Evaluation of the Teaching American History Project: Grant Wood History Institute

Final Report, September 23, 2010

Submitted to:
Elise Fillpot, PhD. Project Director

Submitted by:
University of Iowa Center for Evaluation and Assessment
Julie Kearney, Emily Lai, and Don Yarbrough
210 Lindquist Center
Iowa City, IA 52242
Email: coe-cea@uiowa.edu
Website: http://www.education.uiowa.edu/cea/

The Grant Wood History Institute Project is supported by a grant award to Washington Community Schools, Elise Fillpot, project director, from the United States Department of Education Teaching American History Program. This evaluation was conducted under the auspices of a sub-contract between Washington Community Schools and The University of Iowa Center for Evaluation and Assessment and supported by funds from the above grant, number U215X060173.

Table of Contents

1. Executive Summary

This report is the final evaluation report for the GWHI Project (GWHI) that received funding in 2006-2009 as part of the U.S. Department of Education Teaching American History program. The project continued for a fourth no-cost extension year, but the evaluation concluded at the end of Year Three. The Center for Evaluation and Assessment (CEA) at the University of Iowa evaluated this project (and the previous Bringing History Home Projects, BHH and BHHII, on which it is based) throughout the grant period as a third-party evaluator. Evaluation methodology included observation of teacher professional development workshops, teacher content knowledge assessments, teacher lesson planning surveys, teacher workshop surveys, teacher implementation surveys, student content knowledge assessments, and student historical skills assessments.

The purposes of the GWHI project evaluation were to:

1) Contribute to project improvement
2) Document the actual activities and procedures of the project as implemented
3) Investigate changes over the three years in teachers' attitudes concerning teaching history, their historical content knowledge, their implementation of history curricula, and their use of new pedagogical methods advocated by the GWHI project
4) Examine students' historical content knowledge and historical thinking skills.

During the GWHI grant period, middle and high school teachers from the Grant Wood Area Education Agency (GWAEA), Iowa, received professional development centered around the GWHI paradigm for using historical skills as part of their historical pedagogy and as essential tools for students' learning of history. Approximately 72 teachers from 21 school districts took part in at least some aspect of the GWHI project. There were 36 teachers in Cohort 1 taking part in most of the activities during the first year of programming. During the second year, approximately 30 teachers from Cohort 1 continued and an additional 24 teachers (Cohort 2) joined the project and completed most of the activities during the second year. Some project participants were graduate students or were student teaching during the project, and thus did not have classrooms in which to practice the GWHI approach or collect student data. Additionally, some participants did not participate in enough of the programming to be included in the evaluation activities. The resulting core of 39 teachers participated in evaluation activities and provided curriculum and instruction in US history shaped around the acquisition and practice of the five main GWHI historical skills, 1) Timeline construction, 2) Mapping for understanding, 3) Primary source analysis, 4) Reading for background knowledge, and 5) Synthesizing knowledge from multiple sources.

The GWHI project provided twelve days of professional development workshops for participating teachers during the first two summers and academic years of the project, and ongoing

professional development support throughout the entire grant period. The project's professional development and formative evaluation activities that took place during Years 1 and 2 have been described in detail in previous annual evaluation reports submitted to the U.S. Department of Education. These reports are available at the CEA website at: http://www.education.uiowa.edu/cea/tah/index.html.

During the final year of the project, the primary focus of the GWHI project evaluation was to document and describe teacher and student outcomes. Changes in teachers' historical pedagogy, approach to lesson planning, content knowledge, and motivation to teach and learn history were examined through surveys and assessments throughout the project. Teachers who participated in the GWHI project reported in surveys that their teaching of historical content was more thorough after project participation prior to their involvement in the project. This was true for both middle school and high school teachers, and was also true for teachers whether or not they taught US history as their main assignment or simply as part of another course. Among middle school teachers, increases in thoroughness of teaching historical content were particularly marked if the content had been addressed during GWHI professional development.

Participating teachers also reported increased confidence in their students' abilities. They rated their students' knowledge of history content as higher after project participation than before and large increases were seen in their ratings of students' abilities to perform historical skills. On workshop surveys, participants were highly receptive of the professional development activities offered by the GWHI project and reported that both the content and the GWHI pedagogical paradigm were extremely useful in their teaching. Teachers indicated that it was beneficial for their students to learn history through use of the five skills advocated by the GWHI project. Teachers also reported that they were more motivated to teach and learn history and, to a lesser extent, said that their students were also more interested and motivated to learn history than before their project participation.

Teachers' historical content knowledge before and after project participation was assessed using a measure composed of multiple choice and short answer items; some items were retired NAEP items and some were written by members of the project staff. Although teachers' end of project mean scores on the assessment were slightly higher than their *before* scores, the increase was statistically significant only for Cohort 2 middle school teachers, a group of teachers who were far less experienced in history than the other participants. Examination of the technical aspects of the content assessments suggests that findings of low to moderate increases in historical content knowledge among teachers may have been in part due to the poor quality of the measure for this purpose. Multiple choice items (both NAEP and presenter-generated) tended to discriminate poorly and NAEP multiple choice items were too easy to provide good information about teacher content knowledge. Rubrics used for scoring short answer questions were not straightforward and did not provide useful distinctions between good, adequate, and poor responses.

Although every attempt was made to select items that tapped teachers' content knowledge in areas addressed by the GWHI professional development, since items were chosen before the professional development occurred, it was not possible to be sure that items would be sensitive to the instruction actually provided be project staff.

After each professional development opportunity, teachers provided ratings of their perceived increase in content knowledge and consistently reported increases in content knowledge. For high school teachers, these increases were particularly noticeable in their knowledge of *Segregation in the late 19th Century*, *Liberalism and Conservatism in the 20th Century* and *The United States and Developing Nations*, and for middle school teachers, increases were most dramatic for the topics of *Industry and Labor in the Early Republic*, *Fervor for Change in the mid-19th Century*, and *Reconstruction.*

Teachers' approaches to lesson planning evolved over the course of the project. Lesson planning surveys conducted before and after the project showed changes in the ways that teachers approached lesson planning. Especially among middle school teachers, teachers placed more emphasis at project's end on selecting and making effective use of relevant primary sources and on providing appropriate scaffolding for students' use of the historical skills. Middle school teachers also revealed greater understanding of teaching history as an interpretive, evidence-based process, rather than rote-learning of established truths. Changes in high school teachers' approaches were similar in nature, but the magnitude of the change was not as great over the course of the project as it was for the middle school teachers.

The second emphasis of the GWHI project evaluation during the final year was to collect data to examine student outcomes. Two types of assessments were used to examine changes in 1) student content knowledge and 2) abilities to perform historical skills. Content knowledge assessments were constructed as subsets of the items on the teacher content knowledge tests, with a focus on items at the right difficulty levels for students. Each test included multiple choice and short answer items, some retired NAEP items and some written by project staff. During 2007-08 and 2008-09, middle school and high school students showed significant growth on the content knowledge assessments. In 2007-08, 76% of the students taking the test showed growth from pre- to post-test and in 2008-09, approximately 84% showed growth. While significant growth was seen on these instruments, in the absence of a comparison group, it is difficult to interpret these results.

Historical skills assessments were created by the evaluation team and the project director to provide a means for examining students' abilities to employ the five historical skills that are part of the GWHI paradigm. Each assessment required students to use their own background knowledge in combination with primary source photos or documents and/or tables of statistics to help them answer a short set of questions. Middle school students' performance on the skills assessments increased significantly from pre- to post test with 86% of the middle school students showing growth. However,

high school students' mean performance did not increase, with approximately half of the high school students demonstrating growth on the assessments. Instrument quality may also have played a role in the null findings in this area. In particular, the psychometric quality of the high school skills tests was markedly lower than all of the other instruments.

In summary, teachers who participated in the GWHI project demonstrated growth in their thoroughness of teaching of historical topics, in their strategies for teaching historical skills, and in their approach to planning lessons that incorporate historical thinking skills. They reported perceived growth in their own historical content knowledge, however, only slight gains were seen on content knowledge assessments. Lack of findings in this area may have been due to the lack of good quality assessments for examining teacher content knowledge at the middle and high school level. Teachers perceived that their students were more competent in both their content knowledge and their ability to perform historical skills. Middle school and high school students of participating teachers demonstrated significant growth on content knowledge assessments, and middle school students demonstrated significant growth on historical skills assessments. The difficulty of finding or developing content knowledge and skill assessments that are sensitive to interventions remains a problem for evaluating outcomes of projects like the GWHI project.

2. Project Description

The *Grant Wood History Institute (GWHI)* was a three-year teacher professional development project of the Teaching American History Program (TAH) funded by the U.S. Department of Education. The project in part evolved out of two previous projects, *Bringing History Home (BHH)* and *Bringing History Home II*, TAH-funded elementary history projects completed in 2004 and 2007. The elementary projects emphasized explicit teaching of historical thinking skills within historical curricula designed for the project. The *GWHI* project built upon the historical skills included in the BHH projects to provide middle school and high school teachers with training in using the historical skills as a part of their pedagogy and in providing students with the historical skills to approach new historical content. The project was conducted in the Grant Wood Education Agency (GWAEA) in east central Iowa. The GWAEA comprises 32 public school districts ranging from small rural districts to two of the eight largest urban districts in Iowa. The *GWHI* project provided participating teachers with 1) Professional development workshops for middle and high school teachers on using five historical skills as part of their history pedagogy, 2) Historical content explorations sessions in 16 different topics where historical pedagogy was modeled, 3) Resources to support the teaching, and 4) Ongoing professional support from the project director, project staff, and teacher mentors during the implementation phase.

The GWHI professional development took the form of twelve full days of grade level specific programming conducted by GWHI staff and guest presenters. During each year, there was one day for large group examination of the GWHI paradigm for teaching history and a four-day content and pedagogy institute with concurrent sessions for middle and high school teachers. The first year there was an additional workshop day on using biographies in teaching history and during the second year James Loewen conducted a workshop on his research.

The GWHI project was planned as a three-year project to be concluded during Summer 2009. Because of a variety of factors (including a natural disaster) the project received permission for a no-cost extension carrying remaining funding into a fourth year. Evaluation activities were concluded by the end of Year 3 and although data analysis continued into the no-cost extension year, this report will not include any ongoing project activity during Year 4. Annual evaluation reports filed during Years 1, 2, and 3 include detailed information about project activities and that information will not be repeated in this report. All reports continue to be available at the CEA website (http://www.education.uiowa.edu/cea/tah/index.html)

During all project years, all teachers were invited to receive on-site one-on-one support through visits from the project director. Email and telephone support from project staff and mentor teachers was also available. Additional voluntary large group workshops were conducted during the no-cost extension year of the grant period, however they were not evaluated.

3. Evaluation Methodology

3.1 Evaluation Sample

Data for the evaluation GWHI project were collected from two populations: teachers who participated in the GWHI professional development and the students of those teachers. All participating project teachers were required as part of their project responsibilities to take part in the project evaluation. In all instances, most of the teachers who were asked to complete evaluation instrumentation did so, with response rates on each instrument exceeding 90%. Some project participants were not asked to contribute to various aspects of the evaluation because they did not fulfill some characteristic necessary for the particular evaluation instrument. For example, a few of the participants were not currently classroom teachers – two because they were student teaching, two were graduate students, and one was on maternity leave. Therefore, it did not make sense to ask those teachers to complete implementation surveys concerning their teaching or collect student data. A few teachers joined the project too late to take the pretest content assessments, and a few were present for the pretest, but were unable to attend subsequent workshops and therefore were not eligible for further evaluation activities. Throughout the results section of this report, the response rate for each of the instruments is given as the percentage of those who

completed the instrument out of the number of *eligible participants.* Omission from the eligible sample does not appear to have been related to project participation or participant satisfaction and the response rates among the eligible participants were very high. Thus, the results should be considered to be representative of the entire population.

In addition to the exceptions described above, it is important to describe several other attributes of the GWHI teacher participants. First, in each cohort, in addition to the teachers who teach typical US history courses in middle school or high school, there were participants who teach other topics. These teachers expressed interest in participating in the GWHI professional development because they incorporate US history into their course, usually in English classes or in other social studies courses, such as government or civics. Second, although the project participant recruitment originally employed a voluntary participation model, some of the teachers in middle school Cohort 2 were required by their districts to participate in the GWHI project because they had not taught US history in the past, but were being assigned to teach history during the next school year. These participants were therefore different in two ways; they were novice US history teachers with little or no history background, and they were not volunteers.

The original evaluation design for the GWHI project included a comparison group consisting of teachers who were randomly selected for second-year only participation in Cohort 2 with data collection to take place during the year prior to their project participation. However, not enough teachers were recruited until the project was well under way, so this model could not be used. With the participation of non-US history teaching participants, we considered using that group as a pseudo-comparison group. However, for many reasons, this was not appropriate. For example, when considering using the student data of non-US history teachers, we realized that some of these teachers taught younger students and some taught older students who had already had one or more years of US history instruction. For this reason, data from all teachers are included when looking at teacher outcomes, however student data collected from non-US history teachers are not included in our analysis. GWHI project teachers' demographic characteristics were described in earlier reports.

Evaluation data was also collected from students of eligible teacher participants. All eligible teacher participants collected data from all of their students (or all students in at least two sections of their classes if they taught more than two sections of the same class). For the first year data analysis (2007-08), student data were included only if students had completed both pre- and post-tests, so that change scores at the student level could be explored. The most common reason that student test scores were excluded from analysis was because their school's system did not call for them to continue with the same teacher throughout the school year and so some students had a non-participating teacher at the time of the posttest. Furthermore, pretest scores of students who did not complete the posttest did not differ from the

pretest scores of those who completed both tests. During the second year (2008-09), a random sample was selected from students' matched pre- and posttests. Table 1 shows the demographic characteristics of the resulting 2008-09 sample.

Table 1. Student Demographic Characteristics (2008-09)

Variable	Middle School			High School		
	Level	Frequency	Percentage	Level	Frequency	Percentage
Grade Level	7	131	*58.5*	9	54	*34.4*
	8	93	*41.5*	10	42	*26.8*
				11	56	*35.7*
				12	5	*3.2*
Gender	Male	83	*36.9*	Male	68	*42.2*
	Female	142	*63.1*	Female	93	*57.8*
Age	11	1	*0.5*	14	15	*9.3*
	12	28	*12.6*	15	50	*31.1*
	13	115	*51.6*	16	43	*26.7*
	14	78	*35.0*	17	50	*31.1*
	15	1	*0.5*	18	3	*1.9*

3.2 Evaluation Instrumentation

For the evaluation of the *Grant Wood History Institute*, five data collection methods were used, three for teachers and two for students: 1) Teacher Implementation Surveys, 2) Teacher Content Knowledge Assessments, 3) Teacher Lesson Plan Surveys, 4) Student Content Knowledge Assessments, and 5) Student Historical Skills Assessments. Evaluation team members also communicated regularly with the project director concerning the curriculum implementations, workshop development, and communications with teachers. In this section, each of these will be described in detail including instrument development, administration, scoring and analysis procedures, and technical data concerning the instrument. [Data collection methods used in previous grant years have been described in detail in previous reports available at the CEA website given above.]

3.2.1 Teacher Implementation Surveys

During the fall of 2008 and spring of 2009, evaluators asked all GWHI participating teachers to complete surveys concerning their teaching of history. Only Cohort 1 teachers completed the first survey and both cohorts completed the second survey. Surveys were administered to the teachers using the online survey tool, Websurveyor. Teachers were sent a link that allowed them to access the surveys. The evaluation team designed two surveys, one for middle school teachers and one for high school teachers. Each survey was designed to ask teachers about their teaching of typical curriculum at each grade level

and about their teaching of content and use of pedagogy specifically related to the GWHI teacher professional development activities. Surveys consisted of a combination of scaled and open-ended items concerning 1) the teachers' self-described thoroughness of teaching different content elements, 2) teachers' perception of their students' competence at performing different skills and student knowledge of content related to the history curriculum, 3) teachers' opinions of the benefits for their students of teaching using different aspects of the GWHI approach, 4) teachers' use of different pedagogic techniques during their implementation of the curriculum, and 5) teachers' opinions on teaching history and the skills they believe students need to learn history.

Survey items concerning content taught were constructed after surveying chapter titles and units in several textbooks for teaching middle and high school history, and considering the eight content areas addressed at each level during the GWHI Professional Development Institutes. The textbooks surveyed are included in the References section of this report. Copies of implementation surveys are included in Appendix X of this report.

Responding to the surveys was one of the responsibilities of teachers as project participants and led to a high response rate at each survey administration. Survey response rates for all survey administrations were high, ranging from 90-98%.

3.2.2 Teacher Content Knowledge Assessments

During the year prior to the first teacher GWHI professional development institute, evaluation team members developed two historical content knowledge assessments (one at each grade level) to measure teacher and student content knowledge based on four guiding principles. In particular, items were designed to:

- Address content knowledge related to the eight topics (at each grade level) addressed by GWHI professional development
- Fulfill program requirements for use of nationally standardized test items for assessing content knowledge
- Include items written by project instructors to address specific content areas they planned to include in their instruction
- Employ multiple item formats, including both multiple choice and short answer formats.

In addition, standardized test items were chosen to best reflect content addressed during professional development and *when available* call upon the historical skills addressed by the project (e.g. primary source analysis). Table 2 lists the eight topics originally intended to be addressed by GWHI presenters and emphasized in the content knowledge assessments at each grade level. Because of staff changes, topics addressed during Year 2 did not follow this schedule exactly, so some of the questions chosen were

not as good a match for the topic actually addressed. For example, instead of addressing "Forty Years of Cold War" during the first session of the Year 2 high school institute, the presenters provided content concerning the Vietnam War.

Table 2. Topics to be addressed by GWHI presenters and content knowledge assessments

Year topic addressed		High School Topics	Middle School Topics
Year 1			
	1	Segregation and Discrimination in the late 19th Century	Making of the Constitution and Other Founding Documents
	2	The Gilded Age and the Progressive Response	African Americans and Slavery
	3	Economy and Society (1925-45)	Industry and Labor in the Early Republic
	4	Strategies of the Civil Rights Movements	Westward Expansion
Year 2			
	5	Forty Years of Cold War	A Fervor for Change in the mid-19th Century (Suffrage and Abolition)
	6	Liberalism and Conservatism in the 20th Century	Milestones on the Road to Civil War
	7	The United States and Developing Nations	The Civil War
	8	Post Cold War Society and Politics	Reconstruction

Table 3 describes the three dimensions of each of the two teacher content knowledge assessments.

Table 3. Teacher Content Knowledge Assessment Elements

Item Dimensions	**Item Types**	**Middle School Assessment (k)**	**High School Assessment (k)**
Source	NAEP	12	36
	Presenter-generated	25	13
Format	Multiple choice	16	24
	Short answer	21	25
Topic†	1	6	5
	2	4	8*
	3	3	5*
	4	6	5
	5	3	8
	6	5	4
	7	5	4
	8	5	5

*An additional five items tapped content knowledge from both high school topics two and three.
† Topics are different for each grade level and are listed in Table 2.

Project participants were asked to complete the content knowledge assessment at three times: on the first day of the Paradigm Workshop, at the end of their first summer institute, and at the end of the program. For the first cohort, this meant that they completed the assessment in May of 2007, June 2007, and May 2009. The second cohort completed the assessment in May 2008, November 2008 (because of postponement of summer institute), and May 2009. For both cohorts, the first two assessments were given as paper and pencil instruments in person at the workshop and the final assessment for all participants was completed remotely online via Websurveyor. Table 4 shows the number of participants who completed each assessment.

Table 4. Number of Participants completing Teacher Content Knowledge Assessments

Grade Level	Cohort	Pretest	Posttest 1	Posttest 2
Middle School	1	9	10	10
	2	19	14	10
TOTAL MS		28	24	20
High School	1	29	27	16
	2	6	2	7
TOTAL HS		35	29	23

Scoring of Teacher Content Knowledge Assessments

All data from teacher content knowledge assessments were entered into Excel spreadsheets. Responses to open-ended items were scored by a trained scorer using rubrics from the NAEP and from the presenters who generated the items. The scorer was trained to score the responses for open-ended items until they reached an acceptable rate of agreement (at least 80%) with the evaluator. After all pre-tests and the first posttest were scored, four items were eliminated from the middle school assessment; two were deleted because they were poorly written or too difficult, and two were eliminated because all respondents answered them correctly on the pretest. Eight items were eliminated from the high school content knowledge test because all or nearly all teachers answered them correctly on the pretest. Pretest scores were adjusted to eliminate those items from the analysis provided in this report. Tables 5 and 6 show the item discrimination and difficulty by item type and origin for the middle and high school teacher tests, respectively. At the middle-school teacher level, multiple-choice (MC) items tended to discriminate poorly. Whereas the NAEP MC items appeared to be quite easy for participants, presenter-generated MC items were more difficult. Short-answer (SA) items had higher mean discrimination levels than MC items. Whereas the 2-point NAEP SA items were relatively easy for participants, the 3-point NAEP SA items and presenter-generated SA items were more difficult.

Table 5. Middle School Teacher Content Knowledge Test Item Discrimination and Difficulty (Final posttest only)

Item types			Discrimination and difficulty of items	
Format	Origin	(k)	*Discrimination* Range Mean	*Difficulty*[a] Range Mean
MC	NAEP	(6)	-0.11 – 0.47 0.16	0.80 – 1.00 0.91
	Presenter	(17[b])	-0.44 – 0.56 0.26	0.00 – 0.95 0.51
SA	NAEP 2 pt items	(3)	0.12 – 0.46 0.33	0.75-0.90 0.83
	3 pt items	(2)	0.49 – 0.51 0.50	0.52-0.62 0.57
	Presenter 2 pt items	(4)	0.05 – 0.55 0.39	0.20-0.80 0.49

[a]Difficulty is the average item score; for short answer items, difficulty is computed as the mean score divided by the range of possible points
[b] Also includes one-point short answer items

At the high-school teacher level, the MC items tended to discriminate relatively poorly. Mean item difficulties for the MC items indicated that the NAEP items were easier than the presenter-generated items. For the SA items, NAEP items discriminated poorly, whereas the presenter-generated items performed much better. NAEP SA items, whether two or three points, tended to be easier for participants than presenter-generated SA items.

Table 6. High School Teacher Content Knowledge Test Item Discrimination and Difficulty (Final posttest only)

Item types			Discrimination and difficulty of items	
Format	Origin	(k)	*Discrimination* Range Mean	*Difficulty*[a] Range Mean
MC	NAEP	(19)	-0.13 – 0.72 0.33	0.69 - 1.00 0.92
	Presenter	(12[b])	-0.09 – 0.44 0.24	0.35 - 1.00 0.72
SA	NAEP 2 pt items	(3)	-0.09 – 0.38 0.09	0.75-0.94 0.86

3 pt items	(2)	-0.02 – 0.47 0.23	0.54-0.88 0.71
Presenter 2 pt items	(6)	0.28 – 0.56 0.36	0.33-0.83 0.61
3 pt items	(2)	0.32 – 0.48 0.40	0.41-0.65 0.53

[a]Difficulty is the average score; for short-answer items, difficulty is the average score divided by the number of points possible
[b] Also includes one-point short answer items

3.2.3 Teacher Lesson Plan Surveys

At the Paradigm Workshop (the participants' first contact with the project), participants were asked to complete a Lesson Planning Survey. They also completed the survey at the end of their project participation. The survey was designed to examine teachers' approaches to history lesson planning prior to and following involvement with the GWHI project. The survey was almost completely open-ended: survey participants were given large boxes in which to write "how you might plan a lesson about an important aspect of one of the broad U.S. history topics given below. Use the boxes to demonstrate the structure you will use for your lesson planning." The survey was intentionally open-ended to provide participants with freedom to describe their own process in lesson planning. Teachers were asked to choose between two broad history topics to use as the historical context in which to demonstrate their approach. Data were entered into an Excel file and randomized for scoring purposes.

The project director developed an analytic rubric using eight criteria for scoring the lesson planning surveys. The criteria were: Use of teacher content knowledge, Making Connections (both to today and to students' prior knowledge), Use of Essential Questions, Historical Thinking Skills, Lesson Centerpieces, Activities (with appropriate scaffolding), Assessments, and Constructivist Approach. On each of the eight criteria, scores of 0, 1 or 2 points were possible. The full rubric is attached in Appendix X.

3.2.4 Student Content Knowledge Assessments

During 2007-08 and 2008-09, participating GWHI teachers were asked to administer two assessments to all their students (in at least two sections if they taught more than one). The first assessment was a content knowledge assessment made up of a subset of items from the Teacher Content Knowledge Assessment. For each grade level, we created two parallel forms. This design served two purposes; to shorten the student assessment to facilitate administration in a classroom setting, and to create different forms for pre- and post-testing. In constructing the assessments, we attempted to balance

a number of elements on each version of the assessment: the number of items, the number of points possible on each test (some short answer items are worth two or three points), item format (short answer or multiple choice), item source (NAEP or presenter-generated), and item topic. It was not possible to balance all of these elements. Several items appear on both forms of each level assessment. A few items from the teacher content knowledge assessment were not used on the student content knowledge tests because they appeared to be too difficult for teachers. Table 7 reports the characteristics of the alternate forms of the student content knowledge assessments.

Table 7. Student Content Knowledge Assessment Elements

Item Dimensions	Item Type	Middle School Assessment Form A (k)	Middle School Assessment Form B (k)	High School Assessment Form A (k)	High School Assessment Form B (k)
Source	NAEP	6	7	16	12
	Presenter-generated	16	15	9	13
Format	Multiple choice	10	11	17	18
	Short answer	12	11	8	7
Topic†	1	4	5	2	2
	2	2	2	5	4
	3	2	2	2	3
	4	3	2	1	1
	5	2	1	6	5
	6	4	3	2	3
	7	1	3	2	2
	8	4	4	3	4
	2 and 3 (HS only)	n/a	n/a	4	4
Total Number of Items		22	22	25	25
Total Number of Possible Points		30	30	41	41

† Topics are different for each grade level and are listed in Table 2.

GWHI teachers were asked to administer the assessments during the first and last months of the school year. Teachers were randomly assigned to the form of the test that they were to administer at pretest and then given the other form for posttest. Several problems arose in data collection. During 2007-08, some teachers did not use the form they received, with some teachers using a form given to another teacher in their building, and some using the same form for both pre- and post-testing. Some teachers gave the tests to over 100 students while others had fewer than 20 students, creating an imbalance in the number of students who took Pretest A/Posttest B relative to those who took Pretest B/Posttest A. We also learned that in some schools, it was unlikely that students have the same teacher all year, meaning that they were not present for both pretest and posttest. Finally, as mentioned previously, some of the project participants were not primarily history teachers; for example, some were English teachers who incorporated historical topics and literature into their curriculum and some taught government or other social studies topics and participated in the project because they were interested in

using the GWHI approach for history topics they address in their classes. Although from a project perspective it made sense to include these teachers, it did not make sense to expect their students to gain the content knowledge expected during the course of a year learning US history. During the 2008-09 data collection cycle, we emphasized that teachers were to use only the test they had been given. After we received the data for analysis, we selected random samples of no more than 20 students per teacher, including only those students with both pre- and post tests. In addition, for the 2008-09 data collection cycle, we analyzed only the data collected from teachers who taught regular American History courses.

Teachers were asked to allow students plenty of time to complete the assessment and all teachers reported that all students were allowed to finish the assessment. Teachers used postage-paid mailing labels to return their student assessments to CEA and/or delivered the assessments to CEA personnel at GWHI sessions. Table 8 reports the number of assessments that were scored and analyzed during 2007-08 and 2008-09.[1]

Table 8. Number of Student Content Knowledge Assessments Used in Analysis, 2007-08 and 2008-09

Grade Level	2007-08 Data used for analysis	2008-09 Data used for analysis
High School	372	164
Middle School	453	229

Scoring of Student Content Knowledge Assessments

Multiple choice items were scored using SAS. For scoring purposes, short answer items on the content knowledge assessments took two forms; objective and extended response. Objective items were scored correct or incorrect by one evaluator and no partial credit was given. Extended response items were scored by a trained scorer. For each item, scorers used a rubric that either accompanied retired NAEP items or that had been written by one of the GWHI presenters. Rubrics and training materials included actual student responses marked with criterion scores. (Rubrics and training procedures are provided in Appendix X). The scorers completed a training protocol which involved scoring actual student responses and the percent exact agreement with the evaluator's criterion scores was calculated. A minimum of 80 percent exact agreement on the training sample was required before the scorer was allowed to score the remainder of the assessments. During training, the evaluator discussed discrepancies in scoring and answered questions. Once scoring was completed for a particular item, rater accuracy was checked by computing the percent exact agreement with criterion scores for a random sample of approximately 10% of all student responses (i.e., the "reference" sample). Table 9 reports the

[1] The total number of student content knowledge assessments that were collected during the two data collection years has been described in previous reports.

range of percent exact agreement for the middle school and high school tests during the 2007-08 and 2008-09 scoring cycles.

Table 9. Percent Exact Agreement in Scoring Middle and High School Short Answer Items

Grade Level	2007-08 Range of Training %	2007-08 Range of Reference %	2008-09 Range of Training %	2008-09 Range of Reference %
Middle School	81-100%	93-98%	87-98%	91-100%
High School	80-100%	88-99%	81-96%	87-100%

Data analysis was conducted using SAS. Means and ranges of item discriminations and difficulties are given in Tables 10 and11 for the middle and high school assessments. For middle school students, the NAEP MC items were the easiest and the NAEP three-point SA items were the most difficult. The range and mean of item difficulty across all item types suggest that almost all items were quite challenging for students. Discrimination for all MC items was quite low, with presenter-generated MC items in some cases having negative discriminations. Discrimination of the SA items tended to be much better.

Table 10. Middle School Student Content Knowledge Test Item Discrimination and Difficulty

Item types		Discrimination and difficulty of items					
		Form A			Form B		
Format	Origin	Form A (k)	*Discrimination*[a] Range Mean	*Difficulty*[b] Range Mean	Form B (k)	*Discrimination*[a] Range Mean	*Difficulty*[b] Range Mean
MC	NAEP	(5)	0.11-0.30 0.21	0.52-0.58 0.55	(5)	0.21-0.33 0.28	0.35-0.72 0.47
	Presenter[c]	(7)	-0.12-0.31 0.13	0.05-0.55 0.27	(7)	0.00-0.33 0.15	0.02-0.45 0.20
SA	NAEP 2 points	(1)	n/a 0.48	n/a 0.35	(2)	0.41-0.55 0.48	0.21-0.47 0.34
	3 points	(2)	0.26-0.43 0.35	0.06-0.16 0.11	(2)	0.37-0.56 0.47	0.03-0.13 0.08
	Presenter 2 points	(0)	n/a n/a	n/a n/a	(2)	0.25-0.43 0.34	0.09-0.57 0.33

[a] Discrimination is defined as the correlation of the score on this item with the total score
[b] Difficulty is defined as the mean score across all students
[c] Includes one-point short answer items

For the high school student content knowledge assessment, the NAEP MC items were easier than the presenter-generated MC items. Both types of MC items (NAEP and presenter-generated) tended to

have poor discrimination, with a few of the presenter-generated MC items displaying negative item discrimination. Among the SA items, NAEP two-point items were by far the easiest, with most of the other item types being much more challenging for students. Most of the SA items tended to discriminate reasonably well, with the exception of the NAEP SA items on Form A, which tended to have lower discrimination levels.

Table 11. High School Student Content Knowledge Test Item Discrimination and Difficulty

Item types		Discrimination and difficulty of items					
		Form A			Form B		
Format	Origin	Form A (k)	*Discrimination*[a] Range Mean	*Difficulty*[b] Range Mean	Form B (k)	*Discrimination*[a] Range Mean	*Difficulty*[b] Range Mean
MC	NAEP	(14)	0.07 – 0.42 0.25	0.17-0.85 0.47	(12)	0.09-0.48 0.28	0.27-0.80 0.46
	Presenter [c]	(4)	0.12-0.88 0.38	0.06-0.33 0.23	(5)	-0.05-0.24 0.15	0.16-0.37 0.26
SA	NAEP						
	2 points	(3)	0.17-0.31 0.23	0.53-0.70 0.60	(1)	n/a 0.37	n/a 0.61
	3 points	(1)	n/a 0.29	n/a 0.09	(1)	n/a 0.48	n/a 0.40
	Presenter						
	2 points	(2)	0.25-0.40 0.33	0.10-0.21 0.15	(5)	0.20-0.50 0.31	0.04-0.30 0.19
	3 points	(1)	n/a 0.36	n/a 0.07	(2)	0.52-0.58 0.55	0.18-0.44 0.31

[a] Discrimination is defined as the correlation of the score on this item with the total score
[b] Difficulty is defined as the mean score across all students
[c] Includes one-point short answer items

Student Content Assessment Reliability

In order to summarize the internal consistency of the student content knowledge assessments, we computed coefficient alphas for all forms of the content knowledge assessments. Table 12 presents the internal consistency reliability estimates (coefficient alphas) for the assessments. Overall alpha values for the middle school assessments range from 0.36-0.68, and for the high school they range from 0.72-0.76. The most likely explanation for the higher internal consistency estimates at the high-school level is the additional items included on the high school assessments. In particular, whereas each form of the high school assessment included 25 or 26 separately scoreable items for a maximum possible score of between 34 and 38 points , the middle school assessments only included between 15 and 18 separately scoreable items for a maximum possible score of between 20-26 points. In addition, because tests at both grade levels were constructed to sample content knowledge within eight broad topics of US History, items were

quite heterogeneous, which may depress alpha estimates. Further, open-ended items involving trained raters and scoring rubrics introduced subjectivity (and some amount of unreliability) into the scores. Table 12 also presents the test-retest reliability estimate for parallel forms, in the form of the correlation between composite scores on the pre- and post-test. Because learning between tests was expected to occur (and did occur, particularly among middle school students), lower correlations between parallel forms taken at pre- and posttest at the middle school level are not surprising.

Table 12. Reliability Estimates for Student Content Knowledge Assessments

Test		N	Reliability Index: Cronbach's alpha (standardized)	Reliability Index: Test-retest parallel forms
Middle School				
2007-08		453		0.27*
	Form A	674	0.36	
	Form B	640	0.59	
2008-09		229		0.48*
	Form A	229	0.53	
	Form B	229	0.68	
High School				
2007-08		372		0.52*
	Form A	931	0.76	
	Form B	1203	0.76	
2008-09		164		0.55*
	Form A	164	0.73	
	Form B	164	0.76	

*Indicates correlation coefficient significant at p<.0001

3.2.5 Student Historical Skills Assessments

The second student assessment used was the Student Historical Skills Assessment. Four versions of the skills assessment (two at each level) were created by the GWHI Project Director with the assistance of CEA evaluation staff. Skills assessments were designed to be sensitive to both the historical content taught at each level and the historical skills promoted by the GWHI methods. Each of the four skills assessments included two items, each embedded within a historical context, which required students to examine a photograph, read a primary source passage, and/or examine a table of statistics and then answer several open-ended questions about the content, using their own historical background knowledge to help them answer the questions. Since background knowledge was to be used, all items required students to synthesize what they could glean from the sources with which they were presented, with their own historical background knowledge. Table 13 reports the historical content and skills addressed by each item.

Table 13. Skills Assessment Items Content and Skills

Item	Content	Skill
Middle School		
A1	Reconstruction	Document Analysis

A2	Civil War causes	Use of Statistics
B1	Native American removal	Document Analysis
B2	Civil War	Photo Analysis
High School		
A1	15th Amendment/Voting Rights Act	Document Analysis
A2	Dust bowl	Photo Analysis
B1	Vietnam/My Lai Massacre	Use of Statistics/Document analysis
B2	Sharecropping	Photo Analysis

Skills tests were administered by all teachers in 2007-08 and 2008-09 as both-pre and posttests. As with the content knowledge tests, teachers were randomly assigned to administer either version A or B at pre-test and the other version at post-test. Because of the complexity of scoring the skills tests, we selected a sample of 10 teachers (six middle school and four high school) whose students' tests would be scored. We sampled 30 students (or fewer if there were fewer than 30 students in class) from only the teachers who taught US history and who had at least 20 students taking both pre- and posttests. The final sample included 107 high school and 170 middle school students. Responses to all items were entered into Excel files and then all pre- and post responses were randomized at the item level for scoring.

All items were scored by a process of assigning each subtest item a score of 1-4, corresponding to the relative strength of the response, with "4" meaning the response was among the "strongest" in the set of responses, and "1" meaning the response was among the "weakest" in the set. As part of the process of assigning scores to the responses, the project director derived analytic rubrics for each of the eight items to clarify the scoring process for the second scorer. The second scorer was trained in using the rubrics; a minimum of 75 percent exact agreement on the training sample was required before the scorer was allowed to score the remainder of the assessments. On half of the items, this standard was met on first training. For those items where 75 percent exact agreement was not reached during training, rubrics were revised and items were rescored by the trained scorer and the evaluator, establishing again a minimum of 75% exact agreement. Table 14 reports the percent exact agreement at initial training and at retraining on the revised rubric (as necessary) for each of the eight items.

Table 14. Percent Exact Agreement in Scoring Middle and High School Skills Assessment Items

Item	Initial Percent Exact Agreement % (k)	Retraining Exact Agreement % (k)
Middle School		
A1	77% (60)	n/a
A2	55% (31)	86% (36)
B1	65% (31)	88% (36)
B2	72% (40)	95% (40)
High School		
A1	87% (30)	n/a

A2	87% (30)	n/a
B1	67% (30)	83% (30)
B2	90% (30)	n/a

Since the data set included an equal proportion of pre- and post-tests, and the scores were meant to signify the range of responses from weakest to strongest, another requirement of scoring was that the full distribution of possible scores be used. Therefore an additional requirement was that at least 5% of the responses must fall into each category (score). This distribution occurred naturally in all cases except on high school item B2, where many student responses were missing essential elements and were weak enough that the highest category was used for fewer than 5% or the responses. Table 15 reports the frequency distribution for each score on each of the skills test items.

Table 15. Frequencies on Student Skills Assessments

Item	Frequency of Score n (%)			
	1	2	3	4
Middle School				
A1	36 (21)	40 (24)	56 (33)	38 (22)
A2	33 (19)	40 (24)	37 (22)	60 (35)
B1	26 (15)	59 (35)	57 (34)	28 (16)
B2	67 (39)	25 (15)	30 (18)	48 (28)
High School				
A1	37 (35)	21 (20)	19 (18)	30 (28)
A2	29 (27)	33 (31)	25 (23)	20 (19)
B1	23 (22)	27 (25)	43 (40)	14 (13)
B2	7 (7)	91(85)	7(7)	2 (2)

Reliability estimates in the form of coefficient alpha and correlations between skills test scores on the pre- and post-test are reported in Table 16. Internal consistency was examined by calculating coefficient alpha for all versions of the skills assessment. Internal consistency was low to moderate at the middle school level and was much lower at the high school level. Lower alpha estimates for the high school students are very likely due to the lower variability of high school composite scores, which will be explored more fully in the Results section. Similar to the Student Content Knowledge Assessments, correlations between composite scores on the pre- and post-tests were generally small, which is to be expected as learning occurred during the interval.

Table 16. Reliability Estimates for Student Skills Assessments

Test		Reliability Index		
		N	Cronbach's alpha (standardized)	Test-Retest Parallel Forms
Middle School 2008-09				0.34
	Form A	170	0.58	
	Form B	170	0.44	

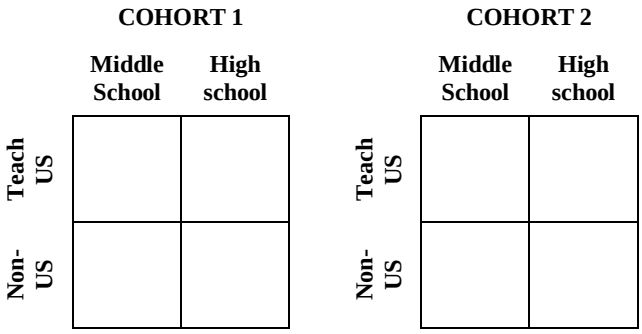
High School 2008-09				0.36
	Form A	107	0.08	
	Form B	107	0.27	

4.0 Evaluation Results

4.1 Teacher Outcomes

In examining teacher outcomes, as mentioned in the Methods section of this report, there are several project participant characteristics to keep in mind. The GWHI project consisted of eight different participant groups: two cohorts, each of which included middle and high school teachers, and each composed of US history and non-US history teachers (See Figure 1). Throughout this report, where relevant, results are provided by cohort, by grade level taught, and/or by whether or not the teachers teach US history. Results are always reported separately for middle and high school teachers and students, but are sometimes pooled by cohort and/or subject taught. For example, the first section of teacher results starts with a report of grand means across particular sets of survey items for middle school and high school teachers, by whether or not they teach US history.

Figure 1. GWHI teacher groups for analysis



4.1.1 Implementation of History Content and Pedagogy

Participating treatment teachers responded to one or two surveys (depending on their cohort) concerning their implementations of history curriculum during the final year of the *GWHI* project: Cohort 1 completed a survey during the fall of 2008 concerning their first year of participation and both cohorts completed implementation surveys during the summer of 2009 concerning their participation during the final year of the GWHI project. Online surveys included scaled and open-ended items. The surveys' scaled items served six main purposes; 1) to determine the teachers' self-described thoroughness of teaching different historical topics (some of which were emphasized during the GWHI workshops, some of which are typically addressed during high school and middle school), 2) to ascertain teachers'

perceptions of their students' content knowledge on specific historical topics, 3) to ascertain teachers' perceptions of their students' competence at performing skills addressed as part of the GWHI paradigm, 4) to gather teachers' opinions of the benefits and drawbacks of the GWHI historical skills for their students, 5) to determine teachers' perceptions of the utility of different pedagogical techniques and instructional practices for teaching history, and 6) to determine the likelihood of participants continuing to use the GWHI paradigm for historical pedagogy. In addition to the scaled items, a number of open-ended questions solicited participant feedback on their project participation including: other history topics and skills taught, their perception of students' additional competencies in historical content and skills, potential barriers to teaching history using the GWHI paradigm, changes in collaboration habits among participants, and other effects of participating in the project.

Although the same categories for items appeared on both surveys, two different versions of the survey reflected the different historical content of middle school and high school history classes. The full surveys and results are included in the appendices to this report. Highlights of these results organized by grade level will be discussed in this section.

Thoroughness of Implementation, Perception of Student Competence in Skills and Content

On each of the grade level surveys, there were three areas in which teachers were asked to think about the ways their instruction and their students had changed over the course of their project participation by providing retrospective pre-post ratings: 1) the thoroughness of their teaching of particular history topics, 2) their perception of their students' competencies at using historical skills, and 3) their perception of their students' content knowledge on specific history topics. In the first section, teachers rated the thoroughness of their own teaching of specific history content areas typical for their respective grade levels through scaled items asking them how thoroughly they taught each topic before and after (written as *before* and *now* on the surveys*)* they participated in the project. The scale for these items ranged from "1" for "Not at all", to "5" for "Very thoroughly." Item stems were phrased to convey the historical topic in a style typical of textbook chapter names.

In the second section, participants were asked to rate their perception of their students' competency at using five historical skills that are central to the GWHI paradigm. These skills included: constructing timelines to show important events and how they relate to each other, using maps to illustrate important concepts, interpreting primary sources to add to their understanding of history, reading for background knowledge to provide a context for new learning, and synthesizing information from various sources to create a narrative. Teachers used a five point scale to rate these skills ranging from "1" for "None are able to do this competently" to "5" for "All or nearly all are able to do this competently."

The third section asked the teachers to use the same competency scale to rate their perception of students' ability to answer questions concerning specific historical topics. The topics on this list were usually more narrowly-defined than the topics in the first section. For example, instead of asking about "The Colonies," this section asked about "Spanish Colonies in the New World." The next section of this report provides selected results from these sections of the teacher implementation surveys. Full results for scaled items and summaries of open-ended item responses are included in the appendices to this report.

Table 17 provides an overall summary of mean results for all scaled items on Thoroughness of Implementation, Teacher Perception of Student Competencies in Skills, and Teacher Perception of Student Competencies in Content from the middle and high school teacher surveys administered at the end of the 2008-09 school year, the final year of the project. (Results for Cohort 1 collected after the conclusion of their first year are not reported here because they were similar to the second year.) Middle and high school teacher mean ratings were averaged over all items within respective item clusters. Table 18 also disaggregates findings with respect to whether or not the responding middle school teachers were US history teachers. The GWHI project included teachers (at both the middle and high school levels) who were not currently US history teachers (typically teachers of other social studies topics or English teachers), but who expressed interest in incorporating history topics into their own teaching and in learning new pedagogical strategies for teaching history.

Table 17. Grand Means for GWHI participants

Grade Level	**Grand means for each item type cluster**					
	Thoroughness of Implementation (k)		Perception of Student Competencies (k)			
			Skills		Content	
(n*)	Before	Now	Before	Now	Before	Now
Middle School	(16)		(5)		(32)	
TOTAL (18)	2.37	3.08	2.39	3.94	2.25	3.49
Teach US (14)	2.53	3.25	2.45	3.98	2.38	3.71
Do not teach US (4)	1.80	2.49	2.20	3.80	1.84	2.78
High School	(21)		(5)		(37)	
TOTAL (21)	2.65	3.32	2.92	3.94	2.66	3.61
Teach US (13)	2.79	3.40	2.91	3.88	2.69	3.57
Do not teach US (8)	2.40	3.21	2.94	4.04	2.60	3.69

* The number of respondents varies slightly for each question and situation throughout the survey results because of non-answered items; the *n* provided here represents the maximum *n* who responded in each cell.

Among both middle and high school teachers, mean retrospective self-ratings of thoroughness of instruction and their perceptions of student competencies on historical skills and content knowledge

increased from *before* to *now* across all three categories. The largest *now* means (and mean differences *before* to *now*, see Tables 20 and 24) were in the middle school teachers' perceptions of their students' ability to use the five historical skills that are part of the GWHI paradigm. Although mean gains from *before* to *now* occurred in all situations, in nearly all cases, the means in the historical topic-related areas (both before and now) for teachers who do not teach formal US history courses were lower than for US history teachers. The two exceptions to this were high school non-US history teachers who rated their students' content knowledge competency slightly higher than did their US history teaching peers, and high school non-US history teachers who also rated their students' historical skills competency higher than their non-history teaching peers.

MIDDLE SCHOOL IMPLEMENTATION FINDINGS

Because each of the Summer Institutes included examination of four specific historical topics, Table 18 shows the mean difference scores for teachers' retrospective ratings of their thoroughness of teaching historical topics, which are grouped by whether the topics were addressed by the GWHI Summer Institutes. The mean difference from *before* to *now* for Cohort 1 middle school history teachers was greater for topics addressed by GWHI than for topics not addressed in GWHI instruction. For Cohort 1 middle school non-history teachers, the mean differences from *before* to *now* were less noticeable than for history teachers in general, and were roughly the same across topic categories. As reported earlier, the Cohort 2 middle school history teachers were different from Cohort 1 middle school history teachers in their experience in teaching history. While most of the Cohort 1 middle school teachers were mid-career history/social studies teachers, most of the middle school history teachers in Cohort 2 had no previous experience in teaching history. They were not new teachers, but had been assigned to teach one or more sections of US history even though they were certified to teach other areas, including English, math, and science. During the workshops, their levels of interest and enthusiasm for teaching history varied, with many teachers expressing resentment for their new assignment. The mean differences *before-now* for Cohort 2 middle school history teachers were smaller than for Cohort 1 and the mean differences between GWHI topics and other history topics were also smaller. Differences between Cohort 2 history and non-history teachers are difficult to interpret because many non-US history respondents left a number of items blank rather than choosing a *Before* option and were therefore left out of the analysis.

Table 18. Mean Difference Before/Now on Retrospective Pre-Post thoroughness of teaching items on middle school GWHI topics and non-GWHI topics (5 point scale)

Group	Mean Difference Score	
	Non-GWHI topics *(k)*	GWHI topics *(k)*

Cohort 1		
Teach history (n=5)	0.63 *(8)*	1.01 *(8)*
Non-US history (n=3)	0.44 *(8)*	0.40 (8)
Cohort 2		
Teach history (n=9)	0.56 *(12)*	0.73 *(4)*
Non-US history (n=1)	1.08 *(12)*	1.50 *(4)*

Table 19 reports the breakdown of the difference scores by specific topics, grouped by whether or not the topics were addressed by GWHI instruction. In Cohort 1, US history teachers indicated covering almost all topics more thoroughly *after* participating in GWHI than before, even the topics that were not addressed as part of their professional development. There were a few exceptions to this rule, including the topic of European Exploration, for which teachers generally suggested their thoroughness of instruction had remained the same, and the topic of A Reforming Age, for which teachers tended to maintain or decrease the thoroughness of their instruction. Non-US history teachers in Cohort 1 also increased the thoroughness of their instruction for most topics, with the exception of European Exploration and African-Americans and Slavery in the US, both of which demonstrated negative mean differences from *before* to *now*. Cohort 2 teachers demonstrated a similar pattern as their Cohort 1 counterparts, with uniformly positive mean scores indicating that all topics were covered more thoroughly *after* participation in GWHI than *before*.

Table 19. Difference scores for middle school teachers' thoroughness of instruction on GWHI topics and other history topics by topic

Items	Difference (Before/Now)		
	Cohort 1		Cohort 2[a]
	Teach US n=5	Non-US n=3	Teach US n=9
Topics *not* addressed by GWHI			
The First Americans	0.80	1.00	0.25
European Exploration	0.00	-0.67	0.14
The Colonies	0.60	0.17	0.50
Events Leading to the Revolution	1.00	0.67	0.50
The American Revolution	0.34	0.67	0.62
The New Government	0.75	0.66	0.56
The Jefferson Era	0.80	0.67	0.50
The Jackson Era	0.80	0.33	0.11
Topics addressed during GWHI Year One			
Making the Constitution and other Founding Documents	1.25	0.17	0.62
Industry and Labor in the Early Republic	1.40	0.33	0.69
Westward Expansion	0.65	0.67	1.11
African Americans and Slavery in the US	1.25	-0.67	1.11
Topics addressed during GWHI Year Two			
A Reforming Age	-0.05	0.34	0.38
Events Leading to the Civil War	1.40	1.00	1.41
The Civil War	1.00	1.67	0.66
Reconstruction	1.20	0.67	0.78

[a] Only one teacher in Cohort 2 was not a history teacher and is not included in this analysis.

In addition to the scaled items, teachers were asked what other historical topics they taught during the previous school year. Thirteen of the 18 middle school respondents answered this item for a response rate of 72%. Four middle school teachers said that they do not teach US history at this time, with two teaching world history and two teaching Civics or Geography. Some of the topics respondents provided were actually part of the list given in the scaled items, including some GWHI topics. The following is a list of the topics given by the other nine respondents (with numbers in parentheses if more than one person listed that topic):

- Immigration (3)
- Election 2008 (2)
- Thomas Jefferson and Sally Hemings relationship
- Big business in the late 1800s
- Branches of the government
- Bill of Rights
- Political cartoons
- Industrial Revolution
- Monroe Doctrine
- Andrew Jackson and the Trail of Tears
- Pioneers
- John Brown
- US Constitution
- Pearl Harbor Day
- Mid 19th century baseball as culture study
- Time period from 1800-1900
- World history topics

Middle school teachers were also asked which historical topics or content addressed by the GWHI project they would continue to teach after the GWHI project concluded. Nine of the 18 respondents replied to this item for a response rate of 50%. Three middle school teachers said that they would use the GWHI methods and teach the historical content more in-depth, but did not mention particular content. Two people said that much of the history they learned was new to them and they will teach "pretty much all of them". Topics mentioned specifically included: pre-Civil War, women's suffrage, Industry and Labor, Reconstruction, Eli Whitney, Civil War, and Westward Expansion. One person said that all of the material covered at the workshops is outside of their content area.

Table 20 reports the mean differences by cohort between middle school teachers' retrospective pre-post self-ratings of perceptions of their students' competency at using the five historical thinking skills that are part of the GWHI paradigm. There does not appear to be a cohort effect in middle school teachers' perceptions of their students' competence in using historical skills, with each group showing

large differences from pre-post, suggesting student competencies on all five skills improved from *before* to *now.*

Table 20. Mean Difference Before/Now on Retrospective Pre-Post of middle school teacher perception of student competence on skills items (5 point scale)

Group	Mean Difference Score *(k=5)*
Cohort 1	
Teach history (n=5)	1.56
Non-US history (n=3)	1.53
Cohort 2	
Teach history (n=9)	1.53
Non-US history (n=1)	1.60

Table 21 reports the mean differences between middle school teachers' retrospective pre-post ratings of their students' competence at answering questions concerning particular historical topics. As with the thoroughness of teaching items, some of these topics were addressed by the GWHI Summer Institutes and some were not, so results are shown for topics addressed by GWHI as opposed to topics not addressed during training, and are disaggregated by cohort. Tables with the complete list of these items and difference scores at the item level are included in the appendices to this report.

Table 21. Mean Difference Before/Now on Retrospective Pre-Post of middle school teacher perception of student competence at answering questions on GWHI and non-GWHI history topics (5 point scale)

Group	Mean Difference Score	
	Non-GWHI topic items *(k)*	GWHI topic items *(k)*
Cohort 1		
Teach history (n=5)	1.09 *(15)*	1.91 *(17)*
Non-US history (n=3)	0.30 *(15)*	0.21 *(17)*
Cohort 2		
Teach history (n=9)	1.07 *(24)*	1.24 *(8)*
Non-US history (n=1)	1.00 *(24)*	1.50 *(8)*

Mean differences were uniformly positive for Cohort 1 teachers, suggesting students' competence in answering questions on a variety of history topics increased from *before* to *now.* As with the thoroughness ratings, for Cohort 1 middle school history teachers, the mean difference from *before* to *now* was greater for topics addressed by GWHI than for topics not addressed by GWHI instruction. For the Cohort 1 non-history teachers, mean differences were much smaller than for history teachers, and about the same across topics. Similar to Cohort 1, teachers in Cohort 2 indicated their students were more competent at answering questions about a variety of history topics *after* participation in GWHI. As with Cohort 1 teachers, mean difference scores for Cohort 2 were higher for GWHI topics than non-GWHI

topics, although the magnitude of this difference for US history teachers was smaller than it was for Cohort 1 US history teachers.

Middle school teachers were also asked in an open-ended item to indicate any other historical topics on which they thought their students would be at least "somewhat competent" at answering questions. Six teachers responded to this item for a response rate of 33%. Other topics listed by single teachers were: Civics/Economics, Lewis and Clark exploration, Manifest Destiny, immigration/Ellis Island, Big business/robber barons, WWI, WWII, War on Terror, War in Iraq, and other world history topics. As with the teachers' responses to the thoroughness items, some of these responses are actually included in the options they were given, not additional topics. One teacher said, "I honestly have no idea. When I reviewed the course this past week, it was very disheartening to see what they didn't remember or what they mixed up."

HIGH SCHOOL IMPLEMENTATION FINDINGS

Table 22 reports the difference scores between *Before* and *Now* for self-reported thoroughness of instruction on GWHI and non-GWHI topics among high school US history and non-US history teachers.

Table 22. Mean Difference Before/Now on Retrospective Pre-Post thoroughness of teaching items on high school GWHI topics and non-GWHI topics (5 point scale)

Group	Mean Difference Score	
	Non-GWHI topics *(k)*	GWHI topics *(k)*
Cohort 1		
Teach history (n=11)	0.54 *(7)*	0.58 *(14)*
Non-US history (n=4)	0.64 *(7)*	0.56 *(14)*
Cohort 2		
Teach history (n=2)	0.73 *(16)*	1.10 *(5)*
Non-US history (n=3)	0.35 *(16)*	0.10 *(5)*

Mean self-ratings indicate that all high school teachers, both history and non-history, increased the thoroughness of their instruction on both GWHI and non-GWHI topics. For Cohort 1 teachers, reported increases in thoroughness were roughly similar across the two types of topics (GWHI and non-GWHI). The largest reported increase in the thoroughness of instruction was for Cohort 2 US history teachers, particularly for the GWHI topics, however this was a very small group, so differences should not be over-interpreted.

Table 23 reports the mean difference scores for high school teachers' retrospective ratings of their thoroughness of teaching historical topics for each topic and grouped by whether the topics were addressed by the GWHI Summer Institutes. Mean difference scores for Cohort 1 teachers suggest that

both history and non-history teachers increased the thoroughness of their instruction on all topics, even those not addressed by their GWHI training. Mean differences ranged in magnitude from modest increases (e.g., Post-Civil War Reconstruction) to more dramatic increases (e.g., Politics, Reform and Immigration). Difference scores for Cohort 2 are more difficult to interpret because there were so few respondents and several left a number of items blank rather than choosing a *Before* option and were therefore left out of the analysis. However, mean difference scores for Cohort 2 non-history teachers suggest their thoroughness of instruction did not change for many topics, and did not appear to increase on any of the topics addressed during the second year of GWHI.

Table 23. Difference scores for high school teachers' thoroughness of instruction on GWHI topics and other history topics

Items	Difference(Before/Now)			
	Cohort 1		Cohort 2	
	Teach US n=11	Non-US n=4	Teach US n=2	Non-US n=3
Topics *not* addressed by GWHI				
The Western Frontier	0.37	0.25	0	0
Toward an Urban Age	0.80	1.75	1.00	1.00
Becoming a World Power	1.00	0.50	1.50	0.50
The United States in World Affairs	0.53	1.00	1.50	0
World War I	0.24	0.25	1.50	0.33
World War II	0.45	0.25	1.00	0.17
The Nixon and Watergate Years	0.40	0.50	1.00	0
Topics addressed during GWHI Year One				
Post Civil War Reconstruction	0.10	0.10	0	0.50
An Industrial Era	0.91	0.50	1.00	0
Politics, Reform, and Immigration	1.90	1.25	1.00	1.67
The Progressive Movement	0.36	1.00	1.50	0.50
The Roaring Twenties	0.49	0.50	1.00	0.50
The Great Depression	0.82	0.50	1.00	0
The New Deal	0.49	0.25	1.50	0.50
The Cold War	0.70	0.50	1.00	0.50
The Post War Years at Home	0.10	0.50	1.00	0
Topics addressed during GWHI Year Two				
The Civil Rights Movement	0.64	0.75	1.00	0
The Turbulent 1960s	0.40	0.50	1.00	0
The Vietnam War Era	0.30	0.50	1.00	0
The Conservative Era	0.22	0.75	1.00	0
Entering a New Century	0.20	0.25	1.50	0

Teachers were also asked to list any additional historical topics they had taught during the past year. Ten of the 21 respondents provided additional US history topics. These included: Protest unit ('60s and '70s), beginnings of women's movement, labor and unions, "fabulous '50s", conservatism, genocide, Korean War, Gulf War, Progressive Era, and "lots of Civil War." As with the middle school

findings, some of these topics were mentioned in the list of topics queried (e.g. The Progressive Movement, The Turbulent 60s), so interpretation is not straightforward.

High school teachers were asked what topics from the GWHI institutes they would continue to address in their instruction after the end of the project. Ten of the 21 respondents provided a response to this item, for a response rate of 48%. The topics mentioned by individual teachers were: Gilded Age; Cold War politics; US foreign policy; eugenics; "new Urban Age"; increasing impact on US of immigration during last century; teaching Civil Rights by categorizing events in terms of their social, economic, political or legal basis and impact; and four teachers did not mention topics, but mentioned using the pedagogical methods used in GWHI.

Table 24 reports the mean differences by cohort in high school teachers' perceptions of their students' competency at using the historical thinking skills. Although all groups reported increased student competency at using the five historical skills, the mean increase for Cohort 1 US history teachers was slightly lower than it was for the other high-school teachers.

Table 24. Mean Difference Before/Now on Retrospective Pre-Post of high school teacher perception of student competence on skills items (5 point scale)

Group	Mean Difference Score *(k=5)*
Cohort 1	
Teach history (n=11)	0.87
Non-US history (n=6)	1.52
Cohort 2	
Teach history (n=2)	1.60
Non-US history (n=3)	1.13

Table 25 reports the mean differences in teacher ratings of student competency at answering questions on a variety of history topics, both GWHI and non-GWHI. Consistent with the teachers' thoroughness ratings, the teachers' retrospective pre-post ratings of students' competency at answering content knowledge questions increased from *before* to *now* in all conditions, although there was no difference in their ratings between the GWHI topics and other historical topics.

Table 25. Mean Difference Before/Now on Retrospective Pre-Post of high school teacher perception of student competence at answering questions on GWHI and non-GWHI history topics (5 point scale)

Group	Mean Difference Score	
	Non-GWHI topic items *(k)*	GWHI topic items *(k)*
Cohort 1		
Teach history (n=11)	0.75 *(18)*	0.81*(19)*
Non-US history (n=4)	1.09 *(18)*	1.00 *(19)*

Cohort 2		
Teach history (n=2)	1.66 *(25)*	1.75 *(6)*
Non-US history (n=2)	n/a*	n/a*

*Nearly all of these items were left blank by respondents so a cluster difference score would not be meaningful.

In an open-ended item, high school teachers were also asked what other skills for learning history they thought their students had acquired as a result of their use of the GWHI paradigm for historical pedagogy. Seven of the 21 respondents replied to this item for a response rate of 33%. Three high school teachers said their students acquired the ability to analyze political cartoons, two teachers said their students better understood cause and effect, and individual teachers said students could analyze primary sources, do some statistical analysis, understand chronological order, evaluate information sources, and discuss the unfairness of US treatment of African Americans, Native Americans, and immigrants.

Benefits for students, utility for teachers, and likelihood of continued use of skills by teachers

Teachers were asked a series of questions concerning the extent to which they thought that the GWHI historical skills were beneficial for their students, were useful in their instruction, and the likelihood that they would continue to use the skills as part of their instruction after the end of the GWHI project.

Teachers were asked to rate the benefit for their students of using the five GWHI historical skills on a five-point scale with "1" meaning "Not at all beneficial" and "5" meaning "Very beneficial". Table 26 reports the mean ratings for both middle and high school teachers, disaggregated by whether or not the teacher was responsible for US history and also by cohort for the middle school teachers. High school results were not reported by cohort because they were uniform across cohorts. All groups rated all skills as being somewhat beneficial. In particular, all skills received mean ratings of at least a 3. The highest ratings across all skill areas were awarded by Cohort 1 middle school US history teachers (mean=4.8) and the lowest ratings across all skill areas were given by Cohort 2 middle school non-US history teachers (mean=4.0). Across all groups of teachers, the skill receiving the highest mean ratings was "Reading for background knowledge" (mean=4.7) and the lowest ratings were assigned to "Mapping for understanding," although the Cohort 1 middle school US history teachers found this skill to be "very beneficial." Interestingly, among high school teachers, the non-US history teachers' mean ratings were higher than the ratings of their US history counterparts.

Table 26. Mean ratings of perception of benefit for students of using GWHI historical skills (5 point scale)

Skills	Middle School		High School	**Grand Mean[a] Across**
	Cohort 1	Cohort 2	Both cohorts	

	Teach US	Non-US	Teach US	Non-US	Teach US	Non-US	**Teacher Groups**
	n=5	n=3	n=9	n=1	n=13	n=8	
Constructing timelines	5.00	4.00	4.56	5.00	4.23	4.63	**4.49**
Mapping for understanding	5.00	3.67	4.33	3.00	4.15	4.25	**4.25**
Primary source analysis	5.00	5.00	4.78	3.00	4.69	4.88	**4.77**
Reading for background knowledge	4.80	4.33	4.67	5.00	4.54	4.88	**4.67**
Synthesizing multiple sources	4.20	4.67	4.44	4.00	4.08	4.88	**4.39**
Grand Mean Across Skills	**4.80**	**4.33**	**4.56**	**4.00**	**4.34**	**4.70**	

[a] Grand means are weighted by sample size.

Teachers were also asked what the most important student skill outcomes from the GWHI project paradigm were. Sixteen of the 18 middle school respondents answered this item for a response rate of 89%. Nearly all of those who responded gave several responses to this question so the sum of the responses is greater than 16. The most common response (given by 13 people) was that primary source analysis and interpretation of primary sources were the most important student skill outcomes. One middle school teacher said, "Teaching students to interpret and formulate opinions using primary sources as evidence." Another teacher said, "It moves away from the textbook and toward students examining pieces of history (like primary and secondary sources) and drawing their own conclusions."

Many of the same middle school teachers (4) and two other teachers also reported that the students' learning to synthesize what they have learned was one of the most important skill outcomes. One teacher said, "That they are able to synthesize all of the knowledge and put it towards one overall essential question each time they study a time period." Three teachers mentioned the active nature of learning that is part of the GWHI method, with one teacher saying, "Using primary documents and getting the students involved makes learning come alive for students." Another teacher said, "I believe students learn to think for themselves and become active learners rather than complacent in their learning." Three teachers mentioned the development of general higher order thinking skills.

Three middle school teachers said timelining skills were an important outcome and two teachers mentioned the practice of students learning to question textbooks rather than accept textbooks as truth. Individual teachers said important outcomes included: understanding cause and effect, mapping skills, reading for background knowledge, cooperation, appraising value of sources, meeting deadlines, and gaining "a sense of self-worth" through their knowledge.

High school teachers were asked the same question regarding the most important skills outcomes for their students. Nineteen of the 21 high school respondents answered this item for a response rate of 90%. Echoing the middle school teachers, the most common response among high school teachers (given by six people) was that the most important student skill outcome is the ability to analyze and effectively

use primary sources. An example of these responses was the high school teacher who said the most important skill was to be "able to examine primary sources with a critical and understanding eye." Five high school teachers said that the process of timeline construction, along with its accompanying skills of understanding chronology and cause and effect, was the most important skill.

Three high school respondents made similar general statements about the importance of analytical skills, including the ability to think critically and the ability to think like a historian. Three high school respondents said that the most important outcome is building students' confidence in their knowledge and their ability to use that knowledge to interpret what they read. One teacher said, "It means you have to make a commitment to an idea/value…and many students will not do that since they lack confidence in their knowledge to take a position or stance."

Two high school teachers said that all five of the processes learned in the GWHI project are the most important. Map interpretation and learning to synthesize information were mentioned by two people each as being the most important skills. One of these respondents said, "The concept of 'putting the pieces together.' This skill applies to so many different aspects/subjects of learning, interacting, living! It provided some great ways to get students to do this."

Individual high school respondents said that important outcomes included students' acquiring more motivation to learn and developing the habit of thinking critically about sources of information.

Teachers were asked to rate the utility of the five GWHI historical skills for their own instruction. The scale they used was a five point scale with "1" being "Not at all useful" and "5" representing "Very useful". Table 27 reports mean utility ratings for both middle and high-school teachers, disaggregated by whether or not the teacher is responsible for US history and also by cohort for middle-school teachers. Once again, high-school teacher responses were not disaggregated by cohort because mean ratings did not differ across cohorts.

Table 27. Mean rating of perception of usefulness of GWHI historical skills in teaching (5 point scale)

Skills	Middle School Cohort 1		Middle School Cohort 2		High School Both Cohorts		**Grand Mean[a] Across Teacher Groups**
	Teach US	Non-US	Teach US	Non-US	Teach US	Non-US	
	n=5	n=3	n=9	n=1	n=13	n=8	
Constructing timelines	5.00	4.33	4.78	5.00	4.50	4.30	**4.59**
Mapping for understanding	5.00	3.67	4.33	5.00	4.33	4.00	**4.31**
Primary source analysis	4.60	5.00	4.89	4.00	4.91	4.88	**4.84**
Reading for background knowledge	4.40	4.67	4.89	5.00	4.90	5.00	**4.84**
Synthesizing multiple sources	4.20	4.33	4.44	4.00	4.64	5.00	**4.57**
Grand Mean Across Skills	**4.64**	**4.40**	**4.67**	**4.60**	**4.66**	**4.64**	

[a] Grand means are weighted by sample size.

Across all skills and all groups, utility ratings were quite high, with all skills rated by all groups as at least a 4 (representing "useful"), with the exception of one skill: Cohort 1 middle school non-US history teachers rated mapping as moderately useful to useful (3.67). Mean ratings across all skills were extremely similar for all groups of teachers (with means of approximately 4.5). Across all groups of teachers, the skill rated the most useful was "Reading for background knowledge" (mean=4.8) and the skill rated the least useful was "Mapping for understanding" (mean=4.3).

Teachers were asked to rate the likelihood that they would continue to use the GWHI historical skills as part of their teaching strategies. Teachers rated the strength of their agreement or disagreement on a 6-point Likert-type scale with the statement, "After the GWHI project has finished, I will continue to use the following as one of my history teaching strategies." Table 28 reports these results for both middle and high school teachers, disaggregated by whether or not the teacher is responsible for US history and also by cohort for middle school teachers. Similar to the above analyses, high school teacher responses were not disaggregated by cohort because there were no cohort differences in mean ratings.

Table 28. Mean ratings of teachers' plans to continue use of GWHI history teaching strategies (6 point scale)

Skills	Middle School Cohort 1		Middle School Cohort 2		High School Both Cohorts		**Grand Mean[a] Across Teacher Groups**
	Teach US	Non-US	Teach US	Non-US	Teach US	Non-US	
	n=5	n=3	n=9	n=1	n=13	n=8	
Constructing timelines	5.00	4.00	5.13	6.00	4.82	5.29	**4.98**
Mapping for understanding	5.00	4.00	4.75	5.00	4.82	5.29	**4.86**
Primary source analysis	5.00	4.33	5.38	1.00	5.50	6.00	**5.30**
Reading for background knowledge	4.60	4.00	5.00	6.00	5.50	6.00	**5.27**
Synthesizing multiple sources	4.40	4.00	4.75	5.00	4.82	6.00	**4.93**
Grand Mean Across Skills	**4.80**	**4.07**	**5.00**	**4.60**	**5.09**	**5.72**	

[a] Grand means are weighted by sample size.

As can be seen, almost all ratings were high. In particular, all skills except one received mean ratings of at least 4 (representing "slightly agree"). A single skill (Primary source analysis) was rated a 1

("strongly disagree") by a single teacher. Across all skills, the highest agreement ratings were assigned by non-US history teachers at the high school level (mean=5.7), whereas the lowest agreement ratings were provided by middle school non-US history teachers in Cohort 1 (mean=4.1). Cohort 1 middle school US history teachers were more likely to plan to continue to use the skills than were their non-US history teaching peers, and the reverse was true among the high school teachers with high school non-US history teachers more likely to continue to use the skills than their US history teaching colleagues. Across all groups of teachers, the skills receiving the highest agreement ratings were "Primary source analysis" (mean=5.3) and "Reading for background knowledge" (mean=5.3). The skill receiving the lowest agreement rating was "Mapping for understanding" (mean=4.9).

Teacher and Student Motivation to Learn History

GWHI participants were also asked to rate their own motivation to teach and learn history and their perceptions of their students' interest in and motivation to learn history. Teachers were asked to use a 6-point Likert-type scale (where "6" was "Strongly Agree" and "1" was "Strongly Disagree") to rate the strength of their agreement or disagreement with four statements concerning their own and their students' motivation and interest in learning history. Table 29 reports the results.

Table 29. Mean teachers' ratings of motivation to teach history and their perception of students' motivation and interest in history (6 point scale)

Items	Middle School				High School		**Grand Mean[a] Across Teacher Groups**
	Cohort 1		Cohort 2		Both cohorts		
	Teach US n=5	Non US n=3	Teach US n=9	Non US n=1	Teach US n=13	Non US n=8	
I am more motivated to teach history	5.00	5.33	5.75	5.00	5.60	5.25	**5.45**
I am more motivated to learn history	5.75	5.33	5.89	6.00	5.64	5.50	**5.67**
My students are more motivated to learn history	4.60	5.33	5.57	4.00	4.64	4.71	**4.90**
My students are more interested in learning history	4.75	5.33	5.57	4.00	4.55	4.71	**4.89**
Grand Mean Across Skills	**5.02**	**5.33**	**5.69**	**4.75**	**5.11**	**5.04**	

[a] Grand means are weighted by sample size.

Teachers' agreement ratings were all quite high. Across all groups, teachers at least moderately agreed that their own motivation to teach and learn history had improved. All groups also agreed that their students' interest and motivation to learn history had increased, although teachers tended to agree less strongly that student interest in and motivation to learn history had increased.

New collaborations, barriers to success, changes in teaching styles, and other outcomes of project participation

All project participants were asked in what ways new collaborations that arose from the GWHI project were important to them. Sixteen of the 18 middle school respondents answered this item for a response rate of 89%. Nearly all of those responding agreed that collaborations were important to them, with one saying it was hard to continue collaborations because everyone is so busy and another saying that five months after the final workshop, they still collaborate on a weekly basis. One teacher who had not previously taught history said they had not formed any new collaborations because they were "too busy trying to learn the new curriculum." A few typical comments from middle school teachers included:

- I really enjoyed working with people who are 'like me' – they get history and are as pumped up about exploring new ways to analyze it and make better lessons. I don't feel like I'm 'alone' out there.
- I have picked up new ideas from others to use in my classroom. I also have a bank of people I can contact if I need ideas.
- I have received many lesson ideas, support and knowledge from these teachers. It is nice to work with someone as passionate as I am about teaching history and making it meaningful for students.

Among the high school teachers, fifteen of the 21 respondents answered this item for a response rate of 71%. The most common response given by eight high school teachers was that collaborating with other teachers was the most beneficial aspect of the project. One teacher said, "They are invaluable. I am sharing ideas with other teachers more now than ever before." Two people said that not only do they continue to work with their GWHI colleagues, but that project participation has also encouraged them to collaborate with other teachers in their school more often. One teacher said, "I have shared with people in class and others not in the class. I just talk to others more about what they do and I do."

Two high school teachers said that collaborating really helps with one of the problems of teaching well – having enough time to do all the research and planning new lessons. Two high school participants said that the collaborations have helped in situations where they need a little help figuring out how to approach a topic or in troubleshooting their lesson ideas. One teacher reported that working with colleagues made them feel more confident and knowledgeable. However, another teacher said although they had felt an initial "strong connection," after going back "to your own world, these relations easily fade."

All survey respondents were asked to describe the general and specific ways in which their teaching of history and social studies had been affected by their GWHI project participation. Sixteen of the 18 middle school respondents replied to this item for a response rate of 89%. As mentioned previously, this group of middle school teachers included several teachers who had never taught history

before and some of the responses reflect this situation. Three teachers said they learned most of what they know about history and teaching history from their participation in the GWHI project. One teacher said, "I learned a majority of what I know and teach in this program – I started it right when I began teaching middle school social studies (I was an elementary education major) and have really grown to love teaching American history because of the rich resources and ideas that this program has offered me!" Another teacher said, "The preparation was invaluable, as I've had no preparation in this."

Several middle school teachers commented on more than one way that their teaching has been affected by program participation, so the total does not add to 19. Other responses are shown in the list below (with the frequency of the response if greater than one in parentheses):

- Increased use of primary sources (5)
- More thoughtful approach to lesson design (4)
- Have more resources and are more able to locate good sources for teaching (3)
- Less use of textbook (2)
- Increased use of timelines (2)
- Teach students to corroborate sources (2)
- Increased use of narratives
- Increased use of maps
- Pleasure in being around people who teach history
- Renewed passion for teaching
- Take more time to explain to students
- More content knowledge
- Feel more competent at teaching
- Increased interest in history
- Use new methods that capture students' interest

One teacher said, "I question my text. I find primary source materials on my own to give depth to the text or in some cases to challenge it. I hope my students 'get that' – don't accept what you are told – independently verify things."

High school teachers also were asked the same question; 19 of the 21 respondents answered this item for a response rate of 90%. The most common answer was a general answer (given by nine people), that they are using a larger variety of methods and tools in their teaching than they did before project participation. Several of the high school teachers expanded upon this answer to give specific examples with individuals saying they use more document-based questions, take a more social/human approach, and use a more active/hands-on approach. One high school teacher mentioned using cowboy songs to

study western life and said, "I consider looking at things now with the idea and eye of the other side of the issue – using the personal history of individuals has made it more interesting for me and my students."

Four high school participants said that they felt more knowledgeable about historical content as a result of project participation. Four participants said they use primary sources more and textbooks less since they have been part of GWHI. One high school respondent said, "I have stopped teaching 90% of what was in the textbook and increased focus on two themes – the Civil War and World War II. I was able to explore each of these in depth." Two high school teachers said that they work harder now on promoting understanding of history. Two respondents said they focus more on "historical thinking." Individual teachers said they are more organized, more motivated to teach, and are now more likely to teach history "as a process" instead of an answer and to give their students increased independence in learning.

Survey respondents were asked about any possible barriers or impediments to continuing to teach using the GWHI methods. Twelve of the 18 middle school survey respondents answered this item for a response rate of 67%. Two middle school teachers said that they did not anticipate any barriers to using the GWHI methods. Two people said that they will not be teaching history next year, with one of these saying "If I did teach social studies then I would continue with the GWHI methods, I am a fan." Six respondents said that the main barrier was time. One middle school teacher said, "Time, time, time – teachers live their lives by the 50-minute period and administrators won't listen."

Three middle school teachers said that the students' abilities and lack of knowledge do not allow them to use the GWHI methods. One teacher said, "Time and student background. If I actually stopped to give the student the background necessary, I could easily spend a year on two or three chapters." Another middle school teacher said that it is the combination of the students' ability levels and the nature of the GWHI methods that are a barrier, saying:

> Students are not always as willing to participate as we want to believe. I have some classes that students will not talk or share. It is very difficult to make lessons interactive. Many of the speakers came in and demonstrated how to create lessons for students. Adults then played the role of students, but was very unrealistic of what we could expect in class from our students. The background knowledge or lack of knowledge for many students is a struggle to overcome, as well as the fact that I have many non-readers in class, so all the activities have to be arranged according to this.

Among the high school respondents, 15 of the 21 survey respondents answered this item for a response rate of 71%. The most commonly mentioned barrier to continuing to use the GWHI methods, given by six high school respondents (40% of those responding), was that it is hard to find the time to do new things. One respondent commented, "For me, finding the time to implement new materials is difficult, but where there is a will there is a way." Two high school teachers said that their challenge would be the amount of content that they needed to cover in short class periods over a year. Two high

school teachers said that their students' reading ability made it difficult to use the primary sources as part of the GWHI pedagogical approach. One teacher said, "Historical documents tend to have an 'era' style of writing and many students get *lost* in following this very *wordy* type of literature or editorial conversation with the reader." Other potential barriers mentioned by individuals were finding appropriate resources may be a barrier, that mapping in particular is difficult to work into the curriculum, and that students are not "aggressive learners" which makes some of the GWHI approaches challenging. Four high school respondents said that they did not foresee any barriers to continuing their use of the GWHI methods.

All survey respondents were also asked if they had any general comments concerning teaching history or participating in the project. Fourteen of the 18 middle school teachers responded to this item for a response rate of 78%. All of the responses to this item were "thank you" or other positive statements about the project. A few typical comments included:

- I have been able to teach many lessons without textbooks and expensive resources – lesson where my students are actively involved and thinking about history.
- Thank you for including me. It has made me a better teacher.
- I would love to be part of a Round 2 if there is one!
- Thanks! You made an impossible year almost doable. [Comment from a teacher who had never taught history before.]
- It was the most beneficial staff development I have ever been involved in.
- I would recommend this program to anyone and everyone – it's great.

Ten of the 21 high school respondents provided additional comments for this question. All but one made positive comments concerning the project. The only negative comment was "I don't think this project was organized very well." One person in addition to making a positive comment said that they would be interested in a workshop geared toward writing curriculum that aligns with the Iowa Core Curriculum. Some excerpts from other comments included:

- It has been a great experience!
- Thanks for all the program has done for us and for your dedicated commitment to positive education. We learn by example and these have been solid.
- Thanks for your time and support it's been tremendous.
- I enjoyed working with all the GWHI staff. Elise has been an angel at all requests and needs to make things work in and out of the classroom. I appreciate Colin also for showing that teachers can be active learners of the subject they are teaching….It has been a first class experience and I would encourage any history teacher to get on board. I appreciate the professional people I have

been exposed to – folks I would normally never had a chance to meet and discuss history topics with.

4.1.2 Content knowledge

All participants in the GWHI project were asked to take a content knowledge assessment at the beginning of their participation and again at the end of the project. The assessment consisted of a combination of multiple choice and short answer items. Items originated from two different sources: selected retired NAEP items and items generated by GWHI presenters, specifically to fit with topics they planned to address during the summer institutes. Between the pretest and the posttest, pretest scores were analyzed and items that were extremely easy were deleted from the posttest and pretest scores were adjusted in this analysis to remove those items. (Scores are therefore relatively lower than they would have been if those items had remained on the test.) Detailed reports concerning item type and origin and item discriminations and difficulties are included in the appendices to this report.

Table 30 reports the content knowledge assessment results for middle school teachers. Middle school teachers in both cohorts demonstrated improvement from pre- to post test. Because of the small number of teachers involved, it is difficult to achieve statistical significance, although Cohort 2 gains from pretest to posttest were significant at the $p<.05$ level. However, Cohort 2 teachers' posttest mean was still lower than the Cohort 1 pretest mean. (Note: Cohort 2 included the group of teachers who had not taught history prior to the start of the GWHI project.)

Table 30. Middle School Teacher Content Knowledge (Raw Scores)

Teacher Group	N	Pre-Test Scores Mean (SD)	Post-Test Scores Mean (SD)	Change Scores Mean (SD)
Cohort 1	9	24.33 (7.12)	27.56 (5.68)	3.22 (5.00)
Cohort 2	8	18.75 (7.98)	23.62 (6.74)	4.87* (4.29)

Note. The maximum number of points possible on the middle school teacher content knowledge assessment was 43. *Indicates difference was significant at $p<.05$ level.

Table 31 reports the content knowledge assessment results for high school teachers. As with the middle school teachers, teachers showed improvement from pretest to posttest, however in neither cohort was the difference statistically significant. The high school teachers' test scores were slightly higher in general than the middle school teachers with posttest scores of about 67% correct as opposed to 55% for the middle school posttest. Cohort 2 high school teachers followed the same pattern as the middle school teachers in that they made a larger mean gain from pre- to posttest than their Cohort 1 peers, however, their mean posttest scores were lower than the Cohort 1 pretest mean.

Table 31. High School Teacher Content Knowledge (Raw Scores)

Teacher Group	N	Pre-Test Scores Mean (SD)	Post-Test Scores Mean (SD)	Change Scores Mean (SD)
Cohort 1	16	45.69 (6.70)	47.44 (6.13)	1.75 (8.18)
Cohort 2	5	36.00 (7.00)	40.60 (10.78)	4.60 (7.57)

Note. The maximum number of points possible on the high school teacher content knowledge assessment was 63.

4.1.3 Teacher Lesson Planning Survey

At the beginning and end of project participation, teachers were asked to complete the Lesson Planning Survey. The open-ended survey provided a means for examining changes in teachers' approaches to lesson planning. Table 32 reports the results for teachers on the Lesson Planning Survey. The survey was scored blind by the project director using an eight dimension analytic rubric with two points possible for each dimension, for a maximum possible score of 16.

Table 32. Teacher Lesson Planning Survey

Level	N	Pre Mean	Pre SD	Post Mean	Post SD
Middle School	13	7.54	2.57	10.46	3.02
High School	20	7.85	2.47	9.70	3.77

Of the 13 middle school teachers, ten teachers' lesson plans improved from pre- to post-survey, two teachers' plans received lower scores at the end of the project than at the beginning, and one stayed the same. The mean difference score for middle school teachers from pre to post was 2.92 (SD=3.86). Of the 20 high school teachers who completed both pre- and post lesson planning surveys, 15 teachers' lesson plans improved, three teachers' plans scored slightly lower at the end of the project, and two were the same. The mean difference score for high school teachers from pre to post was 1.75 (SD=2.81).

In reviewing the lesson plan surveys, the project director characterized the changes in middle school teachers' lesson planning approaches as demonstrating growth in their use of and emphasis on primary sources and historical thinking skills. Middle school teachers were more likely on the post tests to specify the primary sources they would use as lesson centerpieces, and described more specifically the activities and guides they would incorporate to scaffold student engagement in the sources. Middle school teachers at post-test were more likely to reveal their growing sense of history as interpretive and evidence-based by their use of terms such as "analyze" and "perspectives" that appeared less frequently

or not at all on the pre-tests. Middle school teachers' post-test use of the term "timeline" suggested a change in their attention to scaffolding student understanding of chronology and the connection of detailed event to categorical or temporal concept and era.

As suggested by the quantitative scoring of the lesson planning survey, high school teachers' approaches did not appear to change as much over the course of the project, however, some growth was seen, particularly in the specificity of their discussions of primary sources. On the post- test, high school teachers more frequently cited by title the sources they would use in lessons, and aligned the sources with more clearly delineated historic topics than appeared on the pre-test. These trends suggest that a majority of GWHI high school teachers have improved in their understanding of history and how to teach it as an interpretive, evidence-based discipline.

4.2 Student Outcomes

Student outcomes were assessed using two separate assessments that were administered as pre- and posttests by all GWHI participating teachers to students in all of their classes (or at least two sections if they taught more than two sections). The assessments included: 1) a content knowledge assessment, which was a subset of the items from the teacher content knowledge assessment, and 2) a skills-based assessment that required students to use primary source photos or documents, and/or tables of statistics to answer open-ended questions using both the information with which they were provided and their own background knowledge. Both types of assessments are described in detail in the methods section. Tables showing complete results, item discriminations and difficulties for content knowledge items, and complete protocols for scoring open-ended items on both the content knowledge and skills assessments are included in the appendices to this report. The most relevant outcomes are reported below.

4.2.1 Content knowledge

Table 33 reports the results for the Student Content Knowledge Assessment. Classrooms were randomly-assigned to receive either Form A or Form B as the pre-test and the opposite form as a post-test. This design was intended to diminish possible practice effects. Although *a priori* measures were taken to create two versions of the test that were as similar as possible on several dimensions (e.g. item format, item topics, item origin) the two versions of the tests were not equated. Therefore, scores for both tests were converted to standard scores so that they could be pooled across test versions for pre-post comparison. Standard scores for students who took the tests in 2007-08 and 2008-09 at both middle and high school levels are reported in Table 34.

Table 33. Student Content Knowledge (Standard Scores)

Grade Level	Year	N	Pre-Test	Post-Test	Change Scores
			Mean	Mean	Mean

			(SD)	(SD)	(SD)
Middle School	2007-08	453	46.38	54.05	7.67*
			(7.52)	(10.61)	(11.25)
	2008-09	229	45.23	54.38	9.15*
			(7.62)	(9.96)	(9.14)
High School	2007-08	372	47.60	54.13	6.59*
			(7.94)	(9.71)	(8.72)
	2008-09	164	46.00	53.94	7.94*
			(7.60)	(10.57)	(9.01)

Note. All scores were standardized to permit comparisons across forms A and B. Standard scores have a mean of 50 and a standard deviation of 10. *Indicates change is significant at the p<.0001 level

For both years and at both grade levels, students demonstrated significant improvement from pre- to post-test. All difference scores were significant at the p< .0001 level. Table 34 reports the number and percentage of students who demonstrated improvement from pre- to post-test during each of the project years.

Table 34. Number and Percentage of Students Demonstrating Growth on Content Knowledge Tests

Year	Middle School		High School	
	Number	Percentage	Number	Percentage
2007-08	343	76	284	76
2008-09	192	84	137	83

4.2.2 Skills Assessments

All students took a skills test as a pre-test and a posttest during both project years. Once again, classrooms were randomly assigned to receive either Form A or Form B as a pre-test and the opposite form as a post-test. Only a sample of the skills tests completed during Year 2 were scored (See methods section for descriptions of the tests, and sampling and scoring procedures). Table 35 reports the results of the 2008-09 skills tests for the sample of middle school and high school students. Standard scores were computed for each form to enable scores from the two different versions of the tests to be compared. Middle school students' performance on the skills tests improved significantly from pre-to posttest (p<.0001) irrespective of the order of the two forms of the test. High school students' performance did not change significantly from pre- to posttest.

Table 35. Student Skills Tests (Standard Scores)

Student Group	N	Pre-test Scores	Post-test Scores	Change Scores
		Mean	Mean	Mean

		(SD)	(SD)	(SD)
Middle School	170	43.89	54.11	10.22*
		(9.29)	(8.85)	(10.30)
High School	107	50.05	50.03	-0.02
		(9.86)	(10.32)	(11.57)

Note. All scores were standardized to permit comparisons across forms A and B. Standard scores have a mean of 50 and a standard deviation of 10. *Indicates change is significant at the p<.0001 level

Not surprisingly given these findings, most middle school students demonstrated growth on the skills tests (141 of 170, 86%), and only about half of the high school students demonstrated growth (51 of 107, 48%).

5. Discussion and Conclusions

5.1 Teacher Outcomes

The main objectives of the GWHI project concerning teachers were to increase teachers' content knowledge in history, improve teachers' pedagogical knowledge for teaching history, increase teachers' motivation to teach history using methods that enhance students' abilities to think historically, and enhance teachers' abilities to mentor other teachers in history pedagogy.

The evaluation of the GWHI project conducted by the Center for Evaluation and Assessment examined the extent to which these goals were achieved in XX ways, including:

- Surveys concerning teachers' opinions about professional development activities
- Surveys concerning teachers' implementation of historical content and pedagogy
- Historical content knowledge assessments – before and after participation in the GWHI project
- Observation of professional development activities including teacher/mentor led sessions.

The methodology and results section of this report provided detailed information about the evaluation findings and the instrumentation used to examine the project's outcomes. The purpose of this section is to look more closely at the outcomes and describe the implications and limitations of the evaluation.

Teacher Content Knowledge

We looked at teacher content knowledge directly through the use of teacher content knowledge tests administered before and after project participation, and indirectly through survey items (throughout the project) asking teachers to provide feedback concerning the project's effect on their historical content knowledge. Assessments were constructed to measure teacher content knowledge through the use of a combination of retired NAEP items and project presenter-constructed items, chosen to reflect the historical content and skill set addressed by the GWHI project. Tables 30 and 31 report that although all groups of teachers (middle and high school, cohorts 1 and 2) demonstrated improvement from pre- to

post-test on the content knowledge assessments, for only one group, cohort 2 middle school teachers, was the difference statistically significant.

There are several factors to consider when looking at these results. In designing the GWHI assessments to measure content knowledge growth (variations of which would be used for both teachers and students), we chose to include retired NAEP items to learn more about whether growth in teachers' historical content knowledge could be examined using readily-available, nationally standardized test items, and to include items written by presenters specifically to be sensitive to changes in teachers' knowledge of historical content addressed during GWHI professional development. Clearly, NAEP items were written for *students* at the grade levels in which the GWHI participants *teach*. However, widely-reported poor student performance on NAEP tests [in fact, a guiding purpose for the TAH program!] suggested that these items might have performed better as measures of teachers' content knowledge.

However, in fact, the assessments were of poor quality for the intended purposes. Multiple choice items, (both NAEP and presenter-generated) tended to discriminate poorly, and NAEP MC items were too easy for the teachers, despite the fact that the easiest items at pretest were removed from the posttest. Short answer items (both NAEP and presenter-generated) provided somewhat better discrimination, especially at the middle school level, however most short answer items were quite easy for the participants. NAEP SA items discriminated poorly at the high school level, some with negative discrimination.

It is not clear what other problems were inherent in the items or the rubrics for scoring the short answer items. The retired NAEP items that appeared on the test were chosen because of their historical content as a first criterion and/or because of their fit with the skill set emphasized by the GWHI program (e.g. use of primary source document such as photo, cartoon, or quotation), *not* because of their technical qualities. According to the technical data provided by the NAEP tools website, some of the items were either very easy or very difficult for 8th and 12th grade students, and in practice, some of the rubrics were difficult to employ and did not make useful distinctions between good, adequate, and poor responses.

Other factors may have limited the capacity of the content knowledge assessments to provide information about growth in teacher content knowledge. Despite our attempts to coordinate the content tested with the content addressed during professional development workshops, the match between what was presented and what was tested was ultimately not as close as we had hoped. Two of the presenters who wrote the original items (and provided instruction during Year 1) did not return for the second year of programming and the new presenters were not asked to adhere closely to the original content presented. One presented on the Vietnam War rather than the Cold War as originally planned. Another presenter was unable to attend at the last minute and the substitute presented on teaching *September 11*, rather than a more general view of 21st century history.

At the end of each GWHI professional development workshop, participants completed surveys concerning their participation in the workshop, including their self-efficacy concerning particular historical content knowledge. For example, when asked about confidence in their own knowledge of the content areas addressed during the professional development, middle school teachers expressed an average gain in their confidence in their content knowledge of 21% and 32% (2007 and 2008 data, respectively) and high school teachers expressed an average gain of 12% and 20% (2007 and 2008) after attending the four-day professional development institutes each year. These large gains in participants' self-efficacy concerning historical content knowledge suggest that there were outcomes to which the content assessments were not sensitive. [Complete results of workshop surveys have been reported in previous reports.]

Pedagogical Practices

We looked at changes in the content and pedagogy that teachers employed before and after participating in the GHWI project through the use of several surveys that approached teachers' classroom behaviors in different ways. The main survey used to examine changes in classroom practices was the Teacher Implementation Survey. Parallel versions of the survey for middle and high school were written to investigate the historical content and historical skills taught, and the pedagogical strategies teachers used during instruction. As reported in Table 17, mean ratings of thoroughness of teaching increased from before to after participating in the GWHI project, across grade levels and content area. We anticipated that when rating their thoroughness of teaching historical topics, teachers would be more likely to report increased thoroughness of instruction when teaching the topics that most closely aligned with the GWHI instruction, but not necessarily when teaching other topics. On average, this was the case among middle school teachers who taught history, but in general was not true among high school teachers. Irrespective of whether they taught US history or not, high school teachers rated the increase in their thoroughness of teaching GWHI topics about the same as other topics, with modest mean increases across topics.

There are several alternative explanations for these outcomes. First, while retrospective pre-post measures are thought to provide good estimates of change over time, when used in this situation, they may be subject to socially-desirable responding. That is, among this participant group who reported great satisfaction with project participation, their inclination may have been to show uniform increases in thoroughness of implementation (and in perception of student competencies in skills and knowledge) across items. However, there is considerable variation across items in the change from *before* to *now*, and, especially for middle school teachers, between GWHI and non-GWHI topics. Participants also used the entire scale of values to characterize their thoroughness both *before* and *now*. Further, among all teacher groups, increases in teachers' perception of student competencies at using skills from *before* to

now, which were the project's main emphases, were larger than differences in other areas. These findings bolster the argument that teachers were not merely responding in a socially desirable way, but were making real distinctions between items.

Another factor that may be at play in interpreting these outcomes is that teachers *especially at the high school level,* experience conflicting demands on how they are expected to teach history. On the one hand projects like the GWHI project advocate teaching "deeply", placing an emphasis on providing students with the skills to be historians. At the same time however, broader curriculum mandates often encourage more surface "coverage" of vast amounts of historical content. This could explain why high school teachers (whose curriculum may be more likely to be designed to meet specific content standards) were less likely to show variation between GWHI and non-GWHI topics – newly sparked interest in teaching history more in-depth and expanding students' ability to think critically may be balanced with the reality that they must continue to teach all the topics they have taught in the past. As reported in the results section of this report, when asked about barriers to using the GWHI methods, high school teachers' number one response was that they do not feel they have the time to use these methods, with two teachers saying the amount of content they were responsible for covering made it difficult to teach the GWHI way.

Teachers' perceptions of how well their students would perform on both content knowledge tasks and use of historical skills also provides a measure of their self-efficacy for teaching history. Teachers overwhelmingly expressed that their students had improved in their ability to use the historical skills emphasized by the GWHI project. In addition, both middle and high school teachers reported that their students' content knowledge had increased, with middle school history teachers' mean ratings increasing the most. Teachers also reported that the GWHI skills are beneficial for students and useful in their teaching, with middle school and non-US history high school teachers showing particularly strong support for using the skills for teaching and learning history.

Results of the Teacher Lesson Planning Survey revealed that teachers, especially middle school teachers, became more able to explain how they would use primary sources in their instruction. They were more likely to structure lesson plans so that students would analyze multiple sources and be more aware of perspective, and that they as teachers would provide appropriate scaffolding for history explorations. High school teachers' lesson plans did not improve as much, but still showed signs that they were more able to select primary sources that supported students' learning and use sources in a way that encouraged deeper understanding.

5.2 Student Outcomes

The main student objectives of the GWHI project were to increase students' content knowledge in history and to improve their historical thinking skills, defined as the ability to use primary source documents to source, contextualize, and corroborate their own background knowledge and to craft a written response to questions about a particular historical era. Students' historical content knowledge was investigated with a content knowledge assessment that included both multiple-choice and short-answer items. These items were either retired NAEP items or were presenter-generated items and were selected/written to reflect the historical content emphasized during professional development workshops. Historical thinking skills were documented with original assessments designed by the project director. These assessments included two items, each embedded within a historical context, which required students to examine a photograph, read a primary source passage, and/or examine a table of statistics and then answer several open-ended questions about the content, using their own historical background knowledge to help them answer the questions.

Student Content Knowledge

Results from the student content knowledge assessment indicated that students at both grade-levels and during both time periods improved their historical content knowledge, with all student groups demonstrating statistically significant growth from pre- to post-test. For the 2007-08 group of students, this translated to more than three-quarters of the students demonstrating growth and for the 2008-09 group of students, this translated to approximately 83% of students demonstrating growth. Middle-school student gains tended to be larger than the gains of their high-school counterparts.

Student Historical Thinking Skills

Results from the student historical thinking skills assessment were mixed, with middle-school students demonstrating statistically significant improvement from pre- to post-testing and high-school student scores remaining essentially flat from pre- to post-test. One possible reason for such mixed results with respect to historical thinking skills is the generally poor quality of the historical skills test at the high-school level. Whereas the middle-school skills tests featured internal consistency reliability estimates of .58 and .43, respectively, for Forms A and B, the corresponding estimates at the high-school level were only .08 and .21. In general, high-school composite scores on both forms of the test demonstrated less variability than the corresponding composite scores at the middle-school level. In addition, individual items on the high-school skills tests discriminated much more poorly than those on the middle-school skills test (with average item discriminations of .10 for the former and 0.34 for the latter). Whereas all individual items on the middle-school skills tests demonstrated impressive

instructional sensitivity, items at the high-school level were not sensitive to instruction. Finally, all items on the high-school assessments were more difficult than those on the middle-school assessments. This difference in item difficulty may be one reason for the lower item discriminations on the high-school assessments. In turn, differences in difficulty between items on the middle-school and high-school assessments may be due to the specific primary sources used as prompts, with much steeper reading requirements for the documents used on the high-school assessments. Results from teacher implementation surveys are supportive of the hypothesis that high-school students' poor performance on the skills assessments was related to inadequate alignment between what teachers taught and what was assessed. In particular, high school US history teachers indicated at the end of the project that they had taught most of the content sampled by the thinking skills assessments only "moderately well."

Discussion of Student Outcomes

Evaluation results were mixed with respect to student outcomes. Whereas content knowledge assessments composed of multiple-choice and short-answer items were sensitive to the intervention, demonstrating significant improvement from pre- to post-testing among both middle- and high-school students, historical thinking skills assessments that consisted of more extended written responses produced conflicting interpretations between the middle- and the high-school levels. Thus, it appears that middle-school students improved their historical content knowledge and historical thinking skills, whereas high-school students improved only their historical content knowledge. However, there are a few alternative explanations for these results. First, as discussed above, the psychometric quality of the high school skills tests was much lower than the quality of all other assessments. Thus, it is possible that the lack of significant improvement on the high-school skills tests is due to low reliability. Second, high school teachers' instruction may not have been completely aligned with the specific content and skills represented on the high school skills test. Thus, the instrument may not have been sensitive to the intervention. Finally, it is possible that high school students, who may be less intrinsically motivated to try their best on a low-stakes, un-graded test, simply put forth less effort than their peers at the middle-school level. This is certainly consistent with the larger gains in student scores on both types of assessments observed at the middle-school level. Moreover, it seems reasonable to conclude that a lack of student motivation would manifest itself more noticeably on an assessment requiring all written responses than on an assessment consisting of multiple-choice and short-answer items. This argument is consistent with the apparently contradictory high school results, where content knowledge appeared to improve at the same time that thinking skills appeared to stay the same. Thus, the preponderance of evidence suggests that students of teachers who participated in GWHI, particularly at the middle-school level, improved their learning in domains specifically targeted by the intervention.

This conclusion should be qualified, however, by noting that there were no estimates of comparison-group performance to which to compare student gains. Therefore, it is possible that students of teachers who did not participate in GWHI would have demonstrated gain scores of a similar magnitude, simply as a result of their regular classroom instruction. In fact, for the subset of NAEP items embedded within the historical content knowledge assessments it is possible to compare mean item performance for GWHI students and a nationally-representative sample of American students who were not exposed to the intervention.[2] At the high-school level, there were very few differences in mean item performance between GWHI students and the national sample. Items for which differences in group performance did appear were not consistently in favor of the GWHI group. In other words, GWHI high-school students performed better on a few items and worse on a few items when compared with the national NAEP sample. At the middle-school level, there were some differences in mean item performance between GWHI students and the national sample, with differences tending to favor the GWHI sample. In other words, for about half of the NAEP items on the middle-school assessments, GWHI students out-performed students from the national NAEP sample. The other middle-school items did not exhibit group differences. Taken together, student results suggest that the intervention was more successful for middle-school students with respect to improving content knowledge and historical thinking skills.

[2] This information is available on the NAEP website at the individual item level and in the Appendices to this report.