which teachers evaluate assessments. Both researchers and teachers made comments about modifying the method for assessing the markers they nominated. One teacher believed that grading should “be based on improvement as well as achievement.” A researcher felt that students should be evaluated both on writing style and substance: “Ideally, teachers should be able to separate out the historical/social studies nature of the response from surface features, such as the use of Standard English.” Because the students completing the assessment tasks would represent a variety of levels, another researcher suggested there should “be different expectations for sophistication in writing and sophistication in texts read. Students will need different levels of support for reading primary documents.”

A second set of modifications that both teachers and researchers suggested involved the need for teacher support. As a modification for a portfolio assessment, a teacher recommended having students complete most of the work in class. A researcher suggested increased teacher scaffolding to help students build written and oral communication skills.

Teachers and researchers also offered ideas to modify the assessment tasks, such as making them more open-ended and providing adequate time. As one researcher noted, “Classroom performances should be designed in such a way that they are open-ended enough to allow for a variety of student responses, including written, verbal, or visual responses.” Another

researcher highlighted the importance of giving students the appropriate amount of time to complete the task: “They should have time to finish and teachers [sh]ould not just drop this on students but have a schedule to keep that is known.”

Some puzzling differences. What puzzled us were the instances in which teachers and researchers differed. First, only researchers suggested the modification of materials needed to complete an alternative assessment. Some of these modifications included providing a variety of primary sources, varying the reading levels of materials, and taking into consideration students’ facility with the level of English used in the resources. At first, we were surprised that no teacher mentioned altering reading levels among materials. One of the concerns teachers often cite about standardized tests is that they are basically reading tests (Grant et al., 2002), so we thought that varying the reading ability of materials used in alternative assessments would be in the forefront of teachers’ minds. But on further reflection, we suspect that, *because* teachers assume standardized tests to be largely reading tests, this modification may have been too obvious for them to bother identifying.

Another point of interest to us was that only teachers suggested peer support as a modification. As one teacher noted, “[f]or students who wish to explore an area where their confidence is shaky, they could work in a group and be supported.” Another teacher noted that when students work with partners, they may be more likely to complete the task: “Working in pairs helped with the little questions some students had...the little things that often lead to the pencil being put down and the work being left incomplete.” We wondered why researchers would not suggest creating assessment markers that involve a group effort, especially when cooperative learning and pair sharing seem commonplace in the classroom (Slavin, 1996). At this time, we cannot account for this difference.

* * * * *

The points of difference described above should not obscure the many points of convergence between teachers and researchers. In the previous section, we argued that teacher and researcher respondents offered similar assessment tasks; in this section, we see that they are in general agreement about the kinds of modifications necessary to ensure that all students have access to these alternative markers of historical knowing and understanding.

Moreover, we see that the respondents continue to support the notion of alternative assessments. Both teachers and researchers offer a clear-eyed view of the need to make a range of modifications. But they do not stratify their suggestions, holding that only the most able students will be able to complete the kinds of real-world, higher-level thinking tasks that most history educators support (Grant, 2003; VanSledright, 2002). Instead, they implicitly argue that all students need access to tasks that push rather than inhibit their thinking.

Concurrence on Judging Students’ Representations

Our fourth survey question asked the respondents to identify the criteria necessary to distinguish between good and poor representations of students’ ideas. Because we anticipated that the assessments suggested would necessitate more “subjective” judgments than say multiple-choice questions, we wondered what evaluative factors the respondents would identify and whether there would be any coherence between teachers’ and researchers’ criteria.

General agreement on indicators and rubrics. With the exception of student effort (which was noted by only one teacher), no differences surfaced across the teacher and researcher lists of evaluation criteria. One or more respondent in each group listed high quality indicators, low quality indicators, and rubrics as the key considerations.

By “high quality indicators,” we mean references to student work that displays sophistication of thought and evidence of historical thinking and analysis. For example, a teacher noted that a good representation would be “informative, analytical, include relevant examples, and make connections to events that occurred due to that topic.” A researcher offered an expanded list of criteria:

Depth of information; sophistication of interpretation; use of evidence to support interpretation; evidence of historical thinking “skills” (i.e., perspective, cause and effect, agency, and the like); significance: The students’ arguments [should] include attention to/establish the grounds for the historical significance of the questions/investigations pursued.

“Low quality indicators,” by contrast, focus on the accuracy of and the manner in which information is presented. A teacher listed “attention to detail, correctness (is it historically accurate?), is it clear (by virtual of the drawing, painting, drama)” as factors. A researcher identified “clarity in presentation of position, whether written, oral, or visual” as necessary to a good representation.

Finally, “rubric” refers to any mention of that assessment tool. Again, both teachers and researchers cited this instrument as important, although there was considerable variation in how it was described. For example, a teacher coupled the idea of a rubric with largely low quality measures: “A rubric including accuracy, neatness, and prompt completion of the work.” By contrast, a researcher advocated for rubrics based in the high-quality measures Fred Newmann and his colleagues (Newmann, Bryk, & Nagaoka, 2001) describe: “I would use Fred Newmann’s rubrics for authentic intellectual work.”

As the examples above suggest, some differences in references to high and low quality measures surfaced across the teacher and researcher responses. That said, researchers were far more likely to list high rather than low quality measures, while teachers were more likely to do the reverse. Interestingly enough, researchers named rubrics as criteria far more frequently than did teachers.

Some considerations. Looking across these findings, three points surface. The fact that only one of the eight teachers mentioned rubrics is interesting given that, as New York state teachers, they are well aware of the rubrics used on the state tests to assess student performance on both thematic and Document-Based Question essays. Less surprising is the preference by researchers for high quality measures. Most of the researchers we invited to participate have analyzed students’ work and have advocated for more sophisticated tasks and more ambitious evaluations of those tasks. That fewer teachers mentioned high quality measures may be a reflection of the fact that, although the state tests require student essays, the rubrics, in some teachers’ eyes, tend to privilege low-level knowledge and skills (Grant et al., 2002). Finally, we were struck by the general faith in the possibilities of distinguishing between good and poor representations and the uncritical use of rubrics as a means of making that distinction. On the first point, every respondent noted one or more criteria that could be used to evaluate students’ representations, but only one respondent challenged the *idea* of evaluating those representations.

In other words, although there was some variation in the evaluative factors listed, all the respondents seemed to think that it was possible to evaluate students' work fairly.

This result interests us because, although the use of measures like rubrics is ubiquitous in classrooms and on state exams, there are serious challenges to such practices. A number of researchers (Nuthall & Alton-Lee, 1995; Rogers & Stevenson, 1988) conclude that students may "know" different things, depending on the kinds of questions they are asked. Other researchers (Baker, 1994) have demonstrated an absence of agreement among educators as to what counts as a quality response to an essay prompt. Such problems were unacknowledged by all but one respondent. That researcher noted, "Ultimately, the profession might develop some consensus over what it looks like to demonstrate understanding of particular concepts (as with empathy and evidence in Britain, for example), but we're a long way from that in the US." There may be a consensus across teachers' and researchers' evaluation criteria, but, at this point in time, that consensus may be somewhat fragile.

Without diminishing the differences expressed above, we continue to be surprised at the many points of convergence between teachers and researchers. Beyond that convergence, we continue to note the apparent support for more rather than less meaningful assessments. The difficulties in constructing grading rubrics aside for the moment, we were struck that teachers and researchers both appeared to believe that *all* students' representations could be judged in a fair manner.

A Common Sense of Potential Psychometric Problems

The final question of the email survey asked the respondents to list and describe any potential psychometric problems associated with the alternative assessments they suggested. Once again, no particular differences emerged across the teacher and researcher respondents' lists. Researchers were more likely to focus on reliability issues and to cite multiple problems than were teachers, but both groups seemed to think that such problems were no barrier to pursuing alternative assessment approaches.

Considerable agreement on technical problems, validity, and reliability. The list of teachers' and researchers' potential concerns included technical issues, validity, and reliability. Technical problems focused on questions about problems of converting evaluations of alternative assessments into comparable scores. A teacher asked, "How would you compare districts state-wide, based on individual/personal assessments? How to convert the data/scores into percentages?" A researcher bluntly concluded, "Statistical analyses would be a nightmare."

Validity issues also arose frequently across both groups' responses. Validity concerns can take many forms (Horn, 2006) including content, construct, criterion, and consequential (Messick, 1988). Of those four, content and consequential validity, were noted by the respondents. For example, a researcher observed, "Content sampling is the problem for any state-level test, because it's impossible to test the entire universe of what students are expected to learn." A teacher offered an example of consequential validity when she asked, "Is it a valid assessment if the students work together?" A researcher raised the potential problem of valid scores for second language learners:

Validity of constructed-response items and other forms of alternative assessment is impacted by factors such as language ability; it's difficult for a student with limited oral

or written language skills to express his/her understanding, as well as for evaluators to remove the language skills from the evaluation of content.

More often recognized than any other problem, however, was reliability and, by extension, the issue of bias. Although far more researchers raised these concerns than did teachers, the essence of the problem was common—discomfort over who and how alternative assessments would be scored and how those scores might be interpreted. One teacher observed, “The evaluations are subjective for sure.” A second worried about bias, based on the “different levels of teachers’ understanding of historical accuracy.” One researcher pointed to the lack of any apparent consensus on “what constitutes historical/social thinking.” Continuing, he added, “Reliability of scoring of any open-ended item beyond a short constructed-response item is likely to be low.” Another researcher noted the problem of trying to scale up results: “Reliability is a significant issue, as these types of activities are not easily scalable to large groups. No quantitative solution immediately comes to mind.”

These two researchers expressed some doubt about whether the reliability issues they describe can be adequately resolved. Their conclusions were a distinct minority view, however, as most teachers and researchers either indicated confidence that the psychometrics could be worked out or simply stated the problem(s) without judgment. For example, a researcher observed, “History is inferential—the grading of a portfolio opens up issue of reliability/validity, but that is not an issue that should prevent history from becoming more than it currently is.” After listing oral presentations as a useful assessment, a teacher added, “The spoken word could be assessed by rubrics and markbands that would be as acceptably valid as rubrics and markbands that are used to assess the written word.”

A consensus view. Looking across these findings, once again we found a seeming consensus view being expressed. Teachers and researchers not only seemed to identify a similar range of potential problems, but they appeared to take those problems in stride. This conclusion should not surprise readers: No assessment is perfect, so technical, validity, and reliability problems are always possible. Test developers try to minimize these issues, but none would claim to have eliminated them (Horn, 2006).

One other point is worth noting. Our preliminary conclusion that the respondents had given short shrift to the problems of using rubrics to score student representations evaporated, as we read their responses to the question about psychometric concerns. Employing a rubric does not solve the problem of reliability because, as both teachers and researchers pointed out, rubrics can be interpreted differently by different raters.

* * * * *

Thus, again, we see more coherence than difference across teachers’ and researchers’ responses. Neither group is sanguine about the potential psychometric problems, but the fact that there appears to be so much common ground about the nature of those problems speaks to the possibility of coherent and more authentic solutions.

Differences Emerge: Advantages and Disadvantages

In the second survey question, we asked the participants what advantages and disadvantages each of their suggested assessment markers might offer. Of all the survey

responses, it was in the area of pros and cons that the most differences surfaced. Although both groups of participants seemingly had similar responses for advantages and disadvantages of alternative markers, two distinct patterns emerged: Teachers suggested that alternative tasks would engage students in higher-level thinking skills, whereas researchers mentioned the disciplinary nature of those skills; only researchers cited the lack of teachers' pedagogical content knowledge as a disadvantage.

Similarities and differences on advantages of alternative tasks. There were some clusters of commonly-cited advantages (e.g., giving students more choices, offering teachers deeper insights into students' understanding, fostering students' thinking) and disadvantages (e.g., the time and energy necessary to create, administer, and score these assessments; teachers' capacity to teach the necessary knowledge, skills, and dispositions). For both groups of participants, the advantages listed usually related to students while the disadvantages related to teachers.

One of the major advantages teachers and researchers cited for using alternative assessments was the opportunity for students to have a choice in the selection of tasks. For example, one teacher suggested that choice "allow[s] visual and non-visual learners an accessible method of expression and an opportunity to include issues that concern the student or group of students." Similarly, a researcher described the chief benefit of having students complete a term paper was that "often students have some say in the topic." Students commonly perceive history as boring but, as both teachers and researchers intimated, when students have some input into their learning, they may be more attentive. One researcher advocated the idea of a history fair project because it offered students an element of choice: "[It] typically let[s] students select the topic (within a theme each year) and the means of communication (paper, skit, film, etc.) [and] that may enhance student engagement."

Another advantage cited by both teachers and researchers was that alternative markers may provide teachers with multiple ways of assessing students' current and emergent understandings. For example, a researcher noted that a student performance "deviates from single measurement scales and offers student choices and multiple ways of measuring student success." Likewise, a teacher reported that, in the example of a written alternative assessment, "a teacher can monitor not only writing skills in a written assessment but also the content [and] facts." Finally, a researcher suggested that markers such as classroom performances "potentially provide teachers with much deeper insight into the nature of students' thinking." Consistent across teachers and researchers, then, was the view that alternative assessments could offer multiple means by which to understand what students know.

One other commonly-expressed advantage to using alternative markers was that they foster critical thinking among students. Willingham (2007) describes critical thinking as "seeing both sides of an issue, being open to new evidence that disconfirms your ideas, reasoning dispassionately, demanding that claims be backed by evidence, deducing and inferring conclusions from available facts, and solving problems, and so forth" (p. 8). The teachers in this study acknowledged that alternative historical tasks promote these generic critical thinking skills and tended to use phrases like "critical thinking," "writing skills," and "analysis" to describe the benefits. Willingham (2007) notes, however, that there is a distinction between generic critical thinking skills and domain-specific skills: "Then, too, there are specific types of critical thinking that are characteristic of different subject matter: That's what we mean when we refer to 'thinking like a scientist' or 'thinking like a historian'" (p. 8). Researchers tended to focus on

specific disciplinary-like habits that could be developed with children if alternative markers were employed. A researcher captured this belief in the phrase “history-in-action”:

The advantages include assessing “history in action”—students actually employing historical thinking to some useful and important end: answering powerful questions about the past, about connections between past and present, about what it means now and has meant in the past to be human.

One explanation for this difference in emphasis may be that some teachers do not think their students capable of engaging in historical thinking (Fehn & Koeppen, 1998; Yeager & Wilson, 1997). Another reason may be that teachers do not view the development of historical thinking skills as purposeful and meaningful for students. For teachers, the goal may be to prepare students for a participatory democracy, and the fostering of disciplinary skills may or may not help them to reach their goal (Barton & Levstik, 2004). One other possibility may be that teachers intuitively sense historical thinking skills are the same sorts of skills that are used in the real world. As VanSledright (2004) suggests, “Historical thinking is a very close relative to active, thoughtful, critical participation in text- and image-rich democratic cultures” (p. 223). Teachers may not use jargon like “historical thinking” or “doing history,” but instead may focus more on the implications this type of thinking has for students in and out of the classroom. One teacher hinted at this position when she reported the benefits of “inquiry-based” assessments for students were “deeper understanding of material and enhanced ability to apply learned knowledge to similar but different circumstances.”

Similarities and differences on disadvantages to alternative tasks. Turning now to the disadvantages, one main pattern we identified was that of time and energy required of teachers to develop, administer, and score the assessments. Phrases teachers used to characterize the process included “tons of teacher preparation” and “massive time required to properly correct.” A researcher cited the demands of the scoring process for any alternative assessment:

The obvious disadvantages include time and difficulty in scoring, the challenge of comparing very different sorts of products and developing rubrics that assess but do not narrow historical practice to mindless formulae.

Although there was much agreement between teachers and researchers about the time necessary for alternative assessments, there was one striking difference between their responses: Only researchers speculated about the ability of teachers to teach the necessary knowledge, skills, and dispositions to undertake alternative assessments. Several researchers noted that, for teachers to implement alternative markers of historical knowledge, they would need “to know a lot about the process of historical inquiry to help kids,” “to be thoughtful and well-trained,” and to hold the “knowledge of history and historiography to facilitate.” By contrast, no teachers in the study reported feeling unprepared themselves or the possibility of other teachers being ill-equipped to develop and administer alternative markers.

So why might only researchers focus on disciplinary expertise? As noted above, teachers and researchers differed on the types of student thinking that alternative markers may encourage. Teachers focused on generic thinking and writing skills while researchers emphasized the habits of historians and authentic intellectual work. It may be that the teachers in the study do not possess a sophisticated disciplinary understanding of history,

therefore making it improbable for them to imagine authentic historical skills as a disadvantage for developing alternative assessments. However, the kinds of markers teachers suggested indicate interest in what might be considered more discipline-based forms of assessments such as having students interpret primary and secondary sources, analyze political cartoons, and collect oral histories. One other possibility for the discrepancy may be that, until recently, history-education researchers have primarily focused on disciplinary knowledge and pedagogical content knowledge as the key indicators for teachers engaging their students in authentic historical activities (Hartzler-Miller, 2001; McDiarmid, 1994; VanSledright, 1996; Wineburg & Wilson, 1991) .

* * * * *

The differences noted above represent a potential disruption in any emergent consensus across the teacher and researcher respondents. Although some of the differences may be a result of issues like slippery language (e.g., “analysis”), the fragility of a consensus around alternative approaches to assessment cannot be easily dismissed. That said, we remain convinced that both teachers and researchers see real possibilities for students to engage in alternative assessments.

Implications

Although no single consensus emerged on the use of alternative assessments, the potential for such a consensus seems evident. Given the presumption of “two cultures of social studies” (Leming, 1989), the coherence that surfaced across our teacher and researcher data is surprising and is all the more so given the open-ended nature of the survey. Teachers and researchers did display some differences, but we were struck by the general agreement on the kinds of alternative markers that might be developed, the ambitious directions those markers represented, the potential for distinguishing better and worse representations, and the possible psychometric problems. The only differences we could detect were in the areas of modifications and advantages and disadvantages. But even within these two areas, the differences lay side-by-side with several similarities.

In this study, we find no evidence for conclusions (a) that the groups are internally consistent in their views of assessment, (b) that any one group’s views are easily characterized, or (c) that the variation across groups is any less than that within the groups. There may not be a hard consensus in favor of one set of alternative markers, but neither are there any hard lines drawn between the groups surveyed. The uncertainty this situation offers may be read by some as discouraging news. But an alternative view suggests that the absence of hard distinctions between teachers and researchers offers the potential for an emergent consensus.

That consensus faces some real tests, however. First, no panacea exists—the participants offered no one best alternative assessment as each suggestion made came with clear advantages *and* disadvantages and the idea that every assessment would need some modification given different student populations.

Second, the imprecision of assessment language may be a sleeper issue. We found at least two instances in which language differences may cause problems. One instance arose around the terms used to label the assessments both groups offered. In short, one person’s “extended response” may be another’s “essay” and yet another’s “performance assessment.” A second

instance of slippery language surfaced around the difference between researchers' preference for assessing students' ability to think historically and the teachers' preference for assessing students' ability to engage in critical thinking. Such distinctions might dissolve were the two groups able to talk. But without some common meanings for the terms used, any consensus on alternative assessments may falter.

Third, the biggest potential threat to any consensus on alternative assessment markers is less about the markers themselves and more about the knowledge and skills necessary to devise such assessments and who needs to possess the knowledge to do so. A major disadvantage researchers identified was the need for teachers to hold substantial content and pedagogical knowledge, yet no teacher identified this weakness for themselves or their peers. This distinction mirrors the kind of intellectual divide that some (Leming, Ellington, & Porter-Magee, 2003) have argued exists between teachers and academics. Were this distinction to be the first point of conversation, we can imagine a quick and frustrated conclusion.

Finally, the notion of alternatives to typical state-level tests is as much a question of politics as it is pedagogy. There is no doubt that a clear consensus between teacher and researcher communities could influence the direction that state test programs take in the future. But the need to change current programs may not be self-evident to policymakers, legislators, and other officials, even if a professional consensus emerges. So building a constituency for change will be key.

References

- Austin, J. L. (1975). *How to do things with words*. Cambridge, MA: Harvard University Press.
- Baker, E. (1994). Learning-based assessments of history understanding. *Educational Psychologist*, 29(2), 97-106.
- Barton, K., & Levstik, L. (2004). *Teaching history for the common good*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Bernbaum, G. (1962). Language and history teaching. In W. H. Burston & C. W. Green (Eds.), *Handbook for history teachers* (pp. 38-46). London: Methuen.
- Bogdan, R., & Biklen, S. (1982). *Qualitative research for education: An introduction to theory and methods*. Boston: Allyn and Bacon.
- Cimbricz, S. (2002). State testing and teachers' thinking and practice: A synthesis of research. *Educational Policy Analysis Archives*, 10(2).
- Craig, C. (2004). The dragon in school backyards: The influence of mandated testing on school contexts and educators' narrative knowing. *Teachers College Record*, 106(6), 1229-1257.
- Fehn, B., & Koeppen, K. (1998). Intensive document-based instruction in a social studies methods course and student teachers, attitudes and practice in subsequent field experiences. *Theory and Research in Social Education*, 26(4), 461-484.
- Fuller, E. J., & Johnson, F. F. (2004). Can state accountability systems drive improvements in school performance for children of color and children from low-income homes? In L. Skrla & J. J. Scheurich (Eds.), *Educational equity and accountability: Paradigms, policies, and politics* (pp. 133-154). New York: Routledge.
- Gaudelli, W. (2006). The future of high-stakes history assessment: Possible scenarios, potential outcomes. In S. G. Grant (Ed.), *Measuring history: Cases of high-stakes testing across the U.S.* (pp. 321-334). Greenwich, CT: Information Age Publishing.
- Grant, S. G. (2003). *History lessons: Teaching, learning, and testing in U. S. high school classrooms*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Grant, S. G. (2006a). 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.
- Grant, S. G. (Ed.). (2006b). *Measuring history: Cases of high-stakes testing across the U. S.* 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.
- Grant, S. G., Gradwell, J. M., Lauricella, A. M., Derme-Insinna, A., Pullano, L., & Tzetzko, K. (2002). When increasing stakes need not mean increasing standards: The case of the New York state global history and geography exam. *Theory and Research in Social Education*, 30(4), 488-515.
- Grant, S. G., & Salinas, C. (in press). Assessment and accountability in social studies. In L. Levstik & C. Tyson (Eds.), *Handbook of research in social studies education*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Haladyna, T., Nolen, S., & Haas, N. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher*, 20(5), 2-7.
- Hartzler-Miller, C. (2001). Making sense of "best practice" in teaching history. *Theory and Research in Social Education*, 29(4), 672-695.

- Horn, C. (2006). The technical realities of measuring history. In S. G. Grant (Ed.), *Measuring history: Cases of high-stakes testing across the U.S.* (pp. 57-74). Greenwich, CT: Information Age Publishing.
- Kornhaber, M. L. (2004). Appropriate and inappropriate forms of testing, assessment, and accountability. *Educational Policy, 18*(1), 45-70.
- Leming, J. (1989). The two cultures of social studies education. *Social Education, 53*, 404-408.
- Leming, J. (1994). Past as prologue: A defense of traditional patterns of social studies instruction. In M. Nelson (Ed.), *The future of social studies* (pp. 17-23). Boulder, CO: Social Science Education Consortium, Inc.
- Leming, J., Ellington, L., & Porter-Magee, K. (2003). *Where did social studies go wrong?* Washington, DC: Fordham Foundation.
- Linn, R. (2003). Performance standards: Utility for different uses of assessments. *Educational Policy Analysis Archives, 11*(31). Available at: <http://epaa.asu.edu/epaa/v11n31>.
- McDiarmid, G. W. (1994). Understanding history for teaching: A study of historical understanding of prospective teachers. In M. Carretero & J. Voss (Eds.), *Cognitive and instructional processes in history and social sciences* (pp. 159-185). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Messick, S. (1988). Assessment in the schools: Purposes and consequences. In P. Jackson (Ed.), *Contributing to educational change: Perspectives on research and practice* (pp. 107-125). Berkeley, CA: McCutchan.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage.
- Newmann, F., Bryk, A., & Nagaoka, J. (2001). *Authentic intellectual work and standardized tests: Conflict or coexistence?* Chicago: Consortium on Chicago School Research.
- 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 Journal, 32*(1), 185-223.
- Rogers, V., & Stevenson, C. (1988). How do we know what kids are learning in school? *Educational Leadership, 45*, 68-75.
- Schug, M. (2003). Teacher-centered instruction: The Rodney Dangerfield of social studies. In J. Leming, L. Ellington & K. Porter-Magee (Eds.), *Where did the social studies go wrong?* (pp. 94-110). Washington, DC: Fordham Foundation.
- Searle, J. R. (1989). *Speech acts: An essay in the philosophy of language*. New York: Cambridge University Press.
- Shaver, J. (1977). A critical view of the social studies profession. *Social Education, 41*, 300-307.
- Shaver, J. (1981). Citizenship, values, and morality in the social studies. In H. Mehlinger & O. L. Davis (Eds.), *The social studies (80th yearbook of the National Society for the Study of Education, part II)* (pp. 105-125). Chicago: University of Chicago Press.
- Slavin, R. (1996). Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary Educational Psychology, 21*(1), 43-69.
- VanSledright, B. (1996). Studying colonization in eighth grade: What can it teach us about the learning context of current reforms? *Theory and Research in Social Education, 24*(2), 107-145.
- VanSledright, B. (2002). *In search of America's past*. New York: Teachers College Press.
- VanSledright, B. (2004). What does it mean to think historically and how do you teach it? *Social Education, 68*(3), 230-234.

- VanSledright, B., & Grant, S. G. (1991). Surviving its own rhetoric: Building a conversational community within the social studies. *Theory and Research in Social Education, 19*(3), 283-304.
- Wineburg, S., & Wilson, S. (1991). Subject matter knowledge in the teaching of history. In J. Brophy (Ed.), *Advances in research on teaching* (Vol. 3, pp. 305-347). Greenwich, CT: JAI.
- Yeager, E., & Wilson, E. (1997). Teaching historical thinking in the social studies methods course: A case study. *The Social Studies, 88*, 121-126.
- Yeh, S. (2001). Tests worth testing to: Constructing state-mandated tests that emphasize critical thinking. *Educational Researcher, 30*(9), 12-17.

¹ Hereafter, when referring to state-level tests, we use the terms “history” and “social studies” interchangeably as many state exams assess more than historical knowledge and understanding (Grant, 2006b).

² Newmann, Bryk, and Nagaoka (2001) find that differences between teachers and researchers are not unique to social studies.

³ In future studies, we intend to study three additional groups—policymakers, historians, and students—with a similar set of questions.

⁴ We coded as “non-traditional, short-answer questions” those suggestions that indicated a more sophisticated student response was desirable. For example, one researcher noted that, after reading a historical source, students could be asked to explain why the piece is “historically significant...not just a short identifying phrase.”

⁵ We coded as “non-traditional, multiple-choice questions” those suggestions that used language that communicated something more than the typical short, knowledge-level, multiple-choice questions. For example, one researcher called for “non-standardized, multiple choice,” while another referenced the more sophisticated multiple-choice questions Stuart Yeh (2001) has promoted.

⁶ Identifying a suggested assessment task as associated with either teachers or researchers or with both groups was made difficult by the sometimes small number of responses. In general, we used the following coding criteria: (1) Tasks offered by only one group were coded as falling under that group (e.g., “extended response” was only mentioned by researchers so it is coded R), and (2) tasks offered by individuals from both groups were assigned to both groups and coded T/R.

⁷ See Grant (2007) for a discussion of the problems language offers in the assessment of students’ historical knowledge and understanding.