Theory and Research in Social Education
Fall 2011, Volume 39, Number 4, pp. 494-540

Evaluating American History Teachers' Professional Development: Effects on Student Learning

Susan De La Paz
Nathaniel Malkus
Chauncey Monte-Sano
Elizabeth Montanaro
University of Maryland

The United States government has invested nearly one billion dollars in funding to professional historians and history educators across the country since 2000 to strengthen the teaching of American history in elementary and secondary schools, yet we know little about how these programs impact student learning. Using data from one such Teaching American History (TAH) grant, the authors employ multilevel models to investigate the effects of professional development on students' written responses to document-based questions at the fifth, eighth, and eleventh grades, and qualitative analyses of teachers' activities to learn about connections between classroom lessons and student outcomes. Findings indicate that fifth and eleventh grade students whose teachers were involved in ongoing networking activities for at least 30 hours in one year resulted in improved student performance. In addition, during the year, teachers of successful students engaged in activities that allowed them to increase their content knowledge and broaden their approach to teaching with primary documents.

KEYWORDS: *Teaching American History* (TAH) grant program, teacher professional development, document-based questions, historical argumentation, history education

Since 2000, the U.S. Department of Education has awarded millions of dollars in Teaching American History grants to school districts and their partners in an effort to develop teachers' knowledge of U.S. history. Yet, we have little data as to whether these professional development programs actually have a positive impact on students' learning. At the 2009 Annual Conference of the Organization of American Historians, Sam Wineburg critiqued the TAH program for its weakness in producing knowledge on what works in professional development

for history teachers (Shenkman, 2009). In this article, we set out to contribute to the literature on effective professional development by investigating the impact of one Teaching American History (TAH) program on students' learning.

Background

The emergence of TAH grant projects, created in 2000, has provided a viable professional development (PD) opportunity for social studies educators for the past several years. Recent reports (e.g., Westhoff, 2009) describe in detail the results of collaborations between university historians, teachers, and history educators. Westhoff compared teachers' use of primary sources with the ways that a historian might use the texts, explaining recurring difficulties that teachers had in gaining disciplinary understandings; she found missed opportunities in the lessons for students to engage in historical thinking with primary sources. She calls for greater use of analytic frameworks, such as Mandell's "Thinking Like a Historian" (2008), along with time and opportunities for teachers to use primary sources and engage in historical interpretation. Abt-Perkins (2009) also summarizes a TAH grant project, taking a broader perspective on the collaboration that she was involved in by describing the role, length of involvement, types of responsibilities, and evolution of pedagogical principles that guided both teachers' and historians' thinking. Of particular interest here is the organic nature of the grant activities over time—historians changed in their perspectives of how they interacted with and supported teachers, and teachers gained a much deeper appreciation for teaching about historical information as well as an understanding of how to situate and analyze historical sources.

TAH grants have primarily focused on enhancing teachers' content and pedagogical knowledge. Evaluation of their effectiveness initially did not go beyond teacher self-reports (Humphrey, Chang-Ross, Donnelly, Hersh & Skolnik, 2005). Even now that project evaluations examine student learning, little has been written about the impact of TAH on students. Published reports translate processes that trained historians engage in when thinking about the past, in an effort to help both teachers and their students understand and engage in disciplinary thinking (e.g., Mandell, 2008; Mucher, 2007).

TAH evaluations also report teachers' feelings, attitudes, and stages of concern about instructional strategies such as the use of primary documents (Ragland, 2007) and activities that are included in PD (Hall & Scott, 2007). Kortecamp and Steeves (2006) give teachers' perceptions of their impact on students, noting that teachers feel more confident and knowledgeable about both content and pedagogy. In particular, teachers report having "deeper discussions of content fol-

lowing analysis of primary sources, greater student engagement when using primary sources and web-based resources, and positive student feedback" (p. 502). Wayne, Yoon, Zhu, Cronen, and Garet (2008) posit that having a content emphasis is especially important, and that there is much consensus for "sustained" and "intensive" PD – yet more rigorous research designs are needed to determine the "relative effectiveness of PD programs with different durations or different allocation of PD events across time" (p.470). Conversely, Lai, Kearney, and Yarbrough (2009) suggest that *impact evaluations* intend to investigate whether an outcome such as improved student or teacher performance would have occurred in the absence of a grant. Such evaluations require mixed methods or experimental designs with one or more control or comparison groups.

Outside of TAH programs, conclusive research on teacher professional development is limited. Over 10 years ago, Wilson and Berne (1999) argued that relatively little is known about teachers' learning, save that they are able to continue learning about their subject matter, how to teach it, and how to support students' disciplinary thinking. Calls to reform professional development efforts from the late 1990s (e.g., Ball & Cohen, 1999; Hawley & Valli, 1999) are replete with recommendations for systemic, comprehensive overhauls in the ways teachers engage in deep reflection about student learning and the content that they teach. Ball and Cohen speak of the need for teachers to learn from practice and for teachers to use such knowledge to improve their practice—as teachers learn to interpret meanings in student work samples, as well as what happens in the classroom. Hawley and Valli (1999) argue that effective PD "is continuous and ongoing, involving follow-up and support for further learning, including support from sources external to the school," (p.141); moreover, they recognize that teachers need "to adapt their learning to their own students and contexts," (p. 143) to move from developers' suggestions to meaningful and sustained implementation of new practices.

Wilson (2009) led a National Academy of Education committee to investigate what we know about fostering teacher quality at this point in time. According to Wilson and her colleagues, PD has been found to be effective when it enhances teachers' subject matter knowledge, provides extended learning time, actively engage teachers, and links to what teachers are asked to do. Recently, too, teacher educators have argued that teachers must have opportunities to learn and practice the specific teaching strategies that will enable them to support their students' subject matter understandings (Ball & Forzani, 2009; Grossman, Hammerness, & McDonald, 2009).

Yet, attempting to link teacher learning to what they later do in the classroom, and then linking what they do to what their students learn, is a complex proposition that remains elusive in much of the PD literature

(Darling-Hammond, 2006). However, some PD evaluation literature has begun to appear with student outcome data (e.g., Ingvarson, Meiers, & Beavis, 2005; McCutchen et al., 2002) demonstrating that changes in teacher knowledge and classroom practice can improve student learning. Van Hover (2008) explains that in social studies, we have not seen such changes. Instead, she argues studies of social studies professional development tend not to link professional development activities with teacher learning and student outcomes. In this article, we seek to contribute to the current professional development literature by linking social studies teachers' PD activities with their students' learning outcomes.

If research on professional development literature is to link with evidence of student outcomes, then student outcomes must be clearly defined. Studies that focus on assessment of student learning in social studies and history range from reports that summarize student learning over time (e.g., from the National Assessment of Educational Progress) to investigations that analyze student learning over time in instructional contexts (cf. Monte-Sano, 2008; Young & Leinhardt, 1998). We were particularly interested in having students complete written assessments because writing history essays helps adolescents think deeply about historical content (Wiley & Voss, 1999), in contrast to assessments that focus on recall of more basic information. Wineburg (2004) argues that multiple-choice exams that emphasize factual recall have been consistently unsuccessful in capturing student achievement and are inauthentic measures of historical understanding. Therefore, the student outcomes we aimed for involved assessing students' historical thinking via their written arguments (cf. Monte-Sano, 2010).

Writing argument has been compared with other writing genres (e.g., narration and description) and found to promote greater audience awareness and syntactic complexity in high school students (Crowhurst & Piche, 1979). We believed that writing arguments, in particular, may actually help students integrate historical content because students must interpret and organize information from historical documents in a new way (Newmann, 1990). Results from one study with high school students provide support for this idea. Stahl and his colleagues found that writing arguments helped tenth grade students become better at historical reasoning about controversial topics (Stahl, Hynd, Britton, McNish, & Bosquet, 1996). Although writing argumentative essays, in themselves, do not always promote disciplinary thinking (cf. Grant, Gradwell, & Cimbricz, 2004), we chose this genre for assessment in our study.

Researchers have identified the kinds of teaching and activities that tend to foster historical thinking and argumentative writing. Reformers argue that analyzing evidence, developing arguments and conveying interpretations in writing support students' approach to

learning about history as an inquiry into the past by (Bain, 2005; Holt, 1990). In developing students' historical thinking, the kinds of texts with which students work can influence their reasoning processes. For example, students are more likely to think analytically and interact with texts if they read primary documents (Rouet, Britt, Mason, & Perfetti, 1996) or documents with a "visible author" (Paxton, 2002). Further, writing argumentative essays from multiple historical texts has been shown to help advanced students: (a) progress from listing information to synthesizing texts into an argument (Young & Leinhardt, 1999), and (b) develop deep understanding of content (Voss & Wiley, 2000; Wiley & Voss, 1999). Explicit instruction in historical thinking and writing helped middle and high school students with a wide range of incoming skills produce more accurate and persuasive history essays (De La Paz, 2005; De La Paz & Felton, 2010).

Theoretical Framework

Given this literature, the first author and her project collaborators tailored the PD to foster the kinds of teaching strategies that research has found to be supportive of developing students' historical writing. That is, the PD emphasized teaching history as an interpretive discipline in which one develops arguments about the past that are grounded in multiple sources of evidence (Bain, 2005; Holt, 1990; VanSledright, 2002). As such, teaching strategies that involve analysis of primary documents and writing arguments (e.g., De La Paz, 2005; Monte-Sano, 2008; Nokes, Dole, & Hacker, 2007; Wiley & Voss, 1999; Young & Leinhardt, 1998) were foregrounded in the professional development. Consistent with the literature on teacher learning, the PD offered extended time for teachers to participate, opportunities to develop their subject matter knowledge, and examples of specific practices that support teaching with documents and teaching writing (e.g., Ball & Forzani, 2009; Grossman et al., 2010; Wilson & Berne, 1999; Yoon, Duncan, Lee, Scarloss, & Shapley, K., 2007).

Research Questions

In the current study, we wanted to understand relationships between the degree to which teachers participated in our PD and student learning over one academic year. We examined depth and quality of teacher participation by logs that revealed time spent in a variety of PD activities and by what teachers reported that they were doing when engaged in PD activities. Through statistical analyses of teachers' data, we found that our teachers' level of involvement split evenly into two subgroups, one group of teachers who chose to engage in PD more often and another group of teachers who engaged less. Once we found this

marker, which was 30 hours of PD, we used this point of comparison for evaluating students' learning outcomes, as measured by analyses of their written responses to historical questions. We also took a close look at our observations of teachers and information they gave us about their teaching to infer how their teaching activities related to our analyses of student learning. Therefore, to assess the impact of the PD activities on students' learning, we pursued three specific research questions:

1. Did students whose teachers participated in our grant-related follow-up activities (networking group) improve more in their written responses to document-based questions than students whose teachers who had no follow-up participation (PD only group)?
2. Did students whose teachers invested more time in the follow-up activities (high-networking group) have students with improved scores compared to teachers who did not participate in any follow-up activities (PD only group) or who participated minimally (low-networking group) in the follow-up activities?
3. How do classroom observations and self-report data from teachers who participated in the PD relate to results about student learning?

Method

Participants and Setting

The teachers and students in this study were located in northern California in five school districts. A TAH grant, awarded to the first author and a local educational agency (a high school district), allowed us to invite teachers from this district and local elementary and middle schools to our first summer workshop. Our recruiting efforts gave priority to teachers in public feeder districts; however, as space and resources permitted, we extended our reach to teachers of those grades in nearby school districts, especially those with connections to the institution at which the first author was then teaching.

The goals of our professional development were two-fold. First, we intended to provide teachers with information related to historiography and context, through common study of four central themes in U.S. history: the origin and development of the Constitution, the territorial expansion of the nation, the development of the national economy, and the peoples who created the American republic with a primary focus on the African American experience. Second, we wanted to provide information regarding the discovery, analysis, and interpretation of primary documents. To us, discovery included both literal retrieval strategies

(i.e., archival and web search) as well as the purpose of a source, which includes an author's perspective (e.g., an opinion of a participant in the Constitutional Convention about the scope of executive power). Moreover, our goal was for teachers to situate the document historically in a broad context. We hoped teachers would learn to analyze the purposes for which the document might have been created, to whom the document was addressed, how we might think about its accuracy, and what it might tell us about the American experience. Finally, we stressed the importance of having teachers develop classroom experiences for students that integrated disciplinary and content knowledge. To further this goal, we presented argumentation as a foundational genre for writing in history in the workshop and presented teaching ideas for students at different grade levels. During our meetings, we planned for teachers to work together in groups (based on the grades they were teaching) to interpret the documents they had discovered and analyzed individually and decide how to present them in age-appropriate ways for their students.

A total of 53 teachers at three grade levels (fifth, eighth, and eleventh) participated in a four-day, summer workshop that took place before the instructional year. The first author and her colleagues designed the workshop to have a complementary focus on both subject matter and pedagogical content knowledge (Ball, 2000). In this workshop, historians shared content knowledge on one of four previously selected themes that were perceived to connect broadly across grade levels. Educators, including Sam Wineburg, author of *Historical Thinking and Other Unnatural Acts* (2001), modeled practices of "doing history" and suggested how to scaffold reading and writing activities in history to help students think and write in disciplinary ways.

Local university education faculty conducted two additional sessions on economic ideas as applied to the settlement of the Great Plains, and on teaching students strategies to read and analyze primary documents. Three additional sessions were devoted to locating, evaluating, and modifying primary documents using the Internet, reputable history databases, and traditional search engines. Two sessions were devoted to helping teachers incorporate active participation in American history by using historical simulations, period music, and role-play activities. The last two sessions focused on promoting critical thinking through moral dilemmas in history, cultural diversity, and the developmental needs of students.

Teachers. At the end of our summer workshop, we randomly assigned teachers to (a) networking group or (b) a group that would have no further involvement with the TAH grant for the current academic year. Using data from a survey that teachers completed in an application to our program, we matched teachers within each grade

level, using years teaching, credential type, and background in history education as the primary pairing indicators. Interest in U.S. history and gender were secondary variables. After matching, we sorted pairs using a computer-generated list of random numbers, and we assigned the second person to the group that would have no further involvement in TAH activities. It is important to note that each person had an equal chance of being assigned to either group.

Unfortunately, after matching and random assignment, school administrators reassigned three teachers from the networking group and two teachers from the PD only group. Two more teachers from the same group withdrew from the project because they were unwilling to participate in follow-up meetings or did not want students to miss four days of instruction to complete our assessments, and at the end of the project we eliminated one teacher from the PD only group after learning that the conditions under which our assessments were administered differed greatly from our protocol. After these changes, there were 25 teachers remaining in the networking group (11 fifth-grade, 4 eighth-grade, and 10 eleventh-grade) and 20 teachers in the PD only group (9 fifth-grade, 5 eighth-grade, and 6 eleventh-grade).

Networking group. Teachers assigned to this group received ongoing PD activities throughout the rest of the academic year, because we believed that successful PD is continuous and ongoing (Ball & Cohen, 1999; Darling-Hammond, 1997). Moreover, because we recognized the difficulty of sustaining teacher change when teachers work in isolation (Sarason, 1990), we established grade-level cohorts that allowed teachers at each respective level to prioritize ways to align our disciplinary suggestions for teaching historical thinking with standards they were expected to follow. We believed this would give teachers an opportunity to directly apply what they learned to their daily practice.

The following support and activities were offered in grade-level cohorts throughout the school year: (a) seven network meetings that included presentations focused on American history and pedagogy, as well as opportunities for teachers to share ideas; (b) provision of up to 40 hours of paid time per semester so that teachers could develop lessons and carry out other activities to enhance their history instruction; (c) paid substitute teachers to enable teachers to observe other American history teachers or participate in other history-related activities away from their schools; and (d) assistance from two district librarians to help teachers locate documents and identify web sources for American history topics. We also established a yearlong goal for teachers in each grade cohort to individually or collectively develop a unit or lesson that would be presented to other teachers at our second summer workshop. The lesson plans were eventually disseminated to peers on CDs with references to additional materials such as websites, books, documents,

music, and costumes. Incidentally, these presentations were rated highly by new and returning teachers who attended the workshop that occurred after this study ended, including one session in which one high school teacher co-led a session with a university historian.

PD only group. Teachers assigned to this group received no additional mentoring or follow-up activities after the one-week workshop ended. They agreed to participate in the observations and have their students complete assessments. Finally, after the year in which this study took place, teachers in this condition were invited to attend the second summer workshop, and for PD activities that occurred during subsequent years of our grant. Many, but not all, took advantage of this opportunity. In addition, the focus of the second full year of the grant was teacher-to-teacher networking, and teachers who had been in the networking condition invited PD only teachers to participate in district and school events.

Students. In this study, we analyze work from students who participated in our TAH teachers' classrooms, and our analyses are based on the number of students who were present for two assessments, a pretest in the fall and a posttest in the spring. The number of students in our sample (representing 71-84% of the total population at each grade level) was as follows: Grade Five - 525 students, Grade Eight - 611 students, and Grade Eleven - 948 students. It should be noted that a larger sample (about 4, 000 students in 8 districts) benefited from our overall TAH grant activities.

Data Sources

Student-level data. Students completed a document-based writing question (DBQ), requiring them to read four or five grade-appropriate document excerpts and write an essay near the beginning (pretest) and near the end (posttest) of the year. When we developed a format for prompting students to write responses to primary sources, we considered the way the New York state exam is framed. As such, we adopted the practice of providing excerpts from 4-5 documents, with brief historical context, and reminded students to use "specific details from at least three documents" in an introduction, supporting paragraphs, and conclusion that "responded to the (historical) question." While critics (e.g., Grant et al., 2004) contend that the New York state exams typically call for students to describe or explain rather than craft an argument, our prompts avoided this problem entirely. In fact, we asked students to write argumentative essays for reasons stated earlier and because state language arts standards included this genre at each grade we were testing. After this decision, our university

historian, district librarians, and the first author developed the source content collaboratively. Finally, to help students with vocabulary, difficult words were italicized and synonyms were presented in square brackets (e.g., "*orator* [speaker]").

Different DBQ questions were written for each grade level and each time point assessment. The rationale for different DBQ questions was based on the chronological nature of the curriculum. Teachers assigned the assessments before students learned the topic in class (as opposed to afterwards) because we were unable to standardize or observe the nature of teachers' instruction, and we did not want that variable to have an unintended confounding effect on the results of our study. We reasoned that teachers who generally taught students to analyze and write with primary sources should have improved outcomes on these tasks, and hoped that prior learning in the curriculum would give students some understanding of the historical context from which these topics were situated.

Appendix A shows the writing prompts for all DBQs at each grade level. A sample DBQ document set for Grade 8 is shown in Appendix B. All essays were scored using a holistic rubric (ranging from 0 to 5; see Appendix C). This rubric was based on prior work by Ferretti, MacArthur, and Dowdy (2000), and was developed to gauge the writer's ability to (a) interpret the documents and incorporate outside information related to the documents; (b) provide a clear opinion on the topic; (c) support a position with accurate facts, examples and details; (d) weigh the importance, reliability, and validity of the evidence; (e) analyze conflicting perspectives presented in the documents; (f) weave documents into the body of the essay; and (g) include a strong introduction and conclusion. It should be noted that students' content was the primary basis for assigning scores, rather than format of the essay. Thus, students who were not as familiar with this type of writing at the beginning of the year were not likely to receive low scores simply because they had less experience in writing arguments. In addition, we felt this was a reasonable approach to rating students' written arguments, because it had been a focal point during our initial PD with all participating teachers in this study and because others (e.g., Monte-Sano, 2010) report that factual and interpretive accuracy, persuasiveness of evidence, sourcing of evidence, corroboration of evidence, and contextualization of evidence provide markers of historical thinking in writing.

In addition to rating the final essays, short responses to individual documents were also rated holistically (using a simplified rating scheme for each document) in an effort to determine whether students comprehended the documents. Thus, we had two dependent measures for each student: (a) an average comprehension score for each document set, and (b) a writing score for each essay. However, only the writing score for the final essay was analyzed as an outcome measure in the

present study.

The first author employed a series of undergraduate and graduate students majoring in history or earning master's degrees in education to score all essays over a two-year period. Because the original data set was so large, readers were hired for a semester or a year, to score essays on one document set at a time. We held a large group training session at the beginning of each new document-based scoring session, and developed benchmark criteria for scoring using sample student responses. This process took about 7-10 hours for each set of materials and was led by a social studies teacher who was earning a graduate degree in education. Undergraduate readers then were assigned to work in pairs (although they read each paper independently) to score and then reconcile ratings as needed. During training, we established a criterion for acceptable interrater agreement at either exact agreement or adjacent scores at 80% for each topic and grade level. Readers were instructed to average discrepant scores that were adjacent (within one point) and to resolve differences in scores that differed by more than one point by discussion. Final interrater agreement (within one point) was 95% at the fifth grade level, 90% at the eighth grade level and 95% at the eleventh grade level. This report is based on student data from 43 of the teachers in the study, including all but two fifth-grade teachers (20 classes overall), all but one eighth-grade teacher (24 class sections) and all eleventh-grade teachers (50 class sections).

Teacher-level data. We observed all 45 teachers once (see Appendix D for our observation protocol) and asked them to complete surveys regarding their teaching practices. An end-of-the-year survey asked teachers a variety of questions regarding their teaching practices, changes in teaching, and the nature and extent of teaching historical reasoning and writing to students. We also asked the group of networking teachers about the extent to which they benefited from engaging in the yearlong activities. Teachers in this group kept a log of the activities they engaged in during the year, when meeting in groups and working alone. Activities included primary document research, lesson preparation with primary documents, development of assessments that involved the use of primary documents, other American history preparation, etc.

During observations, which were scheduled in advance according to when the teacher wished to demonstrate a model lesson, observers took extensive field notes documenting the extent to which teachers were implementing something they had learned during the workshop, and used a checklist to gauge level and numbers of student engagement, lesson effectiveness, and types of activities used in the lesson. Observers attempted to check for all ties between the TAH workshop content and what they observed in the lesson, and rated the quality of

each activity on a scale from 0 to 5, in addition to writing notes with examples of teacher and student statements and questions.

Further, we provided statements to "anchor" what the ratings might look like. The highest rating, a "5" for the analysis of primary or secondary documents was, "Has students consider the author and helps students discover in-depth meaning in and across documents. Critiques source to create a more focused understanding" whereas a 0 was "Presents documents with little or no historical context or attention to author, source, or evidence. Students accept document as presented." Observers also made ratings regarding the number of students who were engaged in the lesson, the teacher's overall level of preparedness, and the degree to which the lesson met the TAH grant's objectives. Before leaving the classroom, teachers were asked questions regarding how the lesson had been influenced by the workshop, the degree to which it was typical, if there was anything about the particular class being observed that made it easy or difficult to teach American history, and so on.

Our analysis of teacher performance (in particular, our observations of their teaching) was guided by Smith and Niemi's (2001) findings on the connections between specific types of instructional activities that were used by teachers in history classes and positive student outcomes. These authors explored different patterns in instructional practices (as framed on a NAEP survey). They then tested for the effects of teachers' instruction on student NAEP achievement, using a path analysis strategy that controlled for background characteristics (such as whether students were taking AP history classes) that were known to correlate with academic achievement. Four instructional domains were found to be correlated with high student outcomes (a) writing complexity, (b) variety of readings (e.g., reading outside the textbook), (c) use of extensive student discussion, and (d) use of learning tools beyond traditional textbook materials. Their findings indicate that the strongest effect of the history curriculum was tied to the nature of the teachers' instruction, with the above four domains strongly related to higher student scores. Therefore, we chose to descriptively analyze common characteristics of effective teachers in our sample with respect to these domains as well.

Analytic Measures and Statistical Methods

We analyzed our data using two sets of hierarchical linear models (HLM) to account for the nested structure of our data. HLM is the preferred approach because statistical assumptions in regression techniques assume measurements are independent of one another. In reality, however, students in our study were grouped within classes and within teachers; however, there were not enough classes for each

teacher to use a three-level model (with students within classes, within teachers) so we chose to group students within teachers. We believed our PD would impact student learning through teachers, and thus the best way to satisfy the assumption of independence was to nest students within teachers, as opposed to nesting them within classrooms, some of which had the same teacher. This choice reduced the statistical power of our analyses and prevented us from capturing potential peer effects at the class level. However, given the limitations of our data, grouping at the teacher level is the best analytic approach.

A check on the reasonableness of our decision to use HLM is shown by the results in Table 1 where we report the intra-class correlation (ICC) for each grade level analysis. These values ranged from 0.19 to 0.27, indicating that between 19% and 27% of the variance in student scores is between groups (in other words, the degree to which a given student's score is due to his or her group membership alone rather than his or her ability on the dependent measure), and that multilevel modeling was warranted.

Student-level Analytic Variables

Having established a need for a multilevel approach, we completed three separate analyses, one for the fifth-, eighth-, and eleventh-grades for the first two research questions. For all analyses, the dependent variable was the students' standardized score on the spring DBQ. In all analyses, our dependent variables' reliability measures were 0.85 or above. In both sets of models we consider standardized measures of individuals' scores on the *individuals' pretest DBQ score* as a control for students' prior ability. The student's standardized pretest DBQ score serves as a measure of his or her initial performance, and captures differences between students on other measures not included in our data.[1] Therefore, in our model the individuals' standardized pretest DBQ scores function as a control (much like a covariate) for initial writing ability. This allows us to partition or separate the variance that is due to students' incoming abilities from further effects that might have been due to other factors during the academic year. We believed the pretest DBQ measure could be viewed as a measure of initial ability because it was given before networking activities began and it measured the student's performance before his or her teacher had the benefit of our PD. The standardized pretest and posttest essay scores are presented by group and grade level in Table 2. The table clearly shows that there are differences between teachers in their average pretest and posttest scores. These differences were found despite our attempt to randomize the follow-up PD, in part due to the relatively small number of teachers in our sample and in part due to administrators adjusting initial group assignments. These differences underscore the importance of

including individual and teacher level pre-test measures in our statistical models because failing to do so may lead to incorrect (over or underestimation) of the impact of the teachers' PD on learning outcomes with students.

Table 1
Sample Sizes by Teacher Group, Reliability, ICC and Variance Explained by Grade

		Grade					
Sample Sizes		Fifth		Eighth		Eleventh	
Total	Teachers	19		9		16	
	Students	500		611		948	
Any networking activities		PD only	Networking	PD only	Networking	PD only	Networking
	Teachers	10	9	5	4	6	10
	Students	254	246	375	236	280	668
High-networking activities		Low-networking & PD only	High-networking	Low-networking & PD only	High-networking	Low-networking & PD only	High-networking
	Teachers	13	6	6	3	10	6
	Students	327	173	472	139	502	446
Reliability		0.85		0.92		0.89	
ICC		19%		22%		27%	
Level 1 Variance Explained		14%		15%		16%	
Level 2 Variance Explained		40%		51%		73%	
Total Variance Explained		20%		23%		28%	

Note. Variance explained refers to the difference between fully unconditional models and our second set of models for teachers with a high level of participation in networking activities.

Table 2
Mean Pretest and Posttest Essay Scores By Grade and Networking/PD Only Group

Mean Essay Scores		Fifth		Eighth		Eleventh	
		PD only	Net-work-ing	PD only	Net-work-ing	PD only	Net-work-ing
Any net-working activities	Fall Pretest (standard-ized)	0.072	-0.027	0.029	-0.044	-0.213	0.118
	Spring Post-test (stan-dardized)	-0.059	0.079	0.058	-0.092	-0.292	0.134
		Low-Net-work-ing & PD only	High-Net-work-ing	Low-Net-work-ing & PD only	High-Net-work-ing	Low-Net-work-ing & PD only	High-Net-work-ing
High-networking activities	Fall Pretest (standard-ized)	0.095	-0.115	-0.073	0.251	-0.320	0.219
	Spring Post-test (stan-dardized)	-0.072	0.164	-0.100	0.341	-0.256	0.306

Note. The pretest and posttest scores in this table are standardized within each grade such that the average score in each grade is equal to zero with a standard deviation of one. In such a scale, approximately 68% of the students in each grade will score between plus or minus one standard deviation.

Teacher-level analytic variables. For our primary independent variables we created two binary variables that indicated teachers' level of participation in follow-up activities. To answer the first research question, the binary variable indicated whether or not teachers participated at all in the follow-up activities. As shown in Table 1, 9 fifth-grade teachers, 4 eighth-grade teachers, and 10 eleventh-grade teachers participated in the follow-up activities. To answer our second research question we created a binary primary independent variable indicating which teachers invested 30 or more hours in follow-up activities. Our rationale for the 30-hour marker was as follows. The number of hours that group of networking teachers reported in their activity logs had a roughly bimodal distribution, and was thus unfit for use as a continuous variable. The bimodal distribution showed two distinct groups of teacher participation (i.e., a high- and a low- networking group), which split roughly at 30 hours for the year.[2] The low-networking teachers (less

than 30 hours of follow-up activities for the year) averaged slightly less than 20 hours of follow-up activities, and high-networking teachers (30 or more hours) averaged roughly 40 hours of follow-up activities for the year. On this second independent variable (level of networking), teachers with low or no participation were assigned a value of zero, and high-networking teachers were assigned a value of one. As shown in Table 1, a total of 15 teachers were high-networking teachers.

To capture differences between groups, particularly to control for student assignment across teachers and for systematic differences that might influence students' scores (e.g., such as in schools where tracking is used to place students in academic content courses), we used *average pretest DBQ score* for a teachers' students as a teacher-level control variable. Our rationale for analyzing our data in this way was that classroom climate, management, and other dynamics all differ depending on the group. To wit, teachers sometimes recognize this effect, noting things like, "My first period class is much quicker [or slower] than my fifth period class." It is also possible to attribute differences in class performance due to teacher effects, such as when one teacher is much better (or worse) than another, for any of a variety of reasons. Regardless of the reason for the difference between a particular teacher's students, the average pretest DBQ score serves as a control for group effects on student learning. In the context of our study, we were more interested in the effect of the PD than either of these control variables. Our intent was to examine the degree to which our TAH program, as evidenced by PD networking activities, and the length and quality of teachers' involvement in them, had an effect on student learning, above and beyond the effect that existed from students' prior ability and the effect of being in a particular teacher's classroom.

Models. The parallel hierarchical linear models for the two sets of analyses only differed by the level of networking involvement. In the first set of models the level of networking indicated teachers who participated in any follow-up activities (PD only group vs. networking group), and in the second set of models we compared the level of networking involvement by contrasting high-networking teachers who spent 30 or more hours in follow-up activities with teachers who spent less than 30 hours in the follow-up activities including both the low-networking and PD only teachers. The level one, or student-level model is below.

$$\text{Level-1 student:} \quad \text{Posttest}_{ij} = \beta_{0j} + \beta_{1j} * \text{Pretest}_{1j} + r_{ij}$$

The posttest score for student *i* of teacher *j* is equal to the intercept (β_{0j}), which is the mean posttest score for the students of teacher *j*, and the slope (β_{1j}) which is the average relationship between students'

standardized pretest score and the posttest score, and the student level error term (r_{ij}).

The level two, or teacher-level model used in these analyses takes the following form:

$$\text{Level-2 teacher:} \quad \beta_{0j} = \gamma_{00} + \gamma_{01} * \text{Average Pretest}_{1j} + \gamma_{02} * \text{Networking Group}_{1j} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

where β_{0j}, the intercept for the level one equation, is equal to the adjusted group mean posttest score (γ_{00}), and the average pretest score for teacher *j*'s students multiplied by the associated coefficient (γ_{01}), and the networking group effect multiplied by the associated coefficient (γ_{02}), and the teacher level error term (u_{0j}). The relationship between students' pretest and posttest scores was treated as a fixed effect and not allowed to vary across groups. Thus, in the second, level-two equation above for the fixed effect of students' individual pretest, β_{1j} from the level one model is equal to the average relationship between pretest and posttest scores for teacher *j*'s students (γ_{10}) with no associated random error term.

The resulting combined model is below:

$$\text{Posttest}_{ij} = \gamma_{00} + \gamma_{01} * \text{Average Pretest}_{1j} + \gamma_{02} * \text{High-Networking Group}_{1j} + \gamma_{10} * \text{Pretest}_{1j} + u_{0j} + r_{ij}$$

where students' posttest scores are a function of the adjusted group mean posttest score, the average pretest score for a teacher's students, the high-networking group effect, students' individual pretest scores, and a complex error term.

Results

Overview

In general, our results suggest that mere participation in the follow-up networking activities did not have a clear impact on student posttest scores, though the eleventh-grade analysis did reveal a significant difference in student posttest scores for participating teachers. In the analysis of teachers in the high-networking group, there were substantial effects for both the fifth- and eleventh-grade high-networking teachers, and a much larger estimated effect for eighth-grade students, though the difference was not statistically significant. We discuss these results in detail below.

Impact of Networking

The results of the first set of analyses provides answers to our first research question, whether students whose teachers participated in *any* of our grant related follow-up activities improve more in writing essays in response to DBQs than students whose teachers who received only the four-day PD (see Table 3 for descriptive information). Across all the analyses, the individuals' pretest DBQ score variables contributed significantly to the students' posttest DBQ score (the outcome of interest) and indicate that a one standard deviation increase in the pretest DBQ score is associated with an increase in the spring DBQ score (effect size of .36 to .39). This means that students' incoming skills are positively and strongly related to their performance at the end of the year, which is intuitively meaningful even if not the focus of interest in this study. On the other hand, the average classroom pretest DBQ score was not a significant predictor of students' posttest scores (see Table 3), *after* controlling for students' initial writing ability; however, the coefficients are all in the expected positive direction.

Moreover, in the fifth- and eighth-grade analyses, merely participating in some networking activities did not significantly affect students' posttest DBQ scores. However, in the eleventh grade, students with teachers who participated in yearlong PD (as compared to week-long PD) had a substantial increase in their posttest DBQ scores with an effect size of .42.

Table 3
Results of Networking Activities on Student Posttest DBQ Essays

Grade	Fifth		Eighth		Eleventh	
Intercept	-0.03		-0.03		-0.05	
Any Networking (vs. PD only)	0.18		-0.05		0.42	*
Average Classroom Pretest DBQ Score	0.2		0.29		0.21	
Individuals' Pretest DBQ Score	0.36	***	0.39	***	0.38	***

Note. *** = $p < .001$, ** = $p < .01$, * = $p<.05$. All predictors are grand mean centered.

Impact of High Networking vs. No or Low Networking

The results of the second set of analyses provide answers to our question regarding whether students whose teachers invested more time in the follow-up activities (at least 30 hours in one year, when given an opportunity to participate up to 40 hours per semester) have students with improved scores compared to teachers who participated minimally (less than 30 hours) or not at all in the follow-up activities (see Table 4). This model assumes that teachers who did not spend substantial time in the follow-up activities did not receive any additional benefit than the teachers who did not participate at all, and instead assumes that substantial investment in follow-up activities would be required to result in a payoff in the classroom. As in the first set of models, in the second set of models the individuals' pretest DBQ score variables are significant predictors of students' posttest scores across all analyses and are of similar magnitude as the prior set of models. Again, students' incoming abilities have some bearing on their end of year performance. This finding would not surprise most teachers. Also, the average classroom pretest DBQ score is not a significant predictor of student posttest DBQ scores (meaning the average classroom pretest score did not predict students' posttest scores, after controlling for their performance on the pretest), but the coefficients are again in the expected direction because higher group pretest scores did not have a negative association with posttest scores.

It is important to clarify that the results of this second set of analyses differs from the results of our prior analyses in important ways—when teachers participate in 30 or more hours of networking activities, students' scores show a pattern of substantial improvement. In the fifth- and eleventh-grade analyses, a high level of time investments in networking activities are associated with moderate to large significant increases in posttest DBQ scores (Table 4). In the fifth-grade, students with teachers who participated substantially in the PD scored higher than students whose teachers received only the initial, week-long PD, or invested a low level of time in networking activities, by an effect size of .35. Though the p-value of this estimate is 0.64, beyond the 0.05 level of significance, with the marginal power of our sample this represents a substantial difference with a good deal of variation across classrooms (19%, see Table 1).

Eleventh-grade students whose teachers had high networking showed even greater gains at nearly half a standard deviation (.46; $p<0.01$). In sum, these are moderate to large coefficients and a comparison to the pretest effect illustrates their magnitude. In other words, in the fifth-grade level analyses, the benefit students derive from having a teacher who participates substantially in the PD activities is roughly equal to the effect of having a standard deviation higher individual

pretest score (0.35 vs. 0.38), and in the eleventh-grade, the effect is greater than a standard deviation increase in students' pretest scores (0.46 vs. 0.38). A standard deviation difference in pretest scores is a large difference, equivalent to the difference between an average score at the 50th percentile, and a score at the 84th percentile. The fact that the impact of high-networking PD is comparable to such a large difference in pretest scores indicates that the effects of high-networking PD on student outcomes are substantial.

While our analyses revealed significant effects for high-networking teachers, the size or magnitude of the differences are helpful to benchmark against some external frame of reference. Neither the standardized scale of the effect sizes nor the original metric for essay scores (0-5 per the rubric in Appendix C) provides an intuitive scale for comparison. We believe the high-networking effects in the fifth- and eleventh-grade analyses are substantial compared to the effects of students' pretest scores. Further, relative to Cohen's (1988) general benchmarks for effect sizes, the effects we found lie between the small and medium effects of 0.20 and 0.50. Other research on the relative magnitude of effects sizes suggests the effects we found are more substantial. For instance, Hill, Bloom, Black and Lipsey (2008) found average effect sizes in student learning in social studies for an entire school year were approximately 0.30 between fifth- and sixth-grade and even less between eleventh- and twelfth-grade. While we cannot draw any direct comparison between our effect sizes (0.35 and 0.46 in fifth- and eleventh-grade, respectively) and Hill et al.'s benchmarks, they do suggest the differences we found are important.

In the eighth-grade analysis the effect size for high-networking teachers is smaller than in other grades and not significant, but is positive and consistent with the other significant findings. The lack of a significant finding in the eighth-grade analysis may be explained by two factors. First, the eighth-grade had the fewest teachers and thus had poor statistical power compared to the other grades. With a limited number of teachers, only very large effect sizes (i.e., greater than 75% of a standard deviation) would be reliably detected. A second explanation may lie in the DBQ prompts for eighth grade students.

When developing the document-based questions, we considered the California content standards and attempted to create accessible writing prompts that would be supported by documents. We considered some topics, such as the Constitution, as too difficult for students to write about, due to archaic language as well as the difficulty of the underlying concepts (e.g., federalists' beliefs). However, despite our attempt to create equivalent tasks, the eighth-grade prompts might have been harder than those at other grade levels. To illustrate, the pretest question, "Did the U.S. Government have a reasonable (or unreasonable) argument for going to war with Mexico?" assumes that

students know the government's argument for going to war. In addition, the posttest question, "Which arguments do you find most persuasive – that secession was illegal or that it was justified?" assumes that students know arguments for and against the Civil War.[3] One reason the writing prompts might have been more accessible for students at the other grade levels (see Appendix A) is that students could technically answer both the fifth-grade pretest and the eleventh-grade posttest DBQ without reading the documents.

Table 4
Results of a High Level of Participation in Networking Activities on Student Posttest DBQ Essays

Grade	Fifth	Eighth	Eleventh
Intercept	-0.02	-0.03	-0.05
High Networking Teacher (vs. Low-networking & PD only)	0.35 †	0.25	0.46 **
Average Classroom Pre-test DBQ Score	0.06	0.17	0.15
Individuals' Pretest DBQ Score	0.35 ***	0.39 ***	0.38 ***

Note. *** = p<0.001, ** = p<0.01, * = p<0.05, †=p<0.10. All predictors are grand mean centered.

We also ran several alternatively specified hierarchical models using additional covariates. These included a measure of teacher performance from our classroom observation, a measure of the impact of the PD on instruction as measured by self reported changes in instruction from participant surveys, and several demographic measures including teacher experience, education and licensure. These covariates were not significant and made little difference to the rest of the analyses; we removed them from our final models.

One interpretation of the failure of these measures to reach significance is that the measures are not related to students' learning outcomes as measured by their DBQ posttests. However, another possibility is that the lack of significance for these variables may be due to inadequate statistical power, that is, our sample was simply too small to capture these relationships. Another reason why these alternative models showed no effects for these teacher and classroom attributes may be that the measures we used to capture the attributes were inad-

equate. With our data there is no way to determine why these measures showed no significant relationship with students outcomes, but the lack of significance should not be construed as evidence that these teacher and classroom attributes are not related to student outcomes. Given our small sample, the observation and survey data remain worthy of future exploration. As such, we present the following descriptive information as qualitative support for findings that have already been identified.

Teacher Observations

As indicated in our third research question, we wanted to know more about the types of activities among networking teachers with differing levels of involvement. Therefore, we explored teacher observations, surveys, and activity logs to understand the kinds of practices common to teachers who engaged in extensive PD and if those practices were different than networking teachers with little involvement in the follow-up networking activities. We analyzed observation records for 14 of these highly engaged networking teachers and eight networking teachers with low levels of participation[4] using Smith and Niemi's (2001) four instructional domains (writing complexity, reading depth, use of extensive student discussion, and use of learning tools beyond traditional textbook materials), to determine whether common practices could be identified in their lessons.

Our results indicated that the majority of teachers who were in the high-networking group used learning tools beyond the textbook. Of the 14 teachers identified as highly involved, 43% utilized technology and 72% referred to maps or primary source documents. These teachers also tended to have greater student involvement in their lessons, using both teacher-led lessons and student-centered activities. Only 28% of teachers in the high-networking group spent the majority of class time lecturing during the observed period, while 62.5% of the eight teachers in the low-networking group were observed using primarily teacher-led lecture.

The 14 highly involved teachers engaged all or nearly all of their students in activities such as reading and analyzing primary source documents, student presentations, and role-play. One eleventh-grade teacher engaged his students with role play involving individuals in the Civil Rights Movement, while a fifth-grade teacher had her students analyze a letter written by Thomas Jefferson to the Senate regarding the Louisiana Purchase and discuss the point of view and time period in which the letter was written. Another fifth-grade teacher asked her class to write a journal entry from the perspective of Native Americans watching Christopher Columbus and his crews come ashore.

Survey and Activity Log

We used two self-report measures from teachers to gain an appreciation of the teachers' beliefs regarding the value of the networking activities, the degree to which they engaged in teaching activities we prioritized (such as analysis and writing with primary documents), and an overall description of what they did with their time when engaged in networking activities. Results from the survey (see Table 5) show differences in the frequency and methods that teachers used to teach with primary source documents. Moreover, teachers who were in the high-networking group were more likely to report teaching specific lessons on the analysis of primary source documents, and lessons that focused on writing from documents. To illustrate, 71% of the teachers who were more involved in TAH PD activities were likely to teach specific lessons that focused on writing from documents as compared to only 43% of the teachers who were in the low-networking group. On the other hand, all seven of the teachers with little networking involvement reported using primary sources at least every 2-3 weeks, whereas only 79% of the 14 teachers with extended networking reported the same frequency of use. These results may reinforce the idea that what one does with primary source documents in the classroom is more important than simply presenting them to students in lessons. In other words, the use of primary source documents alone is not sufficient to develop students' ability to write historical arguments.

Results from the activity log (see Table 6) show the average number of hours spent by teachers on targeted activities in addition to the time they spent in cohort meetings. The average differences in hours in each category between teachers who participated substantially in follow-up activities and those who did not show how much time it takes beyond meetings to put ideas about historical thinking and writing into practice. Noticeable differences are shown in each targeted category: (a) engaging in research to find primary documents, ostensibly to gain content knowledge about American History topics; (b) preparing to teach lessons that involve primary documents; (c) preparing assessments that incorporate the use of primary source documents; and (d) other lesson plan development, on American History topics in general.

Table 5
Percentage of Teachers With Reported Changes in Teaching

	< 30 hours networking (n=7)	> 30 hours networking (n=14)
Change how frequently you taught with primary source documents?		
Quite a bit	14%	50%
Somewhat	57	50
Not at all	28	0
Change the way you taught with primary source documents?		
Quite a bit	28	50
Somewhat	28	50
Not at all	43	0
How often did you use primary documents with your students this year?		
Weekly	33	29
Every 2-3 weeks	66	50
Once or twice a semester	0	21
Have you taught specific lessons that focused on analysis of documents?		
Yes	71	100
No	28	0
Have you taught specific lessons that focused on writing from documents?		
Yes	43	71
No	57	29

Table 6
Reported Activities Among Teachers with Low and High Levels of Networking Involvement

	< 30 hours networking (n=7)	> 30 hours networking (n=14)
Average Number of Hours in Targeted Activities Over One Year:		
Primary Document Research	< 5	30
Prepare Lessons with Primary Documents	3	22
Develop Assessments with Primary Documents	< 1	3
Other American History Lesson Plan Development	1.5	4

Note. Percentage of time in targeted activities is comparable across groups.

Discussion

Results from our first set of analyses evaluated whether any (versus no) follow-up PD activities resulted in improved student performance. Significant findings were found with teachers at the eleventh-grade. Thus, at one grade level, some degree of participation in follow-up PD activities (as compared to PD that is delivered at a single time, without any follow-up) consistently related to improved student improvement. In contrast, the results from our second set of analyses show a stronger and more consistent pattern of improvement for students whose teachers participated on average about 40 hours in follow-up PD, compared to teachers who participated minimally, on average less than 20 hours, or not at all in such year long opportunities. In the fifth- and eleventh-grade analyses, students whose teachers had substantial involvement had scores with large and significant improvements, with effect sizes ranging from 0.36 to 0.46. The fact that these improvements match the effects of students' pretest performance underscores the powerful influence of significant follow-up PD activities. The results for eighth grade students were not statistically significant, but the effect sizes fit the pattern of results in other grades.

Taken as a whole, the results of our second set of models (as seen in Table 2) provide evidence that sustained investment in professional development, in this case a high level of networking activities,

led to improved student performance, presumably based on changes to instruction. This finding is consistent with other research about the links between professional development and teacher knowledge and practice in the classroom (Yoon et. al., 2007); thus time invested in PD activities allowed teachers an opportunity to make changes to improve their classroom teaching, which in turn benefited student achievement. Our second set of models also explained more variance in the data relative to the first set of models, and had better measures of model fit.[5] These models explain between 40 and 73% of the level 2 variance (i.e., differences in DBQ scores associated with classrooms), and between 14 and 16% of the level 1 variance (i.e., differences between individual students). Overall the models explained between 20 and 28% of the total differences between students' DBQ scores (See Tables 1 and 2).

These results show that teachers who participated in the TAH PD at least 30 hours in one year resulted in improved student performance on at least one outcome measure, a DBQ writing assessment, which is an increasingly common learning outcome in history classrooms. The gains we saw were also consistent with the focus of our PD (i.e., teaching about historical thinking and writing with primary source documents). Results from the HLM models establish the validity of the networking activities, and suggest a potential threshold for teacher involvement to translate into an impact on student performance.

Our qualitative findings add to these results by showing what teachers with high levels of networking did with their time and what their teaching was like in the classroom. The data from our observations, survey, and activity log provide converging evidence that teachers with sustained involvement used the networking group as an opportunity to invest themselves in activities that led to changes in their knowledge of content and pedagogy. Teachers searched for primary sources extensively, in addition to allocating time for planning how to use documents in lessons and assessments. When teaching, this group of teachers routinely found ways to engage students in the learning activities, and they incorporated maps and primary source documents in their lessons. They also believed they changed the way they taught because of the work they had done during PD. Finally, additional self-report data from both groups revealed that only teachers in the networking group taught students specific lessons on analysis of primary documents and writing from primary sources. To be prepared for these lessons, teachers with a high level of PD involvement also spent more time outside of meeting times meeting and discussing teaching ideas with other networking teachers at the same grade level. In sum, their investment of time focused on developing skills in teaching with primary documents, teaching writing, and assessment. This focus on the skills of practice is consistent with what reformers in teacher education are currently saying is the best plan to improving teachers' practice (Ball & Forzani,

2009; Grossman et al., 2009).

We realize that some readers may wonder whether our results might be influenced by the possibility that the teachers who self-selected to participate in follow-up activities to a great degree "bought into" the instructional suggestions that were provided in our PD, as well as the idea that those methods were somehow better aligned with our major outcome measure, the document-based question assessment. It is difficult to refute this possibility with absolute certainty, but the activities reported by teachers on their networking logs seem to suggest that teachers did not focus narrowly, or exclusively on tasks that were the focal point in this study. To illustrate, "primary document research" included searching for primary documents on websites, searching books for primary documents, and reading and evaluating primary documents. Moreover, "lesson preparation with primary documents" included integrating pictures as well as text, and integrating primary documents into content lectures.

Educational Importance

Few published reports examine student outcomes in TAH grants, for which the government has spent millions of dollars over the past seven years. Major outcomes are focused on teachers who are expected to benefit from additional history content knowledge, as well as information on teaching strategies for diverse student populations in their schools. In contrast, our project was designed to meet both objectives: We worked to encourage teachers to infuse primary sources and instruction in writing in their teaching, and to measure the effects of the professional development on tangible student learning outcomes. Clearly there may be other influences on the teachers and students during the year in which the study took place, yet the design and the analytic method provides some justification that grants of this nature can have a positive impact on student learning in measurable ways. We are also encouraged that our main findings—that a reasonable degree of professional development can have a substantial impact on subsequent student performance—is consistent with the results of others who have had considerably more experience in mentoring teachers to engage in new ways of thinking about their teaching.

If a recent review by Yoon et al. (2007) is representative of the literature on student outcomes, in five of six demonstrably good quality experimental or quasi-experimental studies using control groups with pre- and posttest designs that could evaluate impacts on student achievement, sustained and intensive PD was spread out over six to 12 months, ranging from 30 to 100 hours in total. The average level of PD offered in these studies that found significant student improvement was 49 hours of activities in a year. In comparison, in three studies in

which PD ranged from 5 to 14 total hours, no significant impacts on student outcomes were found. Our finding may be tempered by the realization that we did not fully achieve embedded professional learning (Wei, Darling-Hammond, Andree, Richardson, Orphanos, 2009), nor did we ask teachers to construct a shared responsibility for the work they were doing in their schools. Other recent calls for studying the effects of PD on teachers and students (Wayne et al., 2008) suggest additional design issues that we did not consider, such as theories of instruction and theories of teacher change, that might have produced even greater impacts from our PD activities had we attempted to include them in our TAH grant. We did achieve some broader objectives in subsequent years of our TAH grant, as teachers who worked in the networking group shared lessons and content knowledge with others in their own settings. However, during those final years of the grant we did not evaluate the effects of our PD on students.

Recommendations

TAH project directors and evaluators may appreciate knowing which networking activities appeared effective more than knowing the degree to which teachers should be involved in general. Therefore, we recommend first that teachers at each grade level should engage in a sustained focus on the analysis of primary source documents, which may begin by searching for and adapting relevant source materials (cf. Wineburg & Martin, 2009). Teachers can then meet in grade level groups to share the sources they plan to use, in the context of an entire lesson. Primary source excerpts may complement textbook excerpts, or replace the use of textbooks altogether, depending on the content teachers plan to teach. A primary goal is for teachers to situate and analyze historical documents with their students.

Second, many teachers will benefit from learning more historical content. This may be especially true for teachers at the elementary and middle grade levels if they do not have a background in history as part of their credential programs. When possible, we suggest that teachers discuss topics in their curriculum (in grade level groups) that they wish to explore. While it may be hard to hold meetings across school sites, one idea is to share content via email or a Google site, then to use a lesson study approach for real or virtual meetings. Conversely, high school teachers may be able to use department meetings or engage in self-study more easily, especially if they initially possess a fair degree of content knowledge.

Third, there may be more formal avenues to helping teachers access more content knowledge. We recommend asking American history professors to provide content on a select group of themes in American history, in a series of large group lectures, and to follow these meet-

ings with smaller grade-level meetings, to provide time for teachers to reflect on the content they have learned about and to decide how to continue their learning on each topic. This combination may energize teachers by giving them information from which to draw and to allow them freedom to work in teams on real-world lesson plans.

Finally, we suggest additional pedagogical assistance at the grade-level meetings, as teachers think about new content, as well as discuss ways to locate, adapt, and teach with documents. Teachers may then work in groups or work independently, to fully develop teaching and assessment plans. Scheduling a formal presentation venue may motivate teachers to work on lesson plan development because there will be an authentic audience to share materials. In our case, teachers' lesson plans were to be shared in a subsequent summer workshop— this was a way the participants could give back to other teachers.

In closing, as federal funding continues to be spent on projects similar to ours, we echo the call for leaders involved in professional development to collect a variety of information that demonstrates how their efforts influence student learning. This study is one effort to link student achievement and teachers' professional development delivered through a TAH program. Our two sets of analyses suggest that it is not enough to expect that a low level of participation in follow-up activities is sufficient to garner benefits for students. Only teachers who invested substantial amounts of time (in this case at least 30 hours), had students with significantly higher outcomes.

Appendix A

Document-Based Questions for Each Grade Level

	Pretest	Posttest
Fifth grade	Was the United States acting in the best interest of the Indians when they moved them from their land in the east to new land in the west?	Did Captain Preston order his soldiers to fire their bayonets at the colonists?
Eighth grade	Did the U.S. Government have a reasonable (or unreasonable) argument for going to war with Mexico?	Which argument do you find most persuasive – that secession was illegal or that it was justified?
Eleventh grade	Who had the better vision for improving the conditions of African Americans during the early 1900s, Booker T. Washington or W.E.B. Du Bois?	Should the United States have warned Japan before dropping the Atomic Bomb?

Appendix B

Sample Document-Based Question: Eighth Grade Posttest

Civil War

The following essay topic is based on the accompanying historical documents (1-5). Some of these have been edited for the purpose of this task. This task is designed to test your ability to examine and interpret the meaning of historical documents. To complete the task, you will first answer questions related to each document, and then you will write a final essay that uses important information from the documents that you have analyzed.

Directions:

- Write an introduction, supporting paragraphs, and a conclusion.
- Use specific details from at least three documents in Part A.
- You may include outside information that you have learned on this topic.
- Answers from Part A should help you write your essay.

Historical Background:
In 1860 and 1861, eleven southern states seceded from the Union claiming the government no longer protected their interests as members of it. In response to this, the North went to war against the South in order to preserve the Union.

Task:
For Part A, read each document carefully and answer the question after each document. Then, read the directions for Part B and write your essay.

For Part B, use your answers from Part A, information from the documents, and your knowledge of social studies to write a well organized essay. In the essay, you should answer the following question:

Which argument do you find most persuasive – that secession was illegal or that it was justified?

Part A
Document 1:

Dubuque Herald, 1860

The Constitution makes no *provision* [allowance] for secession.... Constitutionally, there can be no such thing as *secession of a State* [leaving the Union]. But it does not follow that because a State cannot secede constitutionally, it is *obliged* [required] … to remain in the Union....If for any *cause* [reason] the Government...should become *inimical* [opposed] to the rights … of the people, instead of giving protection to their persons and property … it is the …right of the people to change the Governmet regardless of Constitutions.

What then is the South to do? Suffer the *compact* [live with the agreement] which brought them into the Union to be *violated* [abused] with *impunity* [pain], and without *means of redress* [a way of complaining]...Who expects...the South to submit to all this?

1. According to this newspaper article, why did the South want to secede?

[six blank lines provided for each item in actual test]

Document 2:

Excerpt from Lincoln's July 4, 1861 speech to Congress:

…Our States have neither more, nor less power, than that *reserved* [given] to them, in the Union, by the Constitution -- no one of them *ever having been* [had ever been] a State ***out*** of the Union. The original ones passed into the Union even ***before*** they cast off their British colonial dependence…the "United Colonies" were declared to be "free and independent States;" but … the object … was not to declare their independence of ***one another***, …but *directly the contrary* [just the opposite], as their mutual pledge. The… *plighting* [agreement] of faith, by each and all of the original thirteen, in the Articles of Confederation… that the Union shall be *perpetual* [everlasting], is most conclusive. Having never been States …***outside*** of the Union, *whence* [from where did] this … claim of power to lawfully destroy the Union itself? … The States have their *status* [identity] **IN** the Union. If they break from this, they can only do so against law, and by revolution....

2. Why does President Lincoln believe that the states can't leave the Union?

[six blank lines provided in actual test]

Document 3:

South Carolina Secession Declaration Debate, Dec. 25, 1860

There is another evil, in the *condition* [present state of affairs] of the Southern towards the Northern States…Our ancestors not only taxed themselves, but all the taxes collected from them were *expended* [spent] amongst them. Had they submitted to…the British Government, the taxes collected from them would have been expended in *other parts of the British Empire* [other foreign colonies]. They were fully aware of the effect of such a policy in *impoverishing* [making poor] the people from whom taxes were collected…To prevent the evils of such a policy was one of the *motives* [reasons] which drove them on to revolution. Yet this British policy has been fully *realized* [put in place] towards the Southern States by the Northern States. The people of the Southern States are not only taxed for the benefit of the Northern States, but after the taxes are collected, three-fourths of them are *expended* [spent] at the North. This cause…connected with the operation of the… Government, has made the cities of the South *provincial* [poor]. Their growth is paralyzed; they are mere suburbs of Northern cities. The agricultural productions of the South are the *basis* [source of wealth] of the foreign *commerce* [trade] of the United States; yet Southern cities do not *carry it on* [participate]…

3. Why does this author believe the South has a right to secede?

[six blank lines provided in actual test]

Document 4:

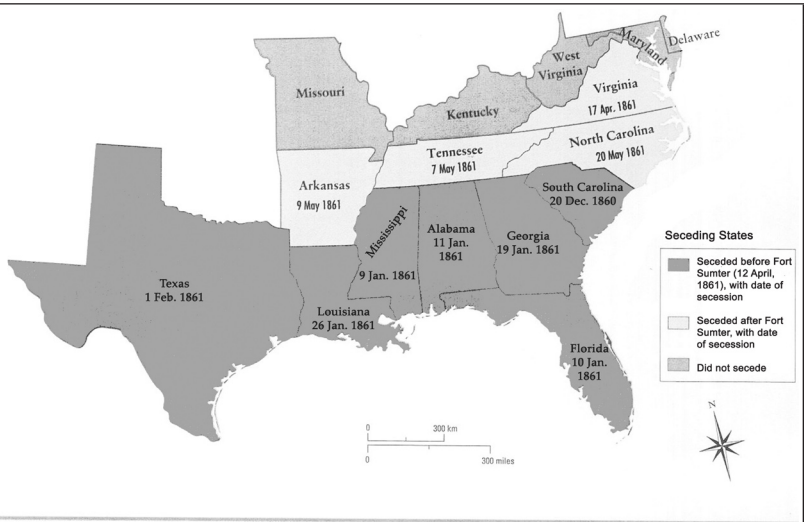
Andrew Johnson, Preserve the Union Speech, Dec 18, 1860

…*We* [Tennessee] deeply sympathize with our sister southern States, and … admit that there is good cause for dissatisfaction and complaint on their part… yet we, as a portion of the people of a slaveholding community, are not for seceding or breaking up the union of these States until every fair and honorable means has been exhausted...The people of Tennessee…declare further: That … no State has the constitutional right to secede from the Union without the *consent* [permission] of the other States which *ratified* [agreed to] the compact. The compact…. formed the Union without making any *provision* [allowance] whatever for its *dissolution* [termination]. *It* [the compact] was adopted by the States …***forever, "without reservation or condition**…"* [A] secession of one or more States from the Union, without the consent of the others…would be revolution, leading in the end to civil…war….We deny the right of a State…to secede from the Union, we admit the … right of revolution … but it is a right which should not be *exercised* [used], except in extreme cases, and in the last resort...

4. What reasons does Johnson provide for keeping the union together?

   [six blank lines provided in actual test]

Document 5: Map of the Secession Vote



Missouri, Kentucky, West Virginia, Maryland and Delaware did not secede from the union. Dates on the remaining states indicate when each voted to secede, with the main difference as before or after the fall of Fort Sumter.

5. Why do you think southern (border) states such as Maryland did not secede from the Union?

   [six blank lines provided in actual test]

Name:_____ Date:_____

Teacher:_____ Period:_____

Part B

Directions: Consider what life might have been like during the time leading up to the Civil War. Write an opinion essay in response to the following question:

Which arguments do you find most persuasive that secession was illegal or that it was justified?

In your essay, remember to:

- Tell which argument you find most persuasive – that secession was illegal or that it was justified?
- Include an introduction, body and a conclusion
- Include details, examples, or reasons to support your ideas
- Use the information from the documents in your answe

STUDENTS WERE GIVEN 3 LINED PAGES TO WRITE RESPONSES

Appendix C

Rubric for Rating Argumentative Essays (Grades 5, 8, and 11)

5 Exceeds Expectations

In addition to containing all of the elements of a strong (4) paper-
The writer deals with an opposing opinion with refutation or alternate solutions. Refutation explicitly recognizes opposing views and provides one or two reasons against those arguments. An alternate solution proposes a compromise position or alternative way of addressing the arguments of the opposition.
OR – The writer connects to broader context such as the influence of the Civil war or the entire time period in which these leaders lived
OR – The writer interprets quotes on a deep level, relating information to other facts.

4 Strong; Above Grade Level

Writes a well-developed essay, logical and clear plan of organization
Paper states a clear opinion and gives a context for that opinion
Analyzes and interprets 2+ documents (embeds reference into text)
All historical information is accurate and clearly relates to the question.
Discusses relevant facts, examples, and details
Essay is clearly written and coherent including an introduction & conclusion

3 Competent; Developed

States an opinion and gives reasons, plus elaboration of at least 1 reason. **OR**
States an opinion and includes 1 well-developed reason using information that could be convincing
OR Three or more reasons without elaboration
Refers to 1+ documents
Uses some facts, examples, and details, but is not analytic
Demonstrates a general plan; may have minor errors in use of document
May simply restate the opinion in the conclusion

2 Paper states an opinion and gives 2 reasons without elaboration. **OR**
Attempts to address some aspects of the task, without use of documents **OR** makes errors when using the documents
Presents few facts, examples, and details
Writes a poorly organized essay, lacking focus
Has vague or missing introduction and/or conclusion

1 Low, Minimally Developed or Undeveloped

Paper states an opinion, but the reasons are not explained. **OR**
Reasons are unrelated to, inconsistent with the opinion, or incoherent
Facts do not relate to the documents (i.e., relates to personal knowledge)
Essay demonstrates major difficulties in organization
Lacks introduction and conclusion

0 Not Rated

Completely ignores the question. **OR**
Paper does not answer the question, or provide an opinion on the issue.
Includes so many indecipherable words that no sense can be made
Ignores or misuses the documents
Lacks organization; little attempt made; blank paper

Appendix D

Teacher Observation Protocol

Teacher _____Grade/class _____ # of students_____
School _____Start time _____ End time _____

1. Check all materials used and briefly describe:
   - ❑ Textbooks_____
   - ❑ Other books _____
   - ❑ Documents_____
   - ❑ Objects (e.g., maps and artifacts)_____
   - ❑ Other (e.g.technology)_____

2. Subject of lesson:
   _____
   _____
   _____

3. California standards being addressed:
   _____
   _____
   _____

5. Check all ties to TAH workshop that you observe in the lesson. For each check, indicate the extent to which it is being used. (Descriptions are of teacher behaviors.)

   a. Analysis of primary and secondary source documents

5	4	3	2	1
Has students consider the author and helps students discover in-depth meaning in and across documents. Critiques source to create a more focused understanding.				Presents documents with little or no historical context or attention to author, source, or evidence. Students accept document as presented.

Comments:
_____
_____
_____

   b. Prewriting strategy for writing essays

5	4	3	2	1
Includes all or most of specific steps (e.g., STOP, DARE) to prepare students for writing.				Gives students writing assignments with few if any prewriting strategies; makes no special preparation for writing.

Comments:

_____
_____
_____

c. Evaluation of web search engine, web subject catalog, and/or a periodical database for use in US history lesson

5	4	3	2	1
Ensures that students select search engine, catalog, and database on basis of objective criteria.				Accepts any search engine, catalog, or periodical database for student use without questioning its accuracy or validity.

Comments:

_____
_____
_____

d. Evaluation of web page related to US history

5	4	3	2	1
Has students evaluate web pages related to US history to establish the validity of the source, the credentials of its author, the type of information, purpose, and timeliness.				Allows students to accept information from web pages related to US history, with little questioning of their source, authors, or purpose.

Comments:

_____
_____
_____

e. Use of objects as part of US history lesson

5	4	3	2	1
Integrates historical objects (e.g., pottery, quilts, pictures, other artifacts) into US history lesson and encourages to question who created the objects, why, and what they mean.				Shows artifacts but does not explain their historical importance clearly.

Comments:

f. Focus on multicultural issues in US history

5	4	3	2	1
Includes the experiences, contributions, and perspectives of a range of racial, ethnic, and cultural groups in US history.				Includes multicultural issues but does not link them to historical context and/or provides a biased perspective of a racial, ethnic, or cultural group.

Comments:

g. Focus on the U.S. Constitution

5	4	3	2	1
Brings in a social context (e.g., slavery, founding fathers) and goes beyond the text to include external factors.				Teaches the constitution as an authoritative text and/or merely focuses on specific powers it establishes (e.g., Congress).

Comments:

h. Inclusion of economics issues

5	4	3	2	1
Presents basic economics ideas and applies them to specific historical events.				Introduces economics ideas with no historical context.

Comments:

i. Focus on moral development

5	4	3	2	1

Presents clearly defined and conflicting moral principles related to US history; has students discuss, interpret, and resolve them.

Permits superficial discussion of moral dilemma that is unrelated to US history.

Comments:

_____
_____
_____

j. Use of role plays of historical figures/situations

5 4 3 2 1

Actively involves students in playing historical roles and encourages them to incorporate their own ideas and knowledge.

Provides students with scripts or has them perform limited, previously defined roles.

Comments:

_____
_____
_____

k. Use of period music and/or costumes

5 4 3 2 1

Uses period music and/or costumes to set context and extend understanding of historical events.

Plays period music or shows period costume with little or no explanation of its historical importance.

Comments:

_____
_____
_____

l. Use of JackDaws materials

5 4 3 2 1

Creates a lesson centered on more than one primary source. Asks critical thinking questions or create own worksheet or activity for students. May link to historical fiction or other reading assignments.

Posts primary sources on bulletin board but does not engage students in discussion regarding the content.

Comments:

_____

_____

_____

5. Student activity: Check all that occur during each 10-minute period. At the end of each period, indicate the most frequent activity for those 10 minutes by circling its check mark.

Student activity	1st 10 minutes	2nd 10 minutes	3rd 10 minutes	4th 10 minutes	5th 10 minutes	6th 10 min.-end
a. Listen to teacher	◎	◎	◎◎	◎	◎	◎
b. Listen to classmates	◎	◎	◎	◎	◎	◎
c. Read/examine content-related materials	◎	◎	◎	◎	◎	◎
d. Work individually at desk/table	◎	◎	◎	◎	◎	◎
e. Work individually at computer	◎	◎	◎	◎	◎	◎
f. Work collaboratively at desk/table	◎	◎	◎	◎	◎	◎
g. Work collaboratively at computer	◎	◎	◎	◎	◎	◎
h. Participate in class discussion (Indicate number of students)	◎ #_____	◎ #_____	◎ #_____	◎ #_____	◎ #_____	◎ #_____
i. Make small-group presentation	◎	◎	◎	◎	◎	◎
j. Make individual presentations, role play	◎	◎	◎	◎	◎	◎
k. Talk off subject to other students (Indicate number of students)	◎ #_____	◎ #_____	◎ #_____	◎ #_____	◎ #_____	◎ #_____
l. Tune out in nondisruptive way (Indicate number of students)	◎ #_____	◎ #_____	◎ #_____	◎ #_____	◎ #_____	◎ #_____
m. Wait for materials, instructions	◎	◎	◎	◎	◎	◎
n. Other	◎	◎	◎	◎	◎	◎

6. Overall, what percentage of the students were actively engaged in this lesson? (Check one.)
   - ❑ All or nearly all
   - ❑ Three-fourths or more
   - ❑ About half
   - ❑ Fewer than half

7. Overall, how well prepared was the teacher to present this lesson? (Check one.)
   - ❑ Very well
   - ❑ Well
   - ❑ Fairly well
   - ❑ Not well at all

8. Overall, how effective was this lesson in meeting the teacher's objectives? (Check one.)
   - ❑ Very effective
   - ❑ Effective
   - ❑ Fairly effective
   - ❑ Not at all effective

9. Overall, how effective was this lesson in meeting the TAH grant's objectives? (Check one.)
   - ❑ Very effective
   - ❑ Effective
   - ❑ Fairly effective
   - ❑ Not at all effective

10. Did any special circumstances affect this lesson (e.g., fire drill, interruptions)?
    - ❑ Yes—please describe the differences briefly:
    - ❑ No

Questions for the teacher (if time and situation permit):

1. (If unable to complete item 3 on observation form): Which California standards was this lesson designed to address?

1. In what ways, if any, was this lesson influenced by the TAH workshop?

2. Was this lesson fairly typical for this class?
   - ❑ Yes
   - ❑ No—please describe the differences briefly:

3. Is there anything about this class that makes it easier or more difficult for you to teach US history?
   - ❑ Yes—please describe briefly:
   - ❑ No

Notes

This research was supported by a Teaching American History grant (award no. U215X030180) from the U.S. Department of Education. We would like to thank Robert Senkewicz and Tom Savage at Santa Clara University, Phyllis DuBois of the American Institutes for Research, and Cathy Giammona and Ana Lomas at East Side Union High School District for their contributions to grant activities. We are grateful to the teachers and students who made this research possible.

[1] Other measures of prior academic performance (such as grades or performance on prior standardized annual assessments) or basic demographic information such as ethnicity or socio-economic status were not available for a large number of students in our sample, and thus not included in these analyses. However, our pretest measure captures the majority of variation between students upon entry into teachers' classes.

[2] To determine the appropriateness of using the 30-hour threshold for high participation teachers we ran the same models using 20- and 40-hour thresholds. In all cases the model fit was superior using the 30-hour threshold.

[3] In addition, this posttest question required students to understand two concepts—illegal and justified—rather than one as in the pretest (reasonable). Something could be illegal and justified so the prompt may have been unintentionally complex in comparison to the pretest which focuses on one idea—the reasonableness of the government's argument.

[4] We had observations for all but one of the teachers in the networking group. One observation for a highly involved teacher was lost.

[5] In other words, fitting the most precise and parsimonious equation to model the effects of networking leads to a better explanation of the different components (e.g., initial ability, class climate) that each contributed to student achievement, without including spurious measures.

References

Abt-Perkins, D. (2009). Finding common ground: Conditions for effective collaboration between education and history faculty in teacher professional development. In R. Ragland & K. Woestman (Eds.), *The Teaching American History Project: Lessons for history educators and historians* (pp. 202-215). New York, NY: Routledge.

Bain, R. B. (2005). "They thought the world was flat?" Applying the principles of how people learn in teaching high school history. In M. S. Donovan and J. D. Bransford (Eds.), *How students learn: History, mathematics, and science in the classroom* (pp. 179-213). Washington, DC: The National Academies Press.

Ball, D. L. (2000). Bridging practices: Intertwining content and pedagogy in teaching and learning to teach. *Journal of Teacher Education, 51*, 241-247.

Ball, D. L., & Cohen, D. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3-31). San Francisco, CA: Jossey-Bass.

Ball, D. L., & Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *Journal of Teacher Education, 60*, 497-511. doi: 10.1177/0022487109348479

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Crowhurst, M., & Piche, G. L. (1979). Audience and mode of discourse effects on syntactic complexity in writing at two grade levels. *Research in the Teaching of English, 13*, 101-109.

Darling-Hammond, L. (1997). *The right to learn*. San Francisco, CA: Jossey-Bass.

Darling-Hammond, L. (2006). Assessing teacher education: The usefulness of multiple measures for assessing program outcomes. *Journal of Teacher Education, 57*, 120-138. doi: 10.1177/0022487105283796

De La Paz, S. (2005). Effects of historical reasoning instruction and writing strategy mastery in culturally and academically diverse middle school classrooms. *Journal of Educational Psychology*, *97*, 137-156. doi: 10.1037/0022-0663.97.2.139

De La Paz, S., & Felton, M. (2010). Reading and writing from multiple source documents in history: Effects of strategy instruction with low to average high school writers. *Journal of Contemporary Educational Psychology, 35*, 174-192.

Ferretti, R. P., MacArthur, C. A., & Dowdy, N. S. (2000). The effects of an elaborated goal on the persuasive writing of students with learning disabilities and their normally achieving peers. *Journal of Educational Psychology*, *92*, 694-702.

Grant, S. G., Gradwell, J. M., & Cimbricz, S. K. (2004). A question of authenticity: The document-based question as an assessment of students' knowledge of history. *Journal of Curriculum and Supervision, 19*, 309-337.

Grossman, P., Hammerness, K., & McDonald, M., (2009). Redefining teaching, re-imagining teacher education. *Teachers & Teaching, 15*, 273-289. doi: 10.1080/13540600902875340

Hall, T. D., & Scott, R. (2007). Closing the gap between professors and teachers: "Uncoverage" as a model of professional development for history teachers. *The History Teacher, 40*(2), 257-263.

Hawley, W. D., & Valli, L. (1999). The essentials of effective professional development. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning professional: Handbook of policy and practice* (pp. 127-150). San Francisco, CA: Jossey-Bass.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172–77.

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

Humphrey, D., Chang-Ross, C., Donnelly, M. B., Hersh, L., & Skolnik, H. (2005). *Evaluation of the Teaching American History grant program*. Retrieved from Department of Education website: http://www2.ed.gov/about/offices/list/opepd/ppss/reports.html#tq

Ingvarson, L., Meiers, M., & Beavis, A. (2005). Factors affecting the impact of professional development programs on teachers' knowledge, practice, student outcomes, & efficacy. *Education Policy Analysis Archives, 13*(10), 1-28. doi: 10.1080/02619760701664151

Kortecamp, K., & Steeves, K. A. (2006). Evaluating professional development of American history teachers. *Theory and Research in Social Education, 34*, 484-515.

Lai, E., Kearney, J., & Yarbrough, D. (2009). How to evaluate Teaching American History projects. In R. Ragland & K. Woestman (Eds.), *The Teaching American History Project: Lessons for history educators and historians* (pp. 265-282). New York, NY: Routledge.

Mandell, N. (2008). Thinking like a historian: A framework for teaching and learning. *OAH Magazine of History, 22*(5), 55-62.

McCutchen, D., Abbott, R. D., Green, L. B., Beretvas, N., Cox, S., Potter, N. S., Quiroga, T., & Gray, A. L. (2002). Beginning literacy: Links among teacher knowledge, teacher practice, and student learning. *Journal of Learning Disabilities, 35*, 69-86.

Monte-Sano, C. (2008). Qualities of historical writing instruction: A comparative case study of two teachers' practices. *American Educational Research Journal, 45*, 1045-1079.

Monte-Sano, C. (2010). Disciplinary literacy in history: An exploration of the historical nature of adolescents' writing. *The Journal of Learning Sciences*, *19*(4), 539-568. doi: 10.1080/10508406.2010.481014

Mucher, S. (2007). Building a culture of evidence through professional development. *The History Teacher, 40*, 265-273. doi: 10.1080/15476880701309906

Newmann, F. M. (1990). Higher order thinking in teaching social studies: A rationale for the assessment of classroom thoughtfulness. *Journal of Curriculum Studies, 22*(1), 41-56.

Nokes, J. D., Dole, J. A., & Hacker, D. J. (2007). Teaching high school students to use heuristics while reading historical texts. *Journal of Educational Psychology, 99*, 492-504.

Paxton, R. J. (2002). The influence of author visibility on high school students solving a history problem. *Cognition and Instruction, 20*(2), 197-248.

Ragland, R. G. (2007). Changing secondary teachers' views of teaching American history. *The History Teacher, 40*, 219-246.

Ravitch, D. & Finn, C. (1987). *What do our 17-year-olds know?: A report on the first national assessment of history and literature.* New York, NY: Harper & Row.

Rouet, J. F., Britt, A., Mason, R. A., & Perfetti, C. A. (1996). Using multiple sources of evidence to reason about history. *Journal of Educational Psychology, 88*, 478-493.

Sarason, S. (1990). *The predictable failure of educational reform*. San Francisco, CA: Jossey-Bass.

Shenkman, R. (2009, April 19). OAH 2009: Sam Wineburg dares to ask if the Teaching American History Program is a boondoggle. Retrieved from George Mason University's *History News Network*: http://hnn.us/articles/76806.html

Smith, J., & Niemi, R. G. (2001). Learning history in school: The impact of course work and instructional practices on achievement. *Theory and Research in Social Education, 29,* 18-42.

Stahl, S., Hynd, C., Britton, B., McNish, M., & Bosquet, D. (1996). What happens when students read multiple source documents in history? *Reading Research Quarterly, 31,* 430-456.

van Hover, S. (2008). The professional development of social studies teachers. In L. S. Levstik & C. A. Tyson (Eds.), *Handbook of research in social studies education* (pp. 352-372). New York, NY: Routledge.

VanSledright, B. (2002). Confronting history's interpretive paradox while teaching fifth-graders to investigate the past. *American Educational Research Journal, 39*(4), 1089-1115.

Voss, J. F., & Wiley, J. (2000). A case study of developing historical understanding via instruction: The importance of integrating text components and constructing arguments. In P. N. Stearns, P. Seixas, & S. Wineberg (Eds.), *Knowing, teaching, & learning history: National and international perspectives* (pp. 375-389). New York: New York University Press.

Wayne, A. J., Yoon, K., S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: Motives and methods. *Educational Researcher, 37*, 469-479.

Wei, R. C., Darling-Hammond, L., Andree, A., Richardson, N., Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad.* Dallas, TX: National Staff Development Council.

Westhoff, L. (2009) Lost in translation: The use of primary sources in teaching history. In R. Ragland & K.Woestman (Eds.), *The Teaching American History Project: Lessons for history educators and historians* (pp. 62-78). New York, NY: Routledge.

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91*, 301–311.

Wilson, S. (2009). *Teacher quality: Education policy white paper*. Education Policy White Papers Project, National Academy of Education, Washington, DC. Retrieved from: http://www.naeducation.org/Teacher_Quality_White_Paper.pdf

Wilson, S., & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge: An examination of research on contemporary professional development. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (Vol. 24) (pp. 173-209). Washington, DC: American Educational Research Association.

Wineburg, S. (2001). *Historical thinking and other unnatural acts: Charting the future of teaching the past*. Philadelphia, PA: Temple University Press.

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

Wineburg, S., & Martin, D. (2009). Tampering with history: Adapting primary sources for struggling readers. *Social Education, 73 (5)*, 212-216.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & Answers Report, REL2007–No. 033). Retrieved from http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=70&productID=31

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

SUSAN DE LA PAZ is an Associate Professor in the College of Education, *University of Maryland*, College Park. She can be contacted at: sdelapaz@umd.edu

NATHANIEL MALKUS is a Doctoral Candidate in the College of Education at the *University of Maryland* and a Research Analyst at the *American Institutes for Research.* He can be contacted at: natmalkus@gmail.com

CHAUNCEY MONTE-SANO is an Assistant Professor in the College of Education, *University of Maryland*. She can be contacted at: chauncey@umd.edu

ELIZABETH MONTANARO is a Doctoral Candidate in the College of Education at the *University of Maryland*. She can be contacted at: bmont22@gmail.com