Grant Title: A Study of the Efficacy of Reading Apprenticeship Professional Development for High School History and Science Teaching and Learning

Cynthia Greenleaf*, Thomas Hanson*, Joan Herman**, Cindy Litman*, Rachel Rosen*, Steve Schneider*, and David Silver**

June 22, 2011
Final report to Institute for Education Sciences
National Center for Education Research
Teacher Quality/Reading and Writing

This work was supported by Grant # R305M050031 to:
Cynthia Greenleaf, PI
Joan Herman, Co-PI
and Steven Schneider, Co-PI

* WestEd
300 Lakeside Drive, 25th Floor
Oakland, CA 94612-3534

** University of California-Los Angeles
National Center for Research on
Evaluation, Standards, & Student Testing
300 Charles E. Young Drive North
GSE&IS Bldg. 3rd Flr./Mailbox 951522
Los Angeles, CA 90095-1522

EXECUTIVE SUMMARY

This report describes findings from an efficacy study examining the effects of professional development focused on integrating academic literacy instruction with content area coursework on teacher instructional practices and student achievement in reading and high school biology and US history. The research builds on a well-tested approach to literacy instruction, Reading Apprenticeship (RA) (Greenleaf, et al., 2001; Greenleaf, et al., in press), which integrates metacognitive inquiry into ongoing content area instruction to make explicit the tacit reasoning processes, strategies, and discourse rules that shape successful readers' and writers' work in the discipline. The instructional model draws together research-based practices in reading instruction, including methods of engaging students in extensive reading; integrating explicit teaching of comprehension strategies; establishing relevance and making personal connections to reading materials and curriculum activities; identifying and using a variety of text structures to support comprehension; and supporting collaborative sense-making activities with written materials. The central dynamic of this instructional model is routine metacognitive conversation; that is, talking about the reasoning and problem solving processes that accompany reading as students carry out learning tasks in the science curriculum.

A highly designed program of professional development in Reading Apprenticeship was the intervention explored in this efficacy study (Greenleaf & Schoenbach, 2004; Greenleaf, et al., in press). The professional development curriculum was designed to involve teachers in inquiry into their own science or history literacy practices and in close analysis of text and task demands, as well as inquiries into student literacy performances through videotapes of class and individual student reading activities, written case studies, and ongoing student assessment. Professional development was also designed to model target instructional approaches, engaging teachers in practicing metacognitive routines, modeling reading and reasoning processes, conducting small group work, engaging and supporting students in extended reading opportunities, and facilitating discussions that focus on how and why to read science or history texts as well as the content of these texts. During the professional development these instructional approaches were tightly integrated with core units of study in biology or US history to illustrate the integration of literacy and science and history learning.

The study targeted schools serving high numbers of students from populations underrepresented in higher education. A group randomized, experimental design was used to assess program impacts. The study relied on multiple measures of teacher implementation and student engagement and learning: (a) pre- and post-intervention surveys of teachers reporting their instructional practices and beliefs about reading, student learning, and student diversity; (b) teacher interview data about instructional practices, beliefs, and student engagement in literacy learning opportunities; (c) teacher practices as reflected by teacher assignments; (d) student opportunity to learn (OTL) surveys; (e) student performance on Integrated Learning Assessments (ILA), half of which embedded the Degrees of Reading Power test of reading comprehension while the other half contained an essay writing task, and (f) standardized test results derived from the California Standards Tests in English language arts, reading comprehension, and biology or US history. Hierarchical linear modeling procedures were used to estimate program impacts on teacher and student outcomes.

The multiple measures of teacher implementation give us a robust corroboration of teacher level outcomes. Teachers in the experimental group demonstrated increased support for history and biology literacy learning, increased use of metacognitive inquiry routines, increased reading comprehension instruction, and increased use of collaborative learning structures. In short, they were more able to integrate science and science literacy learning in classroom instruction.

Although the results for the Degrees of Reading Power (DRP) test and comprehension questions from the Interactive Learning Assessments – did not provide evidence that these differences in teaching resulted in learning differences for students, Interactive Learning Assessments show evidence that students in treatment classes approached reading differently than their counterparts in control classes. In particular, they annotated text more often, showing evidence of reading strategies and discipline-specific reasoning, compared to their counterparts in control classes. There were significant differences in *History Reading Strategies* (es = .76) scores between treatment and control students. African American (es = .99, p < .10), Latino (es = .73, p<.10), and White (es = .82) students in intervention classes demonstrated more frequent use of comprehension-supporting *Reading strategies* than those in control classes. History treatment students whose home language was not English (es = 0.65, p < .10) also showed greater use of comprehension-supporting *Reading strategies* than their counterparts in control classes. In biology, students in treatment schools scored higher on *Metacognition* (es = .27, p < .10). As in history, this impact was more pronounced for Latino and second language learners in biology. Latino students in treatment classes had greater *Metacognition* scores on this measure than their counterparts in control classes (es = .33). In addition, White students in intervention classes demonstrated more frequent use of comprehension-supporting *Reading strategies* (es = .86). Treatment students whose home language was not English (es = .38, p < .10) and students whose home language was English (es = .25, p < .10) both demonstrated greater evidence of *Metacognition* than their counterparts in control classes, but these increases were not statistically significant at conventional levels. Analyses of student annotations showed that students in treatment classrooms annotated the texts in the ILA far more frequently than those in control classes, and that text annotation was highly and positively correlated with comprehension of these texts.

Student Opportunity to Learn (OTL) surveys partially corroborated teacher reports of increased integration of literacy and content instruction and some sub-groups reported increased levels of motivation and effort in class as well as an increased sense of academic identity. The program impact on *Integration of History and Literacy* was positive for students whose home language was not English (es = .29, p<.10), but not for students whose home language was English, suggesting that these students perceived literacy and history instruction to be integrated to a greater degree than did similar students in control classes. The program impact on students' perceptions of the degree of *Integration of Biology & Literacy* instruction was positive and statistically significant for students whose home language was not English (es = .39) as well as for students whose home language was English (es = .37).

Further, the results for state-mandated criterion-referenced test scores offer some evidence that differences in teacher practice resulted in improvements in student academic performance. Two types of state standardized test score data were collected — linked, longitudinal test score data for students for whom we had obtained parental consent; and anonymous, unlinked, cross-sectional data for all students in participating classrooms. To enhance the precision of the impact estimates and to account for potential differences in pre-intervention characteristics between groups, the test score analyses controlled for student and teacher characteristics.

For the longitudinal test data, history students in treatment schools exhibited higher levels of performance on state standardized assessments in reading comprehension (es=0.16, p < .10) and history (es=0.19) than students in control schools. Biology students in treatment schools exhibited similar levels of performance on state standardized assessments as their counterparts in control schools, based on analyses of the longitudinal data. For the cross-sectional data, which was a more representative sample of the students in the study, both history and biology students in the treatment schools performed better than their counterparts in control schools on the state standardized assessments. For the history sample, students in treatment schools exhibited higher scores in history (es=0.25), reading comprehension (es=.22), and English language arts (es=0.26). An analysis of scores by demographic group found increases across all demographic groups in *English language arts* for students in history intervention classes, with effect sizes of 0.32 for African American students, 0.22 for Latino students, 0.50 for Asian students, and 0.25 for White students, though these comparisons were only statistically significant at conventional levels for African American, Asian, and White students. Latino and Asian students in treatment classes also show increased test scores compared to their counterparts in control classes on *Reading comprehension* and *History* CSTs. For the biology sample, students in treatment schools exhibited higher scores in biology (es=0.29, p<.10). Overall, the pattern of results for the cross-sectional test data suggests that the impacts are most consistent and robust for Latino students (es = .26 for *English language arts* and .35 for *Biology*) and for students who speak languages other than English at home (es = .29, .28, and .36 for *English language arts*, *Reading Comprehension*, and *Biology*, respectively). The results of the study thus present a positive picture with regards to the effectiveness of the Reading Apprenticeship framework for integrating academic literacy content with biology and history coursework and instructional practices.

A STUDY OF THE EFFICACY OF READING APPRENTICESHIP PROFESSIONAL DEVELOPMENT FOR HIGH SCHOOL HISTORY AND SCIENCE TEACHING AND LEARNING

INTRODUCTION

Our democracy and future economic wellbeing depend on a literate populace, capable of fully participating in the demands of the 21st century (CCAAL, 2010; Rutherford & Ahlgren, 1990). Yet NAEP results indicate that most American youth lack the skills to successfully engage in the higher-level literacy, reasoning and inquiry needed for an information generating and transforming economy (NAEP, 2006, 2007, 2009; Snipes & Horwitz, 2008). Further, a persistent achievement gap persists between mainstream populations and those outside of the mainstream (Donahue, et al, 1999; Gee, 1999; Jencks & Phillips, 1998; Lee, Grigg, & Donahue, 2007). Achievement gaps among different populations of students in reading are echoed by similar gaps in science learning and achievement (Grigg, Lauko, & Brockway, 2006; NCES, 2007). Instructional support for adolescents' literacy development throughout schooling, though long neglected, has recently gained acknowledgement and support through state and national funding and policy mandates (CCAAL, 2010). As a result, in the current policy environment, schools and districts are under increased pressure to place low-performing students into reading intervention programs with the well-intentioned goal of increasing their literacy proficiency, yet some researchers suggest that the skill-based instruction they receive may perpetuate low literacy achievement rather than accelerate literacy growth (Allington & McGill-Franzen, 1989; Greenleaf & Hinchman, 2009; Haycock, 2001; Hiebert, 1991; Hull & Rose, 1989; Lee & Spratley, 2010). Moreover, intervention programs designed to address reading problems increasingly result in lost opportunities to learn in other academic subjects, particularly science (McMurrer, 2007; Rentner, et al., 2006). Withdrawing adolescents from instruction in academic subject areas to remediate reading difficulties threatens to exacerbate historic inequities in achievement for populations of students traditionally underrepresented in post secondary education (Barton, 2003). There is therefore increasing urgency to investigate how the integration of reading instruction into content area learning at the high school level might advance the reading and achievement of underachieving youth.

This study advances the idea that we must think strategically about the integration of development across subject matter domains if we expect to develop students' multiple capacities, particularly those from groups who have been historically underrepresented in the sciences. Science and history classrooms conceivably can contribute opportunities for students to acquire greater literacy proficiency, and greater literacy proficiency also is essential to students' acquisition of deep scientific and historical understandings and inquiry skills. A key premise of this initiative is that scientific as well as historical inquiry practices share important properties with academic literacy practices in general, that make the integration of literacy and science or history learning particularly powerful (Cervetti, Pearson, Bravo, & Barber, 2006; Greenleaf, Brown & Litman, 2004; McMahon & McCormack, 1998; Pearson, Moje, & Greenleaf, 2010; Wineburg, 2001). Participation in investigation-oriented science relies on various kinds of sophisticated literacy skills - the ability to access scientific terminology, interpret arrays of data, comprehend scientific texts, and read and write scientific explanations (Norris & Phillips, 2003; Osborne, 2002). Similarly, the study of history requires the capacity to sift through historical documents with attention to bias and perspective, to construct evidence-based accounts of probable historical events, to place documents and artifacts into larger historical contexts, to evaluate the credibility of different sources of information, and to perceive and have empathy for the experience of others (Bradley Commission on History in Schools, 1995; Levi, 1989). These critical thinking skills represent the higher forms of literacy required in college and the workplace and envisioned in the Common Core State Standards.

In addition to addressing the challenge of developing students' literacy proficiencies in the context of science and history learning, this study also advances the idea that we must find effective means to develop the instructional capacity of secondary teachers if we are to succeed in advancing student capacities. Despite the recognized and widespread need for adolescent literacy development in the upper grade levels, very few classrooms provide the needed academic literacy instruction, particularly in the subject areas where it is most critically absent (CCAAL, 2010; Lee & Spratley, 2010). In large part, high school teachers lack the know-how to simultaneously build students' academic literacy skills and engage them in a rigorous curriculum of subject area study (Heller & Greenleaf, 2007; Shanahan & Shanahan, 2008). In lieu of helping students develop the proficiencies needed to read, write, and reason with the language, texts, and dispositions of science or history, typical instructional strategies for struggling readers involve simplifying, slowing the pace, and often abandoning more rigorous course work, virtually assuring low levels of achievement for students who are already behind (Dweck & Molden, 2005; Kamil, et al., 2008; Ness, 2008; Pearson, et al., 2010).

National policy responses to persistent achievement inequities have recently targeted the improvement of teacher quality by linking federal funding to state and district reform strategies (for example, Race to the Top, Investing in Innovation, and School Improvement grants), with policies focused primarily on recruiting and retaining better prepared teachers and distributing these human resources more equitably to schools serving low achieving student populations (www.ed.gov)). Yet teaching is a complex human endeavor, enacted with specific learners in interaction with specific content, materials, and tasks, and decades of research have demonstrated that effective teaching requires a capacity and disposition to learn through reflection while fully immersed in teaching (Ball & Cohen, 1999; CCAAL, 2010; Darling-Hammond & Sykes, 2003; Darling-Hammond, et al., 2009; Shulman, 1986). In addition to recruiting, retaining and distributing better-prepared teachers, advancing the academic literacy of all students will depend on identifying and providing effective means of building teacher quality through ongoing teacher professional development. Little systematic research has examined the effects of professional development on teacher practice and student learning outcomes in history or science (Garet, et al., 2001; Kennedy, 1998; Killion, 1998; Loucks-Horsley & Matsumoto, 1999), and those studies that exist have been done primarily in the elementary or middle school setting (Calfee & Miller, 2004; Fishman, et al., 2003; Romance & Vitale, 2001; 2006). Further, professional development endeavors vary widely in their structures, formats, and aims (Borko, 2004; Desimone, 2009; Guskey, 2002).

As some researchers have pointed out, professional development opportunities are distributed inequitably, with teachers serving the most vulnerable populations of students receiving training to implement narrow instructional strategies, follow pacing guides, or implement curricula to fidelity rather than professional development opportunities that help them build the flexible repertoires of practice they will need to develop the advanced academic proficiencies of a broader range of the students they serve (Ball & Cohen, 1999; Hargreaves, 2003; Sparks, 2004). As others have argued, professional development is key to standards-based reform (Darling-Hammond, et al., 2009; Fishman, et al., 2003). Yet in an era in which higher academic standards are envisioned for all learners, the narrow training regimens and implementation monitoring now prevalent in high poverty schools will fall far short of preparing teachers to meet the lofty instructional goals put forward in such documents as the new Common Core Standards (NCSSO & NGA, 2010).

In contrast to more narrow conceptions of professional development, the Reading Apprenticeship model of professional development we report on here aims to build teachers' generative capacity to engage with students in collaborative meaning making and problem-solving during ongoing instruction (Greenleaf, Brown & Litman, 2004; Greenleaf & Schoenbach, 2004; Greenleaf, et al., in press). The study focuses on the impact of professional development designed to foster science and history teachers' adaptive expertise (Bransford, et al., 2005; Lai, et al., 2009), building teacher capacity to surface and model effective ways to address comprehension problems that arise as the varied learners in the classroom interact with course materials such as biology textbooks, graphs and diagrams, and lab procedures in science, and history textbooks, primary source documents, graphs, maps, and artifacts in history. This study therefore has the potential to advance our knowledge of the role such a model of professional development can play in developing existing teacher quality, as well as the potential of such instructional improvement to achieve educationally meaningful increases in student learning outcomes at the high school level. The research base on these vital issues is woefully inadequate to inform policy and practice in educational reform (Desimone, 2009; Yoon, et. al, 2007).

Purpose of the Study

The study was designed to examine the effects of an instructional framework, Reading Apprenticeship, and an accompanying professional development model on teachers' ability to integrate disciplinary literacy practices into science teaching in high school biology and US history classes, exploring the resulting changes in teacher knowledge and instructional practices and student achievement in science and reading. This study was a unique collaboration between research and development staff responsible for science programs and literacy programs at a non-profit educational service agency as well as independent assessment and evaluation researchers, to address critical, educational issues related to content area learning and literacy for underachieving high school students. The study targeted high schools in California and Arizona serving high numbers of African American and Latino students as well as students from low socioeconomic groups and investigated the extent to which literacy integration strategies may differentially impact various subgroups. The research team focused the study on high school biology and US history because these are the respective courses most high school students are required to take in science and history. The study utilized a true group-randomized experimental design and multiple measures of both teacher implementation and student learning over multiple time points to gauge program impacts. Data sources and multi-level analytical methodologies enabled the research team to trace the linkages between a generative model of professional development in literacy for science and history teachers, to teachers' instructional practices, and to student engagement and learning in literacy, science, and history.

THEORETICAL BACKGROUND AND RELEVANCE TO THE FIELD

Literacy Proficiency as a Gatekeeper to Academic Achievement

Numerous studies have demonstrated that the literacy proficiency of young adolescents shapes their academic futures through systemic inequities that result in tracking students into college-bound and non-college bound courses of study at the high school level (Hull & Rose, 1989; Knapp, 1995; Oakes, 2005; Sizer, 1992). As early as 3rd and 4th grade, relative success or failure reading subject-area texts begins to shape students' reading engagement and academic achievement (Stanovich, 1986) and differences in reading volume translate to differences in knowledge and vocabulary (Cunningham & Stanovich, 1998). As students move up the grades, continued difficulty comprehending academic texts can shape their choices of courses and their engagement in school (Allington, 1991; Davidson & Koppenhaver, 1993; Guthrie & Greaney, 1991; Guthrie, et al., 1996; Guthrie, Schafer, & Hutchinson, 1991), yet student's ability to handle text complexity is the best predictor of success in post-secondary education (ACT, 2006). Furthermore, students' learning outcomes are often measured through standardized achievement tests that require specific subject-area knowledge as well as skillful reading and comprehension abilities. Students' reading proficiency thus becomes a gatekeeper to their further learning in all academic subjects. Improving their capacity to read and comprehend academic texts may contribute in important ways to narrowing the achievement gap in advanced course taking, learning, and achievement.

Over the past several decades, reading has come to be understood as much more than a collection of basic skills. Rather, all texts are shaped by specific conventions and structures of language, and proficient reading of such texts demands the use of these conventions to navigate layers of meaning (e.g. New London Group, 1996; Scribner & Cole, 1981; Shanahan & Shanahan, 2008). Literacy practices become increasingly specialized throughout the school career, reflecting the broader activities that characterize the academic disciplines (Heller & Greenleaf, 2007; Lee & Spratley, 2010). While reading in all academic disciplines requires advanced literacy skills such as critical thinking, hypothesis-testing, effective oral and written communication, and reading across multiple texts and technologies, norms of evidence, logic and discourse vary widely across content areas (CCAAL, 2010).

Moreover, it is now widely recognized that even skillful reading at early grade levels will not automatically translate into higher-level academic literacy (CCAAL , 2010; Greenleaf, et al., 2001; Heller & Greenleaf, 2007; Lee & Spratley, 2010; Snow, 2002). As students move up the grade levels, text complexity increases and the uses and features of texts vary from subject to subject. As Paris (2005) has argued, in contrast with decoding which requires mastery of a small set of constrained skills, acquiring proficiency in reading comprehension requires attainment of a growing set of relatively unconstrained skills that are increasingly situated in particular texts and reading tasks. Paris's distinction calls for developing reading instruction that is commensurate with the authentic literacy tasks characteristic of advanced academic learning. All considered, the implications of these challenges for the literacy and science learning of diverse populations of students are profound.

Increasingly, students in U.S. schools come from a variety of economic, linguistic, cultural, and ethnic backgrounds, bringing significantly different experiences and expectations about how to initiate and sustain conversations, how to interact with teachers and peers, how to identify and solve different types of problems, and how to go about particular reading and writing tasks (Greenleaf, Hull, & Reilly, 1994; Lee, 1995; Moje, Dillon, & O'Brien, 2000).

Literacy researchers have therefore argued that for all students to learn to perform academic literacy tasks, teachers need to make explicit the tacit reasoning processes, strategies, and discourse rules that shape successful readers' and writers' work in particular disciplines (e.g. Delpit, 1995; Fielding & Pearson, 1994; Freedman, Flower, Hull, & Hayes, 1995; Gee, 1999; Lemke, 2006; Moje, 2009; Pearson, Moje, & Greenleaf, 2010; Shanahan & Shanahan, 2008). These researchers have advocated explicitly showing students how to carry out literacy tasks, building bridges from their cultural knowledge and language experiences to the language and literacy practices valued and measured in school in particular academic subject areas.

Approaching Academic Literacy as Discipline-Specific Thinking

Science relies on various kinds of sophisticated literacy skills - the ability to access scientific terminology, to interpret arrays of data, to comprehend scientific texts, to read and write scientific explanations (Lemke, 1996; Osbourne, 2002). In science, texts are artifacts of past investigations and are used for reasoning about scientific phenomena. Scientists use texts to generate new research questions and to provide the background necessary for research design and investigation, and skillful reading of science texts mirrors the kinds of thinking characteristic of science exploration and reasoning (Greenleaf, Brown, & Litman, 2004; Hynd, 1998; Pearson, Moje, & Greenleaf, 2010; Saul, 2004). The language and texts used to represent and communicate ideas in science present novel challenges of comprehension and interpretation to the science learner (Lee & Spratley, 2010; Osborne, 2010; Snow, 2010; van den Broek, 2010).

However, science educators often find the reading of science texts to conflict with necessary opportunities for students to build science knowledge through hands on scientific exploration (e.g. Bybee, 1995; 1997; NSES, 1995; National Research Council, 1996). Increasingly, science educators call for the use of science reading materials other than textbooks to increase engagement and add relevance to students' daily lives, pointing out that language, reading, and writing can play a significant role in understanding science and that the process skills of reading and science are parallel and mutually supportive of learning (Baker, 1991; Manzo & Manzo, 1990; McMahon & McCormack, 1998). Often, science teachers are uncertain how to integrate science reading experiences with science investigations and are keenly aware of students' difficulty comprehending science texts. In our work in high school education over the past decade, we have witnessed a reduction of reading in secondary science classrooms, precisely as policymakers are raising alarms about the reading proficiencies of adolescents (Rycik & Irvin, 2001).

Similarly, the study of history requires the capacity to sift through historical documents with attention to bias and perspective, to construct evidence-based accounts of probable historical events, to place documents and artifacts into larger historical contexts, to evaluate the credibility of different sources of information, and to perceive and have empathy for the experience of others (Bradley Commission on History in Schools, 1995; Levi, 1989). The teaching of explicit cognitive heuristics such as sourcing documents, corroborating evidence and information in documents by relating them to other documents, contextualizing time frames and conditions, and comparative analysis of events and conditions in other parts of the world can be taught to students to foster historical thinking (Wineburg, 2001; Wineburg & Martin, 2004). Yet traditional history education has fallen short in building student capacities in these areas. Like their science colleagues, many history teachers have turned away from the dense texts of the discipline in favor of films and experiential learning activities to engage students in learning history content.

Moreover, because of unprecedented policies and public concerns about adolescent literacy, low-achieving middle and high school students increasingly are losing opportunities to engage in science or historical thinking, as they are being taken out of academic subjects to be enrolled in special literacy intervention programs (Greenleaf, Jimenez, & Roller, 2002) thus missing the important contributions science and history classrooms can make to developing students' literacy proficiency (Calfee & Miller, 2004; Palincsar, 2000; Wineburg, 2001). Students who have been historically underrepresented in higher education are at particular risk and need help acquiring high levels of reading and literacy proficiency as well as content knowledge to participate fully in the academic enterprise.

Integrating Literacy Apprenticeships into Subject-Area Teaching

There is general agreement that direct and explicit comprehension instruction is essential to effective adolescent literacy instruction (CCAAL, 2010). Yet recent research has identified problems with comprehension instruction as it is currently implemented in many content area classrooms. Conley (2008) has questioned the pervasive practice of applying generic strategies designed for young children to the teaching of complex content area goals and materials. In science, where a single text may communicate misconceptions or provide an incomplete picture, getting the gist of a text may be only the first stage in a more sophisticated process of questioning and deliberation. To be proficient readers of science, students need discipline-based reading strategies that permit them to go considerably beyond literal comprehension characteristic of early literacy tasks, to tackle academic tasks such as reconciling multiple texts with different methodologies, perspectives, interpretations and biases (CCAAL, 2010; Conley, 2008; Kamil, et. al, 2008). Other researchers have likewise underscored the urgency of creating conceptually-rich discipline-based skills instruction aligned with science learning goals to challenge students intellectually while helping them build their skills in high level literacy (Paris, 2005; Schoenbach & Greenleaf, 2009; Umphrey, 2009).

Recent critiques also raise concerns that comprehension strategies instruction can potentially displace attention to learning in the subject areas, becoming an end in itself (Conley, 2008; Kamil, et. al, 2008). Contrasting the use of graphic organizers in two science lessons, Conley (2008) argues that, "there is a significant but overlooked difference between using cognitive strategies as a 'teaching tool' versus using cognitive strategies as a learning 'tool,'" and challenges the assumption that teaching a cognitive strategy through activities such as having students contribute details to teacher-generated graphic organizers will help students learn to organize their own thinking. Instead, he suggests that well-integrated cognitive strategy instruction should function "as a deliberate action to develop in students a critical understanding of subject matter ideas *and* a cognitive approach to learning" (p. 91).

Various studies over the past few decades have demonstrated the value of integrating explicit teaching of comprehension, text structures, and word-level strategies into compelling sense-making activities with texts increases student reading achievement (Baumann & Duffy, 1997; Beck, McKeown, Hamilton, & Kucan, 1997; Guthrie, McGough, Bennett, & Rice, 1996; Kamil, et. al, 2008; Pressley, 1998). A recent study of reading and writing about science at the intermediate level indicates that when upper elementary students are explicitly taught strategies for science reading and writing in a learning environment structured to support collaboration and metacognition, students' reading and writing of science content improves (Miller, 2004). The authors of this study argue that literacy instruction is best when embedded in meaningful content instruction (Calfee & Miller, 2004).

Similarly, recent reports of the National Reading Panel (2000), Institute of Education Sciences (Kamil, et. al, 2008) and Carnegie Council on Advancing Adolescent Literacy (2010) have all concluded that teaching a combination of reading comprehension techniques, rather than teaching individual comprehension strategies in isolation from one another and from content instruction, is the most effective method to increase reading comprehension. Furthermore, these reports support the executive role of metacognition in using strategies effectively. For example, CCAAL (2010) recommends, "Once strategies are introduced, students must also learn how to think metacognitively, that is, to determine which strategy is appropriate for a given reading task" (p. 77). While the authors of this report identify English learners as particular beneficiaries of metacognition, the significant vocabulary load of academic texts (Groves, 1995) suggests that the ability to marshal reading strategies to compensate for the comprehension-inhibiting effects of unfamiliar vocabulary and concepts is an essential component of effective literacy instruction in high school subject areas. Student collaboration and instructionally focused conversation have also been identified as key to improving literacy achievement in major adolescent literacy policy initiatives, including the CCAAL (2010) report, the IES practice guide (Kamil, et al., 2008), the report of the National Reading Panel (2000) as well as the RAND (Snow, 2002) report. A recent large-scale study of science instruction suggests that such collaborative meaning-making activities are rare in high school science classrooms (Weiss, et al., 2003).

In order to develop in students the metacognitive approach to learning required for high level literacy, research on effective comprehension instruction thus calls for a kind of flexibly adaptive teaching that is neither commonplace nor simple (Heller & Greenleaf, 2007; Lai & McNoughton, 2009). Some have adopted the metaphor of "cognitive apprenticeship" to describe teaching designed to assist students in acquiring more expert, or proficient, cognitive processes for particular valued tasks, such as reading comprehension, composing, and mathematical problem-solving (e.g. Bayer, 1990; Brown, Collins, & Newman, 1989; Lave & Wenger, 1991). When the target proficiency is a cognitive practice such as composing or comprehending a text, the invisible mental processes involved in the task must be made visible and available to apprentices as they actually engage in meaningful literacy activities (Donovan, Bransford & Pellegrino, 1999; Pearson, 1996; Freedman et al., 1995). To help students develop as readers and writers, teachers can create "literacy apprenticeships," engaging students in meaningful and complex literacy practices while demystifying and supporting students in practicing these literacy practices (Brown et al., 1989; Lee, 1995; Osborne, 2002).

The Reading Apprenticeship Approach to Integrating Science and Reading Instruction

Based on this research in literacy learning, to support teachers' learning and adolescents' discipline-specific literacy development, Greenleaf and Schoenbach and their colleagues have developed, implemented, and studied the impact of an instructional model for academic reading instruction – the Reading Apprenticeship instructional framework (Schoenbach, et al., 1999; Greenleaf, et al., 2001; Schoenbach & Greenleaf, 2009). While the framework incorporates research-based instructional approaches that have been shown to improve adolescent literacy levels— including vocabulary and academic language development techniques, direct and explicit comprehension strategy instruction, extended discussion of text meaning and interpretation, and strategies to increase student motivation and engagement in literacy learning (Kamil, et al., 2008)—in this model reading instruction is also closely aligned with subject-area learning goals, in this case, science, and integrated into content-area teaching, rather than being an instructional add-on or additional curriculum. Further, unlike approaches that respond to the challenges of reading by simplifying texts or tasks (Greenleaf & Hinchman, 2009; Pearson, Moje, & Greenleaf, 2010), the Reading Apprenticeship framework helps teachers improve their instructional routines around existing academic curriculum and texts and to extend students' reading opportunities with these and ancillary science materials.

The Reading Apprenticeship model is based on research indicating that most students are capable of complex thinking and carrying out scientific and literary inquiry but have not been given the skills or self-confidence to approach these tasks effectively (Greenleaf, et al., 2001; Langer, 2001; Lee & Spratley, 2010; Moje, et al., 2008). In this framework, students are given extended opportunities to read a wide range of texts with instructional support—both textbook and lab materials or primary source documents, and ancillary materials such as journal articles and trade books. Through an "apprenticeship" process, content-area teachers explicitly model and guide students in practicing the tacit reasoning processes, strategies, and discourse rules that shape successful readers' and writers' work. The focus of instruction is therefore on engaging with academic texts to actively construct meaning and flexibly utilizing an array of comprehension tools in order to do so, rather than on learning to carry out a particular comprehension strategy or set of strategies as learning targets in their own right (Conley, 2008; Kamil, et. al, 2008).

The Reading Apprenticeship framework centers on metacognitive conversation, involving explicit metacognitive routines, modeling, small group work, and class discussions that focus on how to read science and history materials and why people read these materials in the ways they do, as well as the content of what is read in academic classes. These discourse routines offer students support to: clarify content, discuss the processes they use in reading and problem-solving, practice comprehension strategies, respond to and elaborate on content, engage in word learning strategies, write to learn and to consolidate learning, and make connections to other related texts and topics. Reading Apprenticeship practices are designed to draw both on what teachers know and do as readers in particular academic domains, and on adolescents' underestimated strengths as learners.

Many of the underperforming students in U.S. high schools have resigned themselves to low literacy and academic attainment (Lee & Spratley, 2010). In this framework, teachers attend to students' affective and identity issues, creating relevant and affectively safe learning opportunities that help students build stamina and dispositions to engage in academic tasks, discipline-based literacy practices, and inquiry, and to develop identities as resilient learners (Benard, 1996; Schoenbach & Greenleaf, 2009). The framework thus aims to support teachers in building students' capacities to carry out close, intellectually engaged reading; gain insight into their own thinking processes; make meaning; acquire academic and disciplinary language; read independently; and set personal goals for literacy development.

Previous studies of the impact of Reading Apprenticeship have demonstrated increased reading achievement and academic engagement across a diverse group of adolescents enrolled in a Reading Apprenticeship Academic Literacy course in ninth grade (Greenleaf, et al., 2001; Schoenbach, Greenleaf, Cziko, & Hurwitz, 1999). These results have been replicated in additional studies, including a randomized, controlled study examining whether supplemental literacy classes improve the reading skills of struggling ninth-grade readers (Corrin, et al., 2008), demonstrating that teachers' implementation of Reading Apprenticeship can result in significant gains for students across varied grade levels and subject areas for a review of these studies, see www.wested.org/ra). Further, explicit support for reading in a chemistry class has been shown to build low-performing students' abilities and dispositions to work through conceptually dense science materials and, ultimately, to participate in science learning in new ways (Greenleaf, Brown, & Litman, 2004; Litman & Greenleaf, 2008). This prior research suggests that implementation of the Reading Apprenticeship instructional framework has the potential to increase students' reading and biology and history engagement and achievement at the high school level.

The Knowledge Base on Effective Methods of Professional Development

A long history of research in reading has demonstrated that reading comprehension strategies are not often taught in subject-area classes, even when teachers are trained to use these strategies during subject-area teaching (Alvermann & Moore, 1991; Duffy et al., 1986; Duke, 2000; Durkin, 1984; Fielding & Pearson, 1994; Richardson, 1994; Snow, 2002). Furthermore, even when teachers do implement literacy strategies, they often have difficulty balancing content and strategy instruction at the same time because a culture of whole class direct instruction makes it difficult for teachers to engage students actively in reading and learning (Reed, 2009). To increase the potential for reading instruction to become tightly integrated into content area teaching, professional development must therefore demonstrate features of high quality learning for teachers that are known to be effective in producing changes in classroom instruction (Kamil, et. al, 2008; Strickland & Kamil, 2004; Wei, et al., 2009).

There is much known and much yet to be known about the elements of effective professional development. In 1999, the National Research Council report on the science of learning identified important themes in teacher learning (Bransford, Brown & Cocking, 1999). Other studies examining the impact of teacher professional development on teacher knowledge and practice in both science and literacy have supported and elaborated these findings (Garet, et. al, 2001; Kennedy, 1998; Loucks-Horsley & Matsumoto, 1999; Reed, 2009; Yoon, et. al, 2007). Desimone (2009) argued recently that there is now sufficient empirical evidence to establish consensus on a set of five core features of effective professional development: a content focus on student learning in particular subjects, active learning opportunities, coherence with teachers' existing practices and policy contexts, sufficient duration, and collective participation. Similar elements appear in various guises in a wide range of research on teacher professional development (Ball & Cohen, 1999; Darling-Hammond, et al., 2009; Little, 2001; Garet, et. al, 2001; Guskey & Huberman, 1996; Kennedy, 1998; Shulman, 1987). Yet surveys of professional development surface large discrepancies between what is known to be effective and teachers' professional development experiences (Fishman, et al., 2003; Loucks-Horsley & Matsumoto, 1999; Richardson, 2003).

While there is a growing consensus around what constitutes effective professional development, the ambitious research agenda inspired by the National Research Council report on the science of learning made only passing mention of the need for studies linking teacher professional development, teacher learning and student achievement (Donovan, et. al, 1999). As Desimone (2009) has outlined, early research on teacher professional development established effectiveness through teacher self-report on such outcomes as teachers' satisfaction with the staff development experience, with later studies measuring changes in participants' beliefs and attitudes, content knowledge, or commitment to change. Some research has also measured changes in participants' practice. Relatively few studies of professional development programs have examined the impact of teacher learning on student performance (Desimone, 2009; Killion, 1998). Of 450 professional development projects identified and reviewed by the Middle Grades Initiative of the National Staff Development Council, fewer than 10% included any measurement of student achievement (Killion, 1998). Likewise, a review of professional development in math/science identified only four science programs that collected data on student achievement (Kennedy 1998). Thus, while research has confirmed that effective professional development can increase teacher confidence in instructing students with diverse abilities (Garet, et. al, 2001), few studies can link professional development to student achievement (Garet, et. al, 2001; Kennedy, 1998; Loucks-Horsley & Matsumoto, 1999; Reed, 2009). A decade later there is still limited evidence to confirm that professional development can increase students' learning in the content areas (Desimone, 2009; Reed, 2009; Yoon, et. al, 2007).

Compounding the lack of attention to student outcome data in many studies of professional development is the methodological challenge of demonstrating an impact of teacher professional development on student achievement (Borko, 2004; Loucks-Horsley & Matsumoto, 1999; Reed, 2009; Supovitz, 2001; Yoon, et al., 2007). A recent review of the effects of teacher professional development on middle school teachers' subsequent implementation of literacy strategies in content area classes yielded only one study that included student outcome data, and that study did not include statistical information suitable for confirming the calculation of effect sizes (Reed, 2009). The most comprehensive review of the impact of teacher professional development on student achievement in science, math, reading and English/language arts identified 9 studies out of more than 1,300 potential studies that met criterion standards of evidence (Guskey & Yoon, 2009). No studies at the middle or high school levels met these standards, nor did any study published after 2004. The authors of the review acknowledged that they were stunned by these findings: "Obviously, these findings paint a dismal picture of our knowledge about the relationship between professional development and improvements in student learning" (p. 497).

Design Elements of Reading Apprenticeship Professional Development

Based on these understandings of the important features of professional development and with reference to Desimone's (2009) proposed conceptual framework for professional development impact studies, below we describe the elements of the Reading Apprenticeship professional development model that constituted the intervention for this study before turning to the study methods and findings. Based on the Reading Apprenticeship framework and commensurate with the contextualized nature of academic literacy, the professional development model utilized in this study is designed to transform teachers' understanding of their role in adolescent literacy development and build enduring capacity for literacy instruction in the academic disciplines (Greenleaf & Schoenbach, 2004). We outline the specific features of teacher learning opportunities in this model using Desimone's proposed conceptual framework.

Content focus and active learning.

Reading Apprenticeship professional development is inquiry-based, subject-area focused, collaborative, and designed to address teachers' conceptual understandings as well as practical implementation needs. The professional development model emphasizes the development of pedagogical content knowledge (Shulman, 1986; 1987). From the outset of the professional development, teachers are immersed in learning experiences integrating science and literacy by engaging in rich investigations into science reading and scientific or historical investigation. The instructional framework positions literacy as inquiry, and professional development activities aim to draw on the similarities between discipline-based inquiry processes and literacy. Professional development routines engage teachers in learning experiences integrating science or history content and literacy by engaging in rich investigations into disciplinary reading and inquiry.

Because the content focus of this professional development model centers on learning and literacy thinking processes, it is tightly linked to the active mode of learning designed for teachers and for their students. To increase teachers' capacities to generatively design and implement the kind of instruction that supports student literacy, the professional development immerses teachers in models of practice that we aim for them to create in their own classrooms: inquiry based, collaborative classroom instruction that engages students actively in metacognitive conversations about reading and learning processes. To this end, in professional development sessions teachers participate in carefully designed inquiries to help them unlock their own disciplinary expertise in relation to literacy. They work to identify the features of disciplinary texts that might present stumbling blocks to learners. Most importantly, they collaboratively investigate student work, videotaped classroom lessons, and case studies of student literacy learning designed to foster new expectations of what their own students can accomplish. In professional development sessions, they practice classroom routines to build student engagement, support student collaboration, and foster authentic discussion and problem solving around course texts, all with the goal of learning new ways to support students' thinking and learning with academic materials.

The inquiry processes for teacher development are the result of an ongoing research and development process involving varied communities of teachers and SLI researchers over the past decade to iteratively refine effective inquiry designs for professional development (Greenleaf & Schoenbach, 2001; 2004). The resulting collaborative inquiries are designed to engage teachers in exploring a) metacognitive processes involved in reading complex texts, b) videotapes of metacognitive process interviews with students reading complex texts, c) evidence of student thinking in samples of student work, d) videotapes of classroom lessons in which teachers attempt to integrate Reading Apprenticeship instructional approaches into context learning; e) the varied types of texts used to represent ideas in the discipline, f) the knowledge and language demands of disciplinary texts, and g) the benefits and potential pitfalls of using specific reading strategies with subject area texts. Elsewhere we have described these tools and approaches in more detail and explicated their specific aims for developing teachers' capacity for responsively adaptive teaching; in brief: to build teachers' conceptions of reading in their disciplines; to build teachers' insights into student learning needs and capacities; and to build teachers' situated and conditional use of reading comprehension strategy instruction (Greenleaf & Schoenbach, 2004).

These inquiries also develop teacher capacity to participate in and facilitate the collaborative metacognitive conversations at the center of the Reading Apprenticeship instructional framework. Metacognitive conversations are framed in social routines that support talk about thinking and reading such as think aloud, think/write-pair-share, and reciprocal small group discussions about written notes or annotations centered on reading processes with disciplinary texts. As teachers read, surface and discuss their problem solving responses to challenges they find in texts and share their reading processes, the distributed knowledge in the room about how to strategically approach reading becomes shared knowledge through visual notetaking procedures. Professional development facilitators help teachers to label their reading processes, and thereby develop declarative knowledge about – that is, a language for describing – reading and thinking processes. Guided practice of reading strategies and discussion of how these strategies support reading comprehension aims to build procedural and conditional knowledge about how and when to monitor comprehension and resolve confusions with academic texts. Routines for metacognitive conversation serve as a model for classroom instruction that is designed to support teachers to sustain an ongoing inquiry into reading processes in their classrooms. These inquiry approaches are meant to support teachers' ongoing learning about reading processes as well as to help their students build and use new strategic approaches to comprehension challenges, all while reading (and learning) academic content.

Duration.

Reading Apprenticeship professional development is designed to support teacher learning over a two-year sequence of day-long sessions, with summer institutes followed by follow up sessions within the school year. Day-long sessions provide teachers with a sufficient stretch of time in which to engage in active inquiry as learners, analyze the pedagogical structure of these opportunities, reflect on the impact of these pedagogies on their learning and that of their students, and plan for instruction. Between professional development sessions, teachers are asked to use new pedagogical tools in their classes and practice new ways of responding to students. They are asked to bring student work resulting from trying out these new instructional approaches to subsequent sessions and in collaboration with their colleagues engage in examination of student work, student thinking processes, and student instructional needs. The aim of the ongoing sessions is to help teachers "diagnose problems in their classrooms and schools, apply evidence-based and often alternative solutions to them and evaluate and analyze the impact of implemented procedures" (Darling-Hammond, et al., p. 29). Through this cycle of inquiry, the aim is for teachers to receive continuing help in surfacing new issues and problems, and ways of understanding and translating them into new practices (Loucks-Horsley & Matsumoto, 1999).

Collective participation.

Instructionally focused conversation is at the heart of the Reading Apprenticeship instructional framework and Reading Apprenticeship professional development is likewise organized to promote discourse about literacy, content, and problems of practice. The texts and reading tasks used in professional development sessions are designed to raise authentic problems for teachers. As they discuss how to resolve the reading challenges they experience, teachers are immersed in a model of inquiry-based strategy instruction—a metacognitive conversation about how to identify and resolve comprehension problems. Following this immersion, participants debrief the experiences, making the pedagogy embedded in the activity apparent in order to build knowledge of how to support metacognitive conversation about reading processes in the classroom.

Teachers are asked to describe the pedagogies that supported their own reading and learning, their thinking about reading, and collaborative talk about thinking and reading. Teachers are invited to reflect on how metacognitive conversation about reading supported their own learning and to extend the conversation to classroom implications of such learning opportunities and needed adaptations for instruction with students. By understanding the impact of the designed inquiries and their implications for instruction, teachers build an understanding of the inquiry processes and their own purposes for learning how to implement these instructional routines. In this way, teachers work to solve problems of practice in collaboration with their colleagues as they take ownership of these problems and

Typically, teachers in Reading Apprenticeship professional development participate in school-based teams organized into cross-site learning networks to provide opportunities for school-wide professional conversation and collective action on literacy in content areas. In the current study, however, constraints of the study design made this organization impracticable, as will be detailed in the study design, below.

Coherence.

By design, Reading Apprenticeship professional development activities confront many deeply held beliefs and commonly accepted practices in traditional secondary education, among them simplistic views of reading, misperceptions about the substantial capabilities of their diverse students, and little appreciation of the roles reading and texts play in content area learning. In order to help teachers build a new set of conceptual understandings and analytical thinking tools for teaching literacy in the discipline and responding productively to students' literacy needs as they read and learn in the subject area classroom, professional development activities engage teachers in examining video and written case studies of student reading designed to encourage evidentiary thinking and to surface and challenge teachers' current beliefs and practices. As teachers share and discuss their observations and interpretations of a student's performance, they begin to consider many differing interpretations of student readings of text. Because the cases are constructed to engage common misconceptions and provoke authentic questions about student reading and thinking, during a case discussion individual teachers will give voice to conflicting perceptions. Case discussions are carefully facilitated to help teachers see the value of classroom conversations about reading as formative assessment data and practice making evidence-based claims about students' reading and learning strengths and needs rather than basing instruction on preconceptions and assumptions about student abilities. Interpreting student thinking based on observations during reading activity helps teachers develop insights into student thinking during the case, and later develop new insights into their own students as they listen to their students discuss science readings in their classrooms or examine their reading-related work.

In addition to these analytical tools, the professional development provides teachers with an experiential knowledge base of literacy routines, strategies and tools that support teachers as they develop adaptive expertise that allows them to solve problems flexibly. This instructional tool kit, while situated in ongoing inquiry activities and explorations, is designed to meet teachers' needs for concrete and practical solutions to the everyday problem of students' limited comprehension of course texts. At the same time, teachers are invited to reflect on and critique instructional techniques and to adapt them to content area teaching and their own students, grounded in principles of instruction from the Reading Apprenticeship instructional framework. Finally, to foster connections to reform initiatives and expectations operating in their school sites, teachers are given opportunity and support to explore the fit between the literacy approaches they already utilize, those they are learning in the professional development sessions, and the curriculum standards for which they are accountable.

Potential Contributions of this Study

Previous studies employing the Reading Apprenticeship professional development model have shown it to be effective in changing teachers' knowledge and classroom practice increasing students' literacy achievement (Greenleaf & Schoenbach, 2001; 2004). These studies of the model, using a mix of qualitative and quasi-experimental methodologies, have suggested that participating teachers change their beliefs about the role of reading in content area instruction; enlarge their conceptions of literacy (enriching what are often impoverished views of the complexities involved in reading and comprehending texts); expand their repertoire of pedagogical practices to support reading development; implement new instructional strategies; increasingly view subject-area reading tasks from the point of view of learners; and listen to students with new insights into their process of learning (Greenleaf & Katz, 2004; Greenleaf & Schoenbach, 2001; 2004). Most recently, a randomized controlled study demonstrated evidence that Reading Apprenticeship professional development resulted in increased literacy integration and student achievement in high school biology (Greenleaf, et al., 2010). The current study extends previous research by explicitly examining links between professional development embedding core features of effective teacher development, teacher development of adaptive expertise needed to meet the high standards envisioned in such documents as the Common Core standards, and student achievement in reading and science and history. Furthermore, the study tests these links utilizing rigorous research methodologies, in schools serving high needs students, using distal measures that acknowledge the "high stakes" at play in the current educational research and reform environment.

Professional development to implement Reading Apprenticeship in US history or in biology served as the intervention for this study. Based on previous studies, we hypothesized that through this professional development experience, teachers would develop new knowledge and resources about text, reading and student thinking that would support their teaching of reading in biology or US history. With practice, teachers would learn to deploy these resources more flexibly, on demand, as students need them. In turn, students would practice these thinking tools as authentic and relevant responses to real reading situations, to make sense of course texts as they build knowledge of the topic, rather than as a set of fixed exercises in isolation from sense making or knowledge building. Through literacy practice situated in authentic learning, we expected both teachers and students to stretch beyond their current ability and gain more expertise and capacity.

Research Hypotheses

We designed a randomized-controlled study to test these hypotheses, examining impacts of the professional development model on teachers' instructional practices as well as on student literacy and science learning.

> H1: Teachers participating in the Reading Apprenticeship professional development program will exhibit greater increases in knowledge and skills regarding the integration

of literacy and biology or U.S. history, and will demonstrate greater integration of literacy into their instructional practice than teachers in control classrooms.

H2: Students in experimental classrooms will demonstrate greater increases in science or history understanding, reading proficiency, and engagement in subject-area learning than their counterparts in control classrooms.

STUDY METHODS

A multi-role study team was assembled to include developers of the intervention and their research and professional development staff, science content experts, and independent evaluators. Roles were carefully delineated such that developers and their staff were involved in designing the professional development and instrumentation for the study, while independent evaluators were responsible for instrument scoring and analysis.[1] Data firewalls prevented developers from having access to data files.

Experimental Design

To test these hypotheses, a true, group-randomized, experimental design (Cook & Campbell 1979, Murray 1998) was conducted to control for most threats to internal validity in order to assess the impact of Reading Apprenticeship professional development on high school biology and US history teaching and learning. A pretest/posttest control group design – based on 2 cohorts of teachers – was used to assess program impacts on teacher and student outcomes. The design was structured to provide professional development and support to U.S. History teachers and Biology teachers serving different cohorts of high school students. As shown in the top panel of **Figure 1**, Group #1 represents the schools with U.S. History teachers trained by WestEd, while Group #2 represents schools in the control group for the U.S. History Reading Apprenticeship intervention. Group #2 serves as the Biology teacher treatment group – as Biology teachers in Group #2 will receive professional development in Year 2. Group #1 represents schools in the Biology teacher control group. If participating Biology and U.S. History teachers were in the same school, schools in Group #1 and Group #2 both received professional development services, albeit in different departments and years. This helped ensure equal participation and buy-in from schools in both groups (Shadish, Cook & Campbell, 2002). Control group teachers were exposed to regular teacher professional development opportunities that occur in the school sample. Both treatment and control group teachers existed in the same

[1] Because one of the PI's for this study is the developer of the Reading Apprenticeship framework and has carried out a program of research and development focused on effective professional development for Reading Apprenticeship, the research team carefully delineated roles to avoid the possibility or the perception of bias in the study of this intervention. The primary role of WestEd's Math and Science program and the Strategic Literacy Initiative project in the Teacher Quality program of WestEd was to provide content expertise in science, literacy, and professional development to inform the intervention and instrumentation of the study. Strategic Literacy Initiative staff were further divided into two formally isolated teams with research vs. professional development responsibilities. The primary role of UCLA's CRESST Center was to develop, field test, and analyze measures of the nature and degree of literacy instruction in biology and US history classrooms, including teacher surveys and teacher assignments. All scoring of these measures was carried out by CRESST and non-participant teachers whom CRESST staff trained to score. As content experts, the Strategic Literacy Initiative staff were involved in the scoring of teacher interviews, alone. The research staff, with Research and Evaluation experts at the Los Alamitos office of WestEd, kept locked data files and codes identifying teachers and students which only they had access to.

schools in 40 of the 120 participating schools that were randomly assigned to experimental condition. It is conceivable that control group teachers were contaminated through interaction with treatment teachers outside of class. Based on our experiences working in high schools, we anticipated a low likelihood of this occurring – as there typically is very little interaction between teachers across different departments. To further guard against contamination, treatment group teachers were informed during face-to-face meetings of the nature of the study, professional obligations with regards the scientific investigation, and the necessity of not sharing the material with other teachers during the study period.

Difficulties meeting recruitment targets necessitated that a second cohort of US History teachers be recruited in the spring of Year 1. To be eligible to participate, Cohort 2 US History teacher participants could not be in schools with participating Biology teachers to avoid student exposure to both Biology and US History teacher participants (see **Figure 1** below). Both cohorts of US History teachers were pooled in the data analyses.

The bottom panel of **Figure 1** depicts the design with respect to *students*. A comparison of the teacher and student panels in **Figure 1** indicates how teachers' professional development is aligned with the high school trajectories of student participants. To maximize the treatment contrast, we required that teachers receive one full-year of professional development and coaching *prior to* instructing student participants. For the US History intervention, we collected baseline 10th grade test score data from school districts and analyzed student outcome data collected at the end of Year 2 – when program impacts on annual growth in reading comprehension between 10th and 11th grade were assessed, as well as program impacts on History achievement. Analogously, for the Biology study, we collected baseline 8th/9th grade test score data and analyzed student achievement data collected at the end of Year 3 to estimate program impacts. Note that the professional development was staggered such that 11th grade history teachers were trained before 9th/10th grade teachers so that students would not be impacted if contamination occurred. Mixed-modeling procedures (see Statistical Analysis Section below) were used to detect treatment effects on teacher- and student outcomes.

Recruitment and Random Assignment of Schools and Teachers

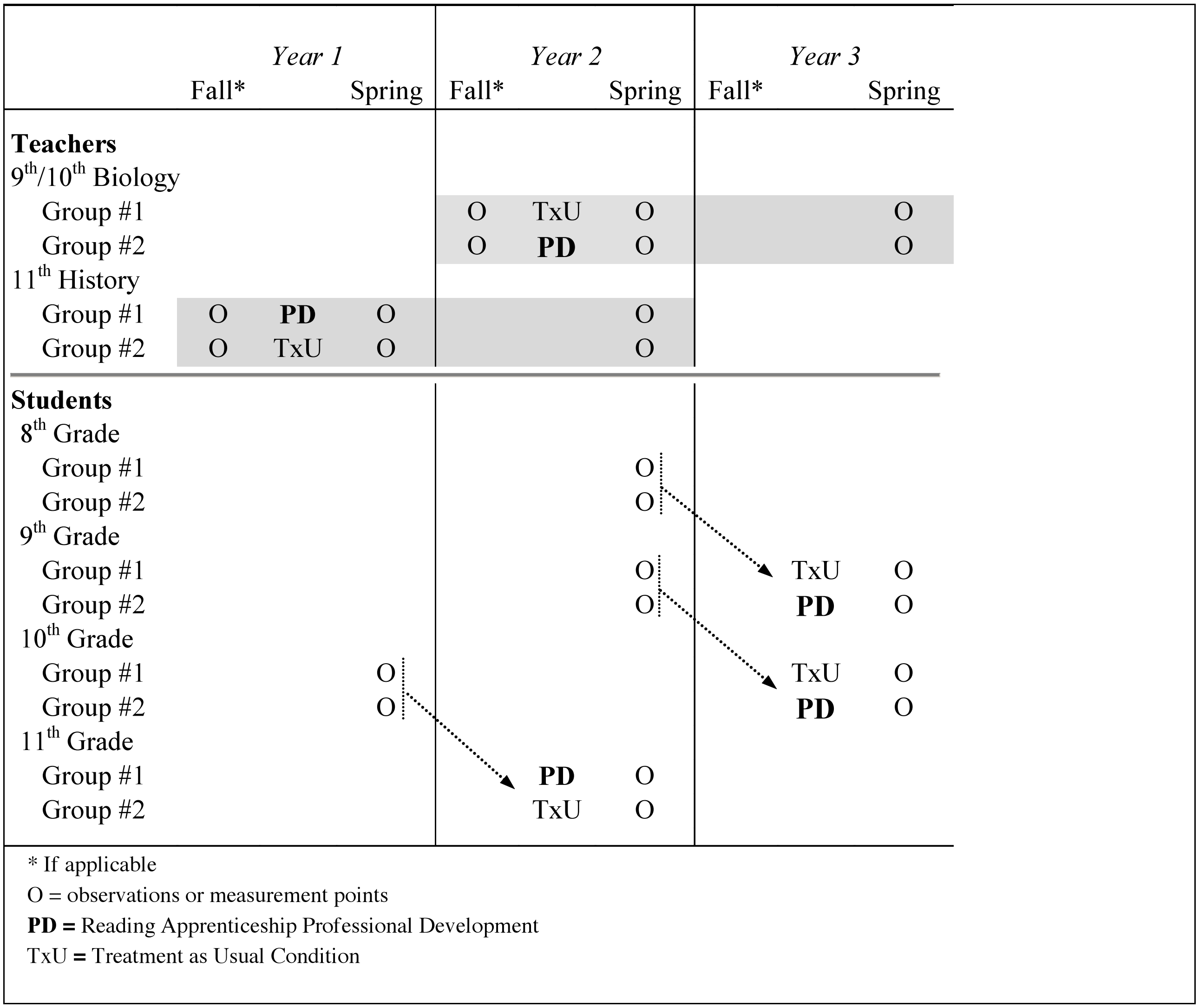
The target population was high school US history teachers or biology teachers and their students in public high schools across California and Arizona. The study took place in high schools that serve populations of students historically underrepresented in post secondary education settings. The sample consisted of schools with high proportions of these students to better ascertain the impact of integration of literacy instruction with US history or biology course-work. Schools, not teachers, served as the unit of randomization to minimize contamination of the control group through teacher interaction. School recruitment efforts involved outreach to high school teachers, principals, and school districts and concluded with teachers and school districts signing a memorandum of understanding during the spring prior to the first program implementation year. A total of 219 teachers (108 history and 111 biology teachers) from 120 schools agreed to participate in the study and were randomized to experimental condition. Difficulties in meeting recruitment targets necessitated that a second cohort of US History teachers be recruited.

For the US history sample, 108 US history teachers in 82 schools were initially recruited, with 59 teachers (45 schools) assigned to the treatment group and 49 teachers (37 schools) assigned to the control group. Note that teachers and schools were recruited and randomized to condition in the spring of prior to the first implementation year,[2] two to three months prior to the scheduled summer professional development institute. Schools and teachers were randomly assigned in batches so that adequate notice could be given to teachers to schedule participation in the summer professional development. For the biology sample, 111 biology teachers in 78 schools were initially recruited, with 56 teachers (39 schools) assigned to the treatment group and 55 teachers (39 schools) assigned to the control group. While biology teachers and schools were recruited and randomized to condition in the spring of 2006, biology teacher professional development was delayed for a year by design.

Schools, not teachers, served as the unit of randomization to minimize contamination of the control group through teacher interaction. Prior to randomization, participating high schools were pair-matched with similar schools based on academic performance and demographic factors. A two-stage strategy was used for matching. First, similar schools were matched based on 3 factors: (1) academic performance, African-American enrollment, and (3) Latino enrollment. Participating schools were located in multidimensional space defined by these factors, and matched with one other school. Schools and the participating teachers within them were randomly assigned to intervention and control conditions within each pair.

[2] Teachers in cohort 1 and cohort 2 were randomized in spring 2006 and spring 2007, respectively.

Figure 1. *Reading Apprenticeship Professional Development Experimental Design for Teachers and Students*



	Year 1			*Year 2*			*Year 3*		
	Fall*		Spring	Fall*		Spring	Fall*		Spring
Teachers									
9th/10th Biology									
Group #1				O	TxU	O			O
Group #2				O	**PD**	O			O
11th History									
Group #1	O	**PD**	O			O			
Group #2	O	TxU	O			O			
Students									
8th Grade									
Group #1						O			
Group #2						O			
9th Grade									
Group #1						O		TxU	O
Group #2						O		**PD**	O
10th Grade									
Group #1			O					TxU	O
Group #2			O					**PD**	O
11th Grade									
Group #1					**PD**	O			
Group #2					TxU	O			

* If applicable

O = observations or measurement points

PD = Reading Apprenticeship Professional Development

TxU **=** Treatment as Usual Condition

Professional Development Intervention

Experimental condition.

Teachers in schools randomly selected to be in the experimental condition for each subject area received a total of 10 days of professional development in Reading Apprenticeship and support to integrate science content and reading instruction. The 10 days of professional development were led by certified Reading Apprenticeship professional development providers and utilized Reading Apprenticeship inquiry tools and approaches to professional development (Greenleaf & Schoenbach, 2004). In addition to these inquiry tools, the institutes were designed to integrate literacy and topics in the high school biology or US history curriculum, engaging teachers in experiential learning with the inquiry approaches and pedagogies they were being asked to implement in their content area teaching, but at a level of complexity suitable for adult and experienced science or history learners (Loucks-Horsley & Matsumoto, 1999; Loucks-Horsley, et al., 2003).

As described above, the professional development curriculum involved teachers experientially in using metacognitive routines such as think-aloud (Kucan & Beck, 1997) and metacognitive logs for reading and for science investigations (Schoenbach, et al., 1999); teacher modeling of reading and reasoning processes with think-aloud and text annotation (Greenleaf, 2006); methods of orchestrating and conducting collaborative small group work such as think-pair-share, jig-saws, and other group protocols involving comprehension routines such as ReQuest (Manzo, 1969) and Reciprocal Teaching (Palinscar & Brown, 1984); and engaging in extended reading opportunities with varied sets of texts on particular topics. These instructional approaches were tightly integrated with core units of study in biology or US history to illustrate science or history and literacy integration. Metacognitive reflection on the impact of these learning opportunities and specific pedagogies, for their own and their students' learning, followed the integrated lessons. The 10 days of professional development were spread over a year, beginning in summer of one year, with follow up mid-year, and a final session before the start of the second year, during which data on student opportunity to learn and achievement was collected. Appendix A gives an overview of the 10 days of professional development for each subject area.

Professional development for history and biology was staggered, beginning history professional development in year 1 and beginning biology professional development in year 2, to allow teachers from both subject areas from the same school to participate without creating dosage effects for students in the schools, as explained above. In addition, to reach recruitment targets for history, it was necessary to conduct professional development for two cohorts of history teachers in years 1 and 2 of the study. In the summer of 2006, the first cohort of US history teachers participated in five days of training. Implementation of reading instruction in their history classes began in the fall of 2006. The professional development coaches made use of informal interviews and/or email interactions with these teachers to plan two follow-up days of training given during Year 1 (2006 - 2007 school year), targeting the teachers' emerging needs for support. A final three-day professional development follow-up occurred in the summer of 2007, prior to the data collection year for this cohort. The data collection year for this first cohort of US history teachers occurred during the 2007 – 2008 school year.

One year after the history cohort began professional development for biology teachers began and followed this same structure and schedule. Simultaneously, a second cohort of history teachers went through the professional development. Thus, biology teachers and History Cohort II teachers assigned to the intervention received an initial 5 days of professional development in the summer of 2007, followed by two days of follow up training in the winter of 2008, and a final two days of training in the summer of that same year. The data collection year for this group of teachers occurred during the 2008 – 2009 school year.

Throughout the study, exchanges among teachers assigned to the intervention took place through a list serve, moderated by professional development coaches. See **Appendix A** for an overview of activities carried out in the 10 days of professional development in each subject area.

To support implementation and to assure equal access across experimental sites to opportunities to read in the subject areas, these teachers were provided funds and a list of reading materials to supplement their locally-adopted textbooks. These materials constituted a classroom library of science or history magazines, trade books, fiction, and non-fiction selections linked to curriculum topics and state curriculum frameworks. Stipends covered teacher participation in the professional development, including travel, food, and housing for the summer institute, honoraria for teacher's time, additional stipends or substitutes for day-long sessions during the school year, and up to $500 for instructional reading materials.

Control condition.

Teachers randomly selected to be in the control condition implemented their usual teaching practices. Thus, the control group represents a treatment-as-usual condition, representing what students would normally receive at schools participating in the study. However, teachers in the control condition were also offered the library of supplemental reading materials given to intervention group teachers so that the difference between groups, if any, was not attributable to whether or not such materials were present in classrooms. Control group teachers were also compensated for their participation in data collection activities. Participation in other professional development activities, changes in teaching practices, acquisition of knowledge and skills, and other changes in conditions and circumstances were tracked and monitored in control groups through yearly surveys.

Data Collection and Analysis Procedures

Study methods included multiple measures of both teachers' instructional practices that shape students' opportunity to learn in the classroom and student achievement, and were designed to enable us to determine the extent to which these instructional methods might have different impacts for groups of students historically underrepresented in higher education. Study measures included a set of pre- and post-intervention survey assessments of teacher knowledge, beliefs, and instructional practices in science or history and literacy; post-intervention interviews; samples of lesson assignments with accompanying student work in particular biology or US history topics; student surveys; and pre- and post-intervention assessments of student learning in biology or US history and reading comprehension. Participating schools were spread across many districts in the large geographically varied state of California as well as Arizona. Sending researchers to these many sites to carry out multiple, direct observations of all classrooms would have been prohibitively expensive in a study of this size. Therefore, we designed data sources and data collection procedures to enable us to corroborate data from various sources, using robust proxy measures of implementation, and carried out a small number of classroom observations in intervention and control classrooms for use in describing classroom implementation.

For example, rather than rely simply on teacher self reports in surveys and interviews, we also used a method of collecting and analyzing lesson assignments and student work samples that has been shown to serve as a good proxy for classroom observations (Clare & Aschbacher, 2001; Matsumura, 2003). Student surveys provided a check on teacher self report regarding classroom practices, as well as a measure of student engagement, self-efficacy in science or history reading, and motivation. Instruments were designed to measure similar constructs to facilitate corroboration across measures. We describe these instruments in more detail below.

Measures of teacher knowledge, belief, and instructional practice.

Teacher survey.

Based on the theoretical constructs underlying the Reading Apprenticeship instructional framework and accompanying professional development model, parallel forms of a teacher survey were designed to assess six global constructs related to effective integration of literacy and biology or US history instruction: (1) science or history reading opportunities, (2) collaboration, (3) metacognitive inquiry, (4) comprehension strategies instruction, (5) a feature of instruction called "negotiating success"—a focus on designing and modifying instruction and assessment supporting response to student learning needs, and (6) teacher beliefs about reading, learning and diversity (**Table 1**). The six constructs were further divided into 14 sub-constructs reflecting aspects of the apprenticeship model -- the range of science or history reading opportunities offered to students, and the nature and degree of teacher modeling, guidance and support for, as well as student opportunities for practice with, key reading and discourse routines, tools, strategies and dispositions.

Related to each of these constructs, we developed a set of items describing instructional practices or in the case of construct 6, a set of value statements. Teachers responded on a 5-point likert scale regarding the degree of emphasis they placed on the item or its frequency of use, or the degree to which they agreed with the item. To pilot the survey, we administered it to a set of teachers not participating in the study and conducted a factor analysis. With few exceptions, items loaded on the constructs they were expected to, and non-loading items were omitted from the survey. The resulting survey constructs had reasonably good psychometric properties, as alpha levels show in **Table 1** for US history and biology below.

Table 1. Internal consistency reliability coefficients for teacher survey measures

	History				Biology			
		Cronbach's alpha reliability				*Cronbach's alpha reliability*		
	Items	Baseline	Year 1	Year 2	Items	Baseline	Year 1	Year 2
(1) Student Reading Opportunities								
Texts	15	0.66	0.77	0.80	13	0.80	0.76	0.71
Learning Structures	4-6	0.69	0.57	0.66	6	0.60	0.54	0.52
Content	4-8	0.54	0.84	0.84	3	0.63	0.63	0.22
(2) Collaboration								
Teacher Modeling	6	0.83	0.85	0.88	6	0.81	0.76	0.75
Student Practice	5	0.84	0.84	0.86	6	0.85	0.82	0.85
(3) Metacognitive Inquiry								
Teacher Modeling	5	0.77	0.75	0.69	5	0.79	0.73	0.72
Student Practice	7	0.80	0.85	0.83	7	0.89	0.84	0.90
(4) Comprehension Strategies								
Teacher Modeling	5	0.81	0.78	0.68	2	0.92	0.94	0.91
Student Practice	15	0.86	0.90	0.88	20	0.94	0.93	0.92
(5) Negotiating Success								
Instruction	9	0.86	0.88	0.86	10	0.84	0.84	0.83
Assessment	6	0.76	0.75	0.77	4	0.71	0.77	0.82
(6) Teaching Philosophy								
Reading	13	0.57	0.59	0.60	11	0.64	0.52	0.65
Learning	14	0.71	0.67	0.75	13	0.65	0.66	0.72
Diversity	9	0.75	0.68	0.71	11	0.49	0.46	0.58
(7) Disciplinary Thinking	7	0.71	0.78	0.74	--	--	--	--

The teacher survey was administered to teachers before the professional development intervention, a year later at the end of their first year of implementation or treatment as usual, and again at the end of the next year. The final survey serves as a post-test to the identical surveys taken in previous summers. Pre-survey responses were analyzed to determine whether there were initial differences between mean responses of treatment and control group teachers on each construct and sub-construct. To examine pre- and post-intervention differences between the treatment and control groups in each subject area, we conducted a regression analysis using individuals' pretest responses as a covariate.

Teacher assignments.

The use of teacher classroom assignments as an indicator of practice is a methodology developed by CRESST researchers (Aschbacher, 1999; Clare, 2000). Teachers are asked to submit sample lesson materials as well as student work from a particular assignment or unit of study. Along with the lesson materials (texts, hand outs, etc.), teachers complete an extensive open-ended questionnaire about the sequence of instruction leading up to the assignment; the kinds of skills and strategies students were asked to demonstrate; what learning activities students engaged in, and how, and with whom, in carrying out the assignment; what instructional support they received; expectations for student performance; and how students would be assessed. Together, the elicited information, lesson materials, and student work samples are given quality ratings based on a rubric. CRESST research supports the validity and reliability of using classroom assignment ratings as an indicator of classroom practice quality and proxy for classroom observation (Clare & Aschbacher, 2001; Matsumura, 2003). Using assignment ratings to assess practice has added benefits of reducing burdens on both teacher time and data collection resources in comparison to other methods.

The original CRESST Teacher Assignment instrument focused on content alone. For this study, we modified the original content dimensions to reflect Reading Apprenticeship's focus on metacognitive inquiry and added new literacy dimensions that measure the quality of the teacher's goals for student literacy learning in the assignment (their purpose, clarity, and elaboration), the degree and nature of the literacy challenge offered to students in the assignment (complexity of texts, degree of challenge in associated tasks, and degree of metacognitive challenge), and the degree and nature of support for literacy challenge present in the assignment (support for engagement with science text, support for metacognitive challenge).

For each construct a 4-point scale (1 = poor, 4 = excellent) was used to rate the quality for each assignment for separate dimensions. The resulting rubric was designed to allow scorers to gauge from lesson assignment materials, student work, and accompanying teacher descriptions of instruction the nature of literacy instruction, opportunities for engagement with challenging biology or US history texts, metacognitive inquiry into reading and thinking processes, and teacher support for the cognitive and metacognitive demands of the literacy task.

Teachers submitted two class assignments with six corresponding samples of student work, representing high, medium, and low quality. The two assignments came from two different topics in US history – Industrialization and WWII — and two different topics in biology – Genetics and Cell Biology. In addition to the lesson plan and student work samples, each teacher also submitted an in-depth coversheet for each assignment, in which they described aspects of the lesson including any in-class support that was provided, reflection about the lesson implementation and success, and student engagement with the material.

A scoring team composed of researchers and teachers who had not participated in the study was trained to score, using anchor (example) lesson assignments and the rubric. Inter-rater reliability was established by scoring anchor assignments in common and working toward consensus on all rubric scales. Once the training process was complete, each Teacher Assignment was scored by at least two raters on the rubric dimensions. Through discussion and using initial independent scores as a focus for these discussions, the raters established final consensus scores for all dimensions. Assignments with more than a one-point difference on at least one dimension were scored by a third rater. The final assignment ratings represent the consensus score across the raters. In addition to these ratings, we also conducted a descriptive analysis of assignment content and activity to allow us to construct concrete pictures of the instruction in treatment and control classes.

The Intra-class Correlation (ICC) was computed to measure inter-rater reliability of all measures that were scored by multiple raters. The ICC is a measure of the variability within raters as a proportion (reported in decimal form, from zero to one) of the total variation across all ratings and all subjects (Shrout & Fleiss, 1979). In the case of perfect agreement, 100% of the variation is accounted for within raters, and the ICC equals 1. As seen in **Tables 2** and **3**, for nearly all dimensions, the average inter-rater reliability was outstanding (>0.8), or substantial (0.6 to 0.79; see Landis & Koch, 1977).

Table 2. Inter-Rater Reliabilities for History Assignments

Rubric Dimension	Immigration, Industrialization, Urbanization	World War II
Reading opportunities	0.73	0.61
Reading comprehension strategies	0.90	0.90
Metacognitive processes	0.85	0.95
Disciplinary Reading	0.92	0.87
Collaborative meaning making	0.86	0.87
Teacher instruction: Support for reading engagement	0.87	0.91
Teacher instruction: Accommodations for reading	0.95	0.97
Cognitive challenge	0.72	0.82
Teacher instruction: Support for cognitive challenge	0.77	0.80
Monitor: Adjusting instruction	0.88	0.93
Assessment: Student feedback	0.92	0.86

Table 3. Inter-Rater Reliabilities for Biology Assignments

Rubric Dimension	Genetics	Cell Biology
Reading opportunities	0.88	0.86
Reading comprehension strategies	0.94	0.92
Metacognitive processes	0.88	0.87
Disciplinary Reading	0.71	0.83
Collaborative meaning making	0.93	0.89
Teacher instruction: Support for reading engagement	0.88	0.90
Teacher instruction: Accommodations for reading	0.92	0.91
Cognitive challenge	0.80	0.75
Teacher instruction: Support for cognitive challenge	0.63	0.62
Monitor: Adjusting instruction	0.96	0.89
Assessment: Student feedback	0.78	0.70

Teacher interviews.

Retrospective telephone interviews were conducted using semi-structured protocols. All intervention and control teachers were interviewed in the spring of the data collection year. The interviews took place by phone and were recorded for analysis. The interviews focused on eliciting and probing the nature and degree of teachers' implementation of classroom practices targeted by the intervention.

The interview protocol is built around six constructs related to effective integration of literacy and history/science instruction targeted by the intervention: Reading Opportunities; Teacher Support for Student Efforts to Comprehend Content from Text; Metacognitive Inquiry into Reading and Thinking Processes; Specific Reading Comprehension Routines, Tools, Strategies and Processes; Collaboration; and Instruction that Promotes Equity. Each of the constructs is assigned an independent score ranging from 1-4 in accordance with the scoring rubric. Half point scores are utilized at the scorers' discretion. In addition to scoring the six literacy constructs using a four-point scale, the rubric scores Inquiry using a dichotomous scale. Because the interview protocol focuses on literacy, there is insufficient data from the interviews to make a fine-grained determination regarding history or science inquiry. Subjects are given a score of 1 if there is significant evidence in the interview of regular history/science inquiry teaching practices, and a score of 0 otherwise. Thus the teacher interviews yield seven scores:

- *Reading Opportunities*: The purpose of this construct is to describe the degree to which the teacher provides students with the opportunities to read disciplinary texts. Subconstructs of reading opportunities include: the role of reading, frequency of reading, volume of reading, and text variety, and accountability for reading.
- *Teacher Support for Reading*: This construct measures support for student efforts to comprehend content from text. Specifically, the construct considers the availability of social support for reading and comprehending; the extent to which teachers support student problem-solving and meaning-making; the extent to which students, rather than the teacher, do the work of comprehending content from text; and the use of formative assessment during reading and sense-making.
- *Metacognitive Inquiry*: The purpose of this construct is to describe the extent to which students engage in metacognitive inquiry into reading and thinking processes. Specifically, the construct considers opportunities for ongoing metacognitive conversation about how as well as what students read; teacher modeling, guidance and support, and student practice of reading and thinking processes, routines and strategies that support students to become self-monitoring and self-governing readers of history; encouragement and support for grappling with challenging texts, tasks and concepts; and teacher assessment of students' reading and thinking processes.
- *Reading Comprehension*: This construct considers the extent to which the teacher provides modeling, guidance, and support, and students practice an appropriate number of high leverage reading comprehension supporting and disciplinary reading strategies. In addition, it considers whether teacher routinely monitors student use of comprehension strategies and provides additional support and reteaching on an ongoing basis.
- *Collaboration*: This construct considers the degree to which the teacher establishes a small number of structures/routines to support collaborative meaning making and equitable

participation; models and supports collaborative processes; monitors group work; and holds students accountable for collaborative processes and learning.

- *Equity*: The purpose of this construct is to measure the teacher's commitment and support for helping all students achieve high levels of literacy learning. The construct considers the extent to which the teacher consistently differentiates instruction and uses factors within the teacher's control and establishes policies aimed at motivating disengaged and unmotivated students.
- *Inquiry:* This dichotomous variable assesses whether instruction and classroom interactions routinely focus on inquiry/investigation—i.e., on gathering and interpreting evidence— or generally focus on acquiring ready-made knowledge.

An exacting training process was developed to establish inter-rater reliability. Training of five scorers was achieved by having scorers independently listen to and score randomly selected interviews. After all scorers had scored each interview independently, they met and discussed their scores for each construct. The goal was for all five scorers to come within a one-point difference range on each of the overall construct scores. Any discrepancy of more than one point was discussed in detail until consensus was reached. Through this process of discussion of scores and interpretation of the rubric, the rubric was refined to clarify distinctions. The five scorers continued to meet at intervals of approximately 2 weeks throughout the scoring process to work toward consensus on the scores of 10 randomly selected interviews (5 treatment, 5 control). Once again, any discrepancy of more than one point was discussed in detail. After each of these meetings, scores of the five scorers were averaged across constructs and sub-constructs, with these averages reported as the final scores for these 10 interviews. In addition to these regular reliability checks, scorers had the option of requesting a second scorer for any interview where they had a question about their judgment.

Classroom observations.

The purpose of the classroom observations was to provide a snapshot of student literacy learning opportunities in classrooms where teachers received professional development and support to integrate reading and content instruction. While the study included teacher survey, interview data and descriptions of lesson assignments, observations permitted us to learn about these classrooms independently of teacher self-report.

For the classroom observations, a subset of 16 classrooms representing 16 schools was selected from the 124 classrooms and 90 schools that remained in the study. The observations included four history treatment (three cohort 1, one cohort 2), four history control (three cohort 1, one cohort 2), four biology treatment and four biology control classrooms. The 16 teachers were contacted by email and invited to participate in the observations. In order to standardize the observed lessons and to ensure that we would witness literacy practices, we asked to observe a lesson "in which reading plays a central role." All 16 teachers agreed to participate. One biology teacher dropped out of the observation when his class schedule changed unexpectedly. Thus the final sample comprised seven biology classrooms and eight history classrooms. Six classrooms were observed in spring 2008; four classrooms were observed in January 2009; and five classrooms were observed in March 2009. Each teacher was observed for one class period. Because participating classrooms had a variety of scheduling configurations, the duration of classes varied.

We used an observation instrument adapted from the Classroom Observation and Analytic Protocol developed by Horizon Research, Inc. (HRI) for *Looking Inside the Classroom: A Study of K-12 Mathematics and Science Education in the United States* (May, 2003). The HRI instrument contains four components that assess the quality of the design and implementation of mathematics and science lessons. Key indicators within the four components — instructional strategies, science content, literacy opportunities, and classroom culture — were modified from the original HRI instrument to better reflect the goals of the current study and the Reading Apprenticeship framework. In addition, we created a parallel observation and analytic protocol for history by modifying the science instrument to reflect best practices in history.

All observations were conducted by two observers. Observers took detailed field notes during the observed lesson, describing what the teacher and students were doing throughout the lesson, and recording the times activities began and ended. The goal of the field notes was to come as close as possible to a verbatim record of the lesson and classroom interactions. Immediately following the observation, the two observers used their field notes to rate and describe the observed lessons in each of the four component areas. Within each component, observers rated each key indicator on a scale of 1-5, with 1 representing "not at all reflective of best practice" and 5 representing "to a great extent reflective of best practice." Half points were used at observers' discretion. In addition to rating each key indicator, observers provided a synthesis rating for each of the four lesson components with a detailed rationale and supporting evidence for each rating. Synthesis ratings for the lesson components were based on qualitative judgments, rather than on the arithmetical average of the key indicators. In addition to lesson ratings, the observation protocols provide basic descriptive information about the teacher, course and students; a narrative description of the lesson; and information about the duration of instructional and non-instructional activities, and time allocated to different learning structures (whole class, pair/small group and individual).

After scoring the four lesson components independently, the two observers compared their ratings and the evidence supporting them, discussed any disagreements, and came up with consensus ratings on each of the key indicators of lesson quality as well as the synthesis rating for each component. Thus each teacher received one consensus score on each key indicator and for each of the four synthesis ratings of lesson quality. Detailed narratives of the lesson and rationales for synthesis ratings were written up as soon as possible after the scoring, and before the next observation.

Data from the classroom observations was analyzed in fall 2010. The analysis focused on ratings of lesson quality. First, we looked at discipline (biology and history) by treatment group (treatment and control) differences in the four components of lesson quality—1) instructional strategies, 2) disciplinary content and instruction, 3) classroom culture, and 4) literacy opportunities and instruction. Second, in order to establish construct validity for the classroom observation protocol and other instruments developed for the study, and to follow the trajectory from teacher implementation to student learning, we examined relationships between ratings of lesson quality and other implementation and student outcome measures, including Teacher Assignment ratings, teacher interviews, teacher surveys, the student Opportunity to Learn survey, and the Integrated Learning Assessment..

Measures of student learning opportunities and outcomes.

Parental consent forms and student outcome measures, including Opportunity to Learn (OTL) Surveys and state standardized test scores for the baseline and intervention year, were collected for one focus class, for each participating teacher in the control and intervention conditions, to enable us to link baseline scores, intervention year scores, OTL surveys, and teacher implementation measures. Teachers were instructed to administer student surveys and assessments to third period, if they taught the target course at that time, or to the period closest to it that they taught this course. Our intent was to maximize complete data and minimize absenteeism by avoiding first period or periods after lunch. Non-identifiable student standardized test score data were collected for each participating teacher to broaden the sample and its representativeness.

Student Opportunity To Learn (OTL) survey.

Based on prior surveys developed at WestEd for the Performance Assessments in Science (PASS) assessments, student reading surveys developed by Greenleaf and colleagues (Greenleaf, et al., 2001), and CCSSO's Survey of Enacted Curriculum (www.ccsso.org/projects/Surveys_of_Enacted_Curriculum), we developed an Opportunity to Learn (OTL) survey. Parallel forms of the survey asked students about classroom practices related to the integration of literacy and US history or biology, but it also included items related to student engagement, motivation and students' perceptions of themselves as readers and learners. Six key constructs were assessed by the survey and used as outcomes: (1) class emphasis on reading in US history or biology, (2) integration of US history or biology and literacy activity, (3) identifying as a reader, (4) student identity, (5) motivation in class, and (6) course consequences on reading history or science. Using pilot data, a series of exploratory and confirmatory factor analyses were used to validate the factor structure of the student survey items. **Table 4** below provides Cronbach alpha levels for each construct (see **Appendix B** for an item map showing which items measure which construct). The survey was administered to students in both treatment and control groups in spring of the data collection year.

Table 4. Internal consistency reliability coefficients for student opportunity to learn survey measures

	History		Biology	
	Items	Cronbach alpha	Items	Cronbach alpha
Class emphasis on reading in History/Biology	10	0.80	10	0.82
Frequency of integration of content and literacy activity	5	0.76	5	0.74
Motivation in focal class	5	0.78	5	0.77
Academic identity	18	0.94	18	0.94

Integrated learning assessments.

In tracing the effects of the professional development on student reading and learning, we were eager to capture the complexity of student outcomes targeted by the intervention—increased engagement and use of metacognitive and comprehension supporting routines that support students to become self-monitoring and self-governing readers of history or science texts, as well as increased achievement in these subject areas and literacy. We obtained supplementary funding from IES to develop an end-of-year Integrated Learning Assessment (ILA) that affords a closer look at aspects of students' literacy and history or science learning related to the treatment and where we would expect to see differences between treatment and control groups.

We administered a WWII and genetics version of the assessment, respectively, to students in the US history and biology classrooms in the spring of the data collection year in both treatment and control classrooms. The ILA is a performance-based assessment that integrates CRESST's previous work on model-based assessment and explanation tasks with an existing Reading Apprenticeship curriculum-embedded assessment, the CERA. Students read a set of complex, grade-level history or science texts, describe their reading and thinking processes, and answer a set of comprehension questions. Scoring attended to students' comprehension and conceptual understanding, as well as students' literacy, historical or scientific thinking and discourse processes. A report is appended describing the development, scoring, and analysis of the ILA.

The Reading Strategies rubric was based on a 4-point scale. The profile of a score point can be broken down into three main criteria: consideration of the frequency of annotations, the variety in the annotations, and the types of reading strategies used (i.e., general versus discipline-specific). The strategies assessed were drawn from the RA theory of content area reading. **Table 5** provides additional information about the evidence raters looked for while rating and the types of reading strategies students used.

Table 5. Description of Annotations and Reading Strategies

Text Annotations *(evidence found only on text passage)*	Types of Reading Strategies: General	Types of Reading Strategies: Discipline-specific
• *Markings, including:* o Underlines o Highlights o Circlings/Boxings o Connecting lines and arrows o Symbols • *Comments, including:* o Questions o Statements	• Identifying key vocabulary • Identifying unknown vocabulary • Attempting to define unknown vocabulary (e.g., through identifying root words, looking ahead in the text for a definition) • Identifying the main ideas of the text • Paraphrasing • Summarizing • Predicting the content of text sections • Identifying confusions • Using context clues to build understanding	• Making connections to prior history knowledge • Linking ideas together within a document and/or across documents (intertextual reading) • Evaluating the source of a document • Determining bias or point of view • Considering the document in historical context • Identifying cause and effect

A student who received a score of 4, for example, would have shown strong use of reading strategies that may have been demonstrated through annotations seen throughout the text set, a variety of different annotations being used, and evidence of at least one discipline-specific reading strategy. Alternately, a student receiving a score of 1 would have shown evidence of no or minimal use of reading strategies. The annotations may have been minimal, disconnected, or indiscriminate with large sections of the passage highlighted or underlined without apparent purpose.

The new Metacognition rubric was adapted from previous RA work and was piloted during the first scoring session. Like the other ILA rubrics, Metacognition rubric was based on a 4-point scale; the same rubric was used to score both metacognitive items. The profile of a score point could be broken down into three main criteria: the degree to which the student engages with complexities in the text or ideas that require attention, describes thinking processes that occur while reading, and explains an approach to how he/she thinks about the reading. Additionally, raters considered how aware students were of their thinking, the degree of self-monitoring, and finally executive control.

As explained above, the ILA History sample includes two cohorts of history teachers. Cohort 1 treatment teachers participated in the initial professional development during the summer of 2006 and then practiced and attended follow-up sessions during the school year. All Cohort 1 teachers administered the History ILA at the end of the 2007-2008 school year. Similarly, Cohort 2 treatment teachers participated in the initial professional development during the summer of 2007 and then practiced and attended follow-up sessions during the school year. Then, all Cohort 2 teachers administered the History ILA at the end of the 2008-2009 school year.

CRESST researchers trained a team of high school history teachers to score the ILA reading and writing components during the summers following the two ILA administrations. In 2008, CRESST researchers led a 9-day training and scoring session. In 2009, three raters, including two repeat raters and the history content expert on CRESST's research team, completed the training and scoring over the course of four days. To minimize rater bias, all identifying information (student names, teacher names, and school names) was removed from the student assessments. Responses were randomly distributed and divided into packets.

All raters underwent intensive training to introduce and practice scoring procedures, address questions, and ensure that the scoring rubrics were clear. Training began with a focus on the Content and Language rubrics. All student writing responses were then scored by three different raters to achieve greater consistency. The final scores for student responses represent the arithmetic mean of the three raters' scores. After the writing section of the ILA was scored, training focused on the Reading Strategies rubric, and finally training for the Metacognition rubric took place on the final day. Similar to the writing responses, the reading strategies were scored by three different raters. However, in 2008 the metacognitive responses were scored by only two raters. Based on the 2008 results, the 2009 raters only scored the first of the two metacognitive items and were able to triple score this item. The final scores for reading strategies and metacognition represent the arithmetic mean of the raters' scores.

The Intra-class Correlation (ICC) was computed to measure inter-rater reliability of all measures that were scored by multiple raters (i.e., Reading Strategies, Metacognition, Writing Content, and Writing Language). The ICC is a measure of the variability within raters as a proportion (reported in decimal form, from zero to one) of the total variation across all ratings and all subjects (Shrout & Fleiss, 1979). In the case of perfect agreement, 100% of the variation is accounted for within raters, and the ICC equals 1. For all measures, the average measure inter-rater reliability was outstanding (>0.8), with the exception of the second metacognitive score[3], for which it was slightly lower (Landis & Koch, 1977).

Table 6. Inter-Rater Reliabilities for History ILA

	Inter-Rater Reliabilities
Reading Strategies	0.97
Metacognition Q1	0.83
Metacognition Q2	0.75
Writing Content	0.83
Writing Language	0.79

Table 7. Inter-Rater Reliabilities for Biology ILA

	Inter-Rater Reliabilities
Reading Strategies	0.97
Metacognition Q1	0.86
Writing Content	0.97
Writing Language	0.97

In addition to reading the documents provided for US history and biology topics, annotating these documents, and responding to comprehension questions, half of the students in each participating class were asked to carry out a writing task while the other half completed the Degrees of Reading Power test of reading comprehension (described below). Writing tasks required students to integrate information from the texts they had read in the ILA into a coherent essay. A scoring rubric for the essay attended both to students' treatment of the content, as well as their command of language and writing conventions.

Degrees of Reading Power test of reading comprehension.

We administered a standardized reading comprehension test in order to validate the ILA, which was under development at the time of the study. The Degrees of Reading Power test is a

[3] Students received one score for each of the metacognitive items on the ILA. These items were piloted for the first time during this ILA administration. Therefore these scores were not averaged to create a single metacognition score and rater reliability was calculated for each score.

modified cloze test requiring students to read page-long passages of expository texts. Scoring of the test gives an indication of the complexity level of texts students are able to read, based on the surface features of language in the text. The test is thus both a criterion and norm-referenced test.

State standardized test scores.

To broadly assess student performance in US history or biology and reading comprehension, we relied on available, state mandated criterion-referenced tests. Because Arizona and California do not have the same tests, we limited this data to California schools alone. The California state standardized tests in English and US history or biology are not particularly well suited to the intervention of this study. The vast majority of items, for instance, are concept identification or factual recall questions on content that require very little reading. Conversely, while the English test requires reading, the vast majority of items focus on literature. Nevertheless, these tests, however distal a measure of student achievement in the areas targeted, represent both a readily available and critical measure of the impact of Reading Apprenticeship professional development, given the increasingly high stakes attached to state standardized measures.

Standardized test data collection proved quite difficult across multiple school districts with varied research capabilities and was not completed until the Spring of 2011. Data were requested from every district from which we had had a teacher participate, even if the specific teacher was not retained in the study. Our final data collection resulted in two types of data. For students for whom we had obtained parental consent, we collected linked, longitudinal test score data. We also collected anonymous cross-sectional data linked to teachers, but not to specific students, for those for whom we did not obtain parental consent. For linked students we collected: CST English Language Arts and ELA Reading Comprehension for the year that students were exposed to study teachers (posttest) and the year prior to *student exposure* to study teachers (baseline). The baseline measures were used as covariates in the longitudinal impact analysis models. For students who are unlinked in the dataset, we collected CST English Language Arts, ELA Reading Comprehension, and Biology/History scores for students in participating teachers classrooms in the year of random assignment (pretest), the first implementation year, and the second implementation year (posttest). The pretest student test score data were aggregated to the teacher level and used as covariates in our cross-sectional student test score impact analysis models.

The English Language Arts and Reading tests are not vertically scaled and thus do not have the same meaning across different grade levels. To convert the scores to an identical metric so that test score data from all of the grades can be analyzed together, within each grade, test score data were normalized by subtracting the state mean from each student's score and dividing by the state (ELA) or sample (reading comprehension) standard deviation. Normalized in this way, the test score data represent the *relative ranking* of students in the state rather than the absolute level of performance, and the impact estimates reflect the impacts relative to the state ranking.

Table 8 shows the schedule for the various data collected to gauge evidence of teacher implementation and student learning outcomes.

Table 8. Data Collection Schedule

	Baseline	*Year 2*	*Year 3*
Teacher Practice			
Teacher Surveys (Instructional Beliefs/Practice)	Summer	Summer	Summer
Teacher Assignments			Fall/Spring
Teacher Interviews			Spring
Student Opportunity to Learn Surveys			Spring
Student Outcome Measures			
Integrated Learning Assess			Spring
Degrees of Reading Power			Spring
State Test Scores (Biology, History, & ELA)	Spring	Spring	Spring

Data Analysis

To estimate program impacts, outcomes for teachers and students in treatment classrooms were compared to the outcomes for their counterparts in control classrooms. We analyzed the effectiveness of intervention using hierarchical regression models to account for clustering of the data by school (Goldstein, 1987; Raudenbush & Bryk, 2002; Murray, 1998). In each of the impact analyses, we controlled for baseline (pre-test) measures of outcome variables when available, randomization strata (i.e., pairs), and student-level covariates when analyzing student outcomes.

FINDINGS

Retention of Schools and Teachers

Tables 9 and 10 below show the number of teachers and schools randomly assigned to US history and biology treatment and control groups, respectively, as well as the data retention rates for each data source.

US History

As shown in **Table 9**, 86.4 percent of history treatment teachers and 95.2 percent of history control teachers provided responses on the baseline teacher survey, 78 and 79.6 percent of treatment and control teachers participated in the 1st-year post-implementation teacher survey, and 50.8 and 55.1 percent participated in the 2nd-year post-implementation survey. Return rates for other types of data after the 2nd study year were slightly lower than those for the 2nd-year post-implementation survey.

Teacher interviews were conducted with 55.6 percent of teachers and lesson assignment data were collected from 49.1 percent of randomized teachers. Integrated Learning Assessment

(ILA) data were collected from 45.4 percent of history teachers, student OTL survey data were secured from 50.9 percent of randomly assigned history teachers. Longitudinal student test score data were collected from 43.5 percent of teachers, while cross-sectional student test score data were collected from 45.4 percent of teachers. Approximately half of the recruited sample of teachers was retained in the study by the end of the second year, and student test score data were secured for less than half of the initial sample. Although the attrition levels are relatively high, there were no statistically significant differences in data return rates across treatment and control teachers. The school participation chart at the right side of **Table 9** shows similar data return rates for schools as that for teachers.

Table 9. History teacher and school retention by data source

	Teachers			Schools		
	Overall	Treatment	Control	Overall	Treatment	Control
Recruited teachers/schools	108	59	49	82	45	37
Teacher baseline survey	98 (90.7%)	51 (86.4%)	47 (95.2%)	75 (91.5%)	39 (86.7%)*	36 (97.3%)
Teacher year 1 posttest survey	85 (78.7%)	46 (78.0%)	39 (79.6%)	66 (80.5%)	35 (77.8%)	31 (83.8%)
Teacher year 2 posttest survey	57 (52.8%)	30 (50.8%)	27 (55.1%)	50 (61.0%)	25 (55.6%)	25 (67.6%)
Teacher interview	60 (55.6%)	32 (54.2%)	28 (57.1%)	52 (63.4%)	27 (60.0%)	25 (67.6%)
Lessons assignment	53 (49.1%)	27 (45.7%)	26 (53.1%)	46 (56.1%)	22 (48.9%)	24 (64.9%)
Student Integrated Learning Assessment	49 (45.4%)	25 (42.4%)	24 (49.0%)	42 (51.2%)	20 (44.4%)	22 (59.5%)
Student OTL survey	55 (50.9%)	29 (49.1%)	26 (43.1%)	45 (54.9%)	22 (48.9%)	23 (62.2%)
Student Degrees of Reading Power test	43 (39.8%)	21 (35.6%)	22 (44.9%)	38 (46.3%)	17 (37.8%)*	21 (56.7%)
Student cross-sectional test scores	49 (45.4%)	27 (45.8%)	22 (44.9%)	40 (48.8%)	22 (48.9%)	18 (48.6%)
Student longitudinal test scores	47 (43.5%)	26 (44.1%)	21 (42.9%)	38 (46.3%)	20 (44.4%)	18 (48.6%)

Notes.
* Significantly different from zero at the .10 level, two-tailed test.

Biology

As shown in **Table 10**, 69.6 percent of biology treatment teachers and 92.7 percent of biology control teachers provided responses on the baseline teacher survey, 50 and 83.6 percent of biology treatment and control teachers participated in the 1st-year post-implementation teacher survey, and 44.6 and 60 percent participated in the 2nd-year post-implementation survey. The initial loss of biology treatment teachers continued in the ensuing years for both treatment and control teachers at a rate similar to that of the history teachers. Return rates for other types of data after the 2nd study year were slightly lower than those for the 2nd-year post-implementation survey. Teacher interviews were conducted with 52.2 percent of teachers and lesson assignment data were collected from 42.9 percent of treatment teachers and 52.7 percent of control teachers. Integrated Learning Assessment (ILA) data were collected from 45 percent of biology teachers, student OTL survey data were secured from approximately 49.5 percent of randomly assigned biology teachers. Longitudinal student test score data were collected from only 23.2 percent of biology treatment teachers and 49.1 percent of biology control teachers, while cross-sectional student test score data were collected from32.1percent of treatment teachers and 50.9percent of control teachers. Approximately half of the recruited sample of teachers was retained in the study by the end of the second year, and student test score data were secured for less than half of the initial sample, and only about one quarter of the treatment teachers. These differences in data return rates across treatment-group and control-group teachers were statistically significant, and such differences could bias impact estimates if participation is associated with outcome measures. However, as shown below, there was little evidence of selective participation based on the analyses of baseline differences across treatment and control schools. The school participation chart at the right side of **Table 10** shows similar data return rates for schools as that for teachers.

Table 10. Biology teacher and school retention by data source

		Teachers			Schools	
	Overall	Treatment	Control	Overall	Treatment	Control
Recruited teachers/schools	111	56	55	78	39	39
Teacher baseline survey	90 (81.1%)	39 (69.6%)**	51 (92.7%)	69 (88.5%)	32 (82.0%)*	37 (94.9%)
Teacher year 1 posttest survey	74 (66.7%)	28 (50.0%)***	46 (83.6%)	58 (74.4%)	25 (64.1%)**	33 (84.6%)
Teacher year 2 posttest survey	58 (52.2%)	25 (44.6%)	33 (60.0%)	47 (60.3%)	22 (56.4%)	25 (64.1%)
Teacher interview	58 (52.2%)	24 (42.9%)*	34 (61.8%)	47 (60.3%)	21 (53.8%)	26 (66.7%)
Lessons assignment	53 (47.7%)	24 (42.9%)	29 (52.7%)	43 (55.1%)	21 (53.8%)	22 (56.4%)
Student Integrated Learning Assessment	50 (45.0%)	22 (39.3%)	28 (50.9%)	41 (52.6%)	19 (48.7%)	22 (56.4%)
Student OTL survey	55 (49.5%)	23 (41.1%)	32 (58.2%)	46 (59.0%)	21 (53.8%)	25 (64.1%)
Student Degrees of Reading Power test	49 (44.1%)	21 (37.5%)	28 (50.9%)	40 (51.3%)	18 (46.1%)	22 (56.4%)
Student cross-sectional test scores	46 (41.4%)	18 (32.1%)	28 50.9%)	38 (48.7)	14 (35.9)	24 (61.5%)
Student longitudinal test scores	40 (36.0%)	13 (23.2%)**	27 (49.1%)	33 (42.3%)	11 (28.2%)**	22 (56.4%)

Notes.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test

Equivalence of Treatment and Control Groups

Although retention and data return rates among teachers randomly assigned to condition were relatively low, attrition patterns were fairly similar for treatment and control schools in the History sample. For the Biology sample, however, retention and data return rates were lower for treatment group participants than for control group participants. To describe treatment/control group equivalence at the time of random assignment and at subsequent data collection periods, we present school-, teacher-, and student characteristics by data source. **Table 11** shows school characteristics by **history** treatment/control status for the randomized sample, teacher baseline survey sample, and teacher final survey sample. Note there were no initial significant differences between treatment and control sample detected, nor did significant differences between the samples emerge over time as the sample was decreased in size. Overall, the randomized and teacher baseline and final survey samples show a high degree of similarity, with few meaningful differences in school performance and demographic characteristics.

Table 12 shows school characteristics by **biology** treatment/control status for the randomized sample, teacher baseline survey sample, and teacher final survey sample. Similar to the history teacher sample, the biology randomized and teacher baseline and final survey samples show a high degree of similarity, with few meaningful differences in school performance and demographic characteristics in the attrited sample.

Tables 13 and 14 show pre-intervention characteristics of students in treatment and control schools based on the longitudinal test score, cross-sectional test score, and student OTL survey samples, for history and science, respectively. Note that parental consent was required to collect student-level longitudinal test score and OTL data, so group differences in student characteristics reflected in these tables could be due to differences in teacher participation rates, student participation rates, or both factors.

For the History sample (**Table 13**), statistically significant differences between treatment and control schools were present; as indicated by the longitudinal test score sample, treatment schools had higher proportions of Asian students (12% vs. 4%). Treatment schools also exhibited baseline test scores that were between .01 and .29 of a standard deviation higher than those in control schools, although only group differences in reading comprehension approached statistical significance at conventional levels. No treatment/control differences were apparent for the cross-sectional and OTL History samples. For the Biology sample (**Table 14**), no substantial treatment/control differences were apparent across each of the samples, although there was a noticeably higher proportion of Latino students in the treatment group than the control group (49 percent compared to 32 percent, $p < .10$). Overall, there was little evidence of selective participation based on the analyses of baseline differences across treatment and control schools. Although few statistically significant baseline differences in teacher or student characteristics were detected, unobserved differences across groups could still be associated with biases in estimated program impacts.

Table 11. Pre-intervention characteristics by treatment/control status for randomized sample, baseline survey sample, and final survey sample – History Sample

	Treatment	Control	Difference	p-value	Diff/SD
Randomized sample					
School characteristics (82 schools)					
African American (%)	7.60	5.97	1.63	0.28	0.24
Hispanic (%)	44.38	50.22	-5.84	0.33	-0.12
White (%)	34.20	32.86	1.34	0.81	0.04
English learners (%)	16.55	18.99	-2.44	0.44	-0.14
Free/reduce priced meals (%)	39.35	40.67	-1.32	0.83	-0.03
CA Academic Performance Index	720.03	694.76	25.27	0.35	0.04
CA History standardized test	337.68	333.94	3.75	0.60	0.01
AZ Reading standardized test	678.50	677.86	0.64	0.93	0.00
Teacher baseline survey sample					
School characteristics (75 schools)					
African American (%)	7.23	5.91	1.32	0.38	0.20
Hispanic (%)	42.48	49.53	-7.05	0.27	-0.15
White (%)	35.54	33.39	2.15	0.71	0.06
English learners (%)	16.11	18.38	-2.27	0.48	-0.13
Free/reduce priced meals (%)	36.87	39.66	-2.79	0.66	-0.07
CA Academic Performance Index	727.00	704.82	22.18	0.40	0.03
CA History standardized test	339.38	333.94	5.44	0.47	0.02
AZ Reading standardized test	679.67	677.86	1.81	0.83	0.00
Teacher characteristics (98 teachers)					
Female	0.56	0.45	0.11	0.27	0.23
Years teaching History	9.43	9.88	-0.45	0.63	-0.08
Years teaching at school	7.11	8.94	-1.83*	0.09	-0.31
History major	0.07	0.11	-0.04	0.53	-0.14
Teacher year 2 posttest survey sample					
School characteristics (50 schools)					
African American (%)	7.05	6.29	0.77	0.66	0.11
Hispanic (%)	43.22	45.08	-1.85	0.81	-0.04
White (%)	37.26	37.94	-0.68	0.92	-0.02
English learners (%)	14.16	14.99	-0.83	0.82	-0.06
Free/reduce priced meals (%)	35.08	36.14	-1.06	0.89	-0.03
CA Academic Performance Index	739.25	706.47	32.78	0.36	0.05
CA History standardized test	342.14	334.26	7.88	0.42	0.02
AZ Reading standardized test	675.40	680.20	-4.80	0.63	-0.01
Teacher characteristics (57 teachers)					
Female	0.62	0.41	0.21	0.11	0.42
Years teaching History	9.10	10.39	-1.29	0.42	-0.21
Years teaching at school	7.33	9.52	-2.19	0.15	-0.36
History major	0.12	0.10	0.02	0.79	0.08

Notes. Effect size calculated by dividing difference by sample standard deviation.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test

Table 12. Pre-intervention characteristics by treatment/control status for randomized sample, baseline survey sample, and final survey sample – Biology Sample

	Treatment	Control	Difference	p-value	Diff/SD
Randomized sample					
School characteristics (78 schools)					
African American (%)	7.36	7.66	-0.29	0.89	-0.04
Hispanic (%)	51.55	45.00	6.55	0.26	0.14
White (%)	28.70	33.12	-4.42	0.45	-0.14
English learners (%)	18.57	18.56	0.01	0.99	0.00
Free/reduce priced meals (%)	42.90	39.76	3.14	0.58	0.08
CA Academic Performance Index	707.66	703.66	4.00	0.88	0.01
CA Biology standardized tests	335.31	335.03	0.28	0.97	0.00
AZ Reading standardized test	673.89	680.71	-6.83	0.25	-0.01
Teacher baseline survey sample					
School characteristics (69 schools)					
African American (%)	8.57	8.07	0.50	0.83	0.06
Hispanic (%)	50.57	43.86	6.71	0.27	0.14
White (%)	29.57	33.26	-3.69	0.56	-0.12
English learners (%)	16.35	18.59	-2.24	0.47	-0.13
Free/reduce priced meals (%)	42.01	39.26	2.74	0.66	0.07
CA Academic Performance Index	702.00	712.13	-10.13	0.73	-0.01
CA Biology standardized tests	333.31	337.39	-4.08	0.57	-0.01
AZ Reading standardized test	673.89	680.71	-6.83	0.25	-0.01
Teacher characteristics (90 teachers)					
Female	0.53	0.71	-0.18*	0.09	-0.37
Years teaching science	5.00	5.44	-0.44	0.63	-0.10
Years teaching at school	6.19	5.99	0.20	0.72	0.04
Biology major	0.33	0.33	0.01	0.94	0.02
Teacher year 2 posttest survey sample					
School characteristics (47 schools)					
African American (%)	9.64	6.68	2.96	0.34	0.37
Hispanic (%)	48.21	41.60	6.61	0.36	0.15
White (%)	29.82	34.34	-4.52	0.55	-0.14
English learners (%)	15.93	17.48	-1.55	0.67	-0.09
Free/reduce priced meals (%)	43.62	34.64	8.98	0.22	0.23
CA Academic Performance Index	712.53	737.21	-24.68	0.47	-0.03
CA Biology standardized tests	338.11	343.57	-5.46	0.52	-0.02
AZ Reading standardized test	677.50	679.50	-2.00	0.77	0.00
Teacher characteristics (58 teachers)					
Female	0.56	0.76	-0.20	0.11	-0.42
Years teaching science	5.14	5.38	-0.24	0.88	-0.05
Years teaching at school	6.02	5.41	0.61	0.41	0.12
Biology major	0.35	0.33	0.01	0.92	0.03

Notes. Effect size calculated by dividing difference by sample standard deviation.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.

*** Significantly different from zero at the .01 level, two-tailed test

Table 13. Pre-intervention Characteristics of Students in Treatment and Control Schools – History Sample

	Treatment	Control	Difference	p-value	Diff/SD
Longitudinal Test Score Sample (38 Schools)					
Student Characteristics					
Female	0.51	0.51	0.00	0.98	0.00
English Learner	0.23	0.21	0.01	0.86	0.03
African American	0.06	0.05	0.01	0.50	0.05
Asian	0.12	0.04	0.09**	0.01	0.30
Latino	0.36	0.41	-0.05	0.70	-0.11
Other	0.32	0.36	-0.04	0.36	-0.09
White	0.14	0.14	-0.01	0.68	-0.01
English Language Arts CST 05 (std)	0.11	-0.13	0.24	0.25	0.23
Reading Comprehension 05 (std)	0.13	-0.26	0.39*	0.09	0.29
Cross-sectional Test Score Sample (40 Schools)					
Student Characteristics					
Female	0.47	0.43	0.03	0.50	0.06
English Learner	0.17	0.30	-0.13	0.36	-0.30
African American	0.05	0.07	-0.02	0.87	-0.09
Asian	0.15	0.06	0.09	0.25	0.29
Latino	0.26	0.47	-0.21	0.72	-0.43
Other	0.49	0.36	0.13	0.86	0.26
White	0.04	0.04	0.01	0.66	0.04
Baseline English Language Arts CST	0.10	0.05	0.06	0.52	0.06
Baseline Reading Comprehension	0.14	-0.05	0.19	0.18	0.29
Baseline History CST	331.33	330.86	0.46	0.54	0.01
Student OTL Survey Sample (45 Schools)					
Student Characteristics					
African American	0.04	0.04	0.00	0.50	0.01
Latino	0.31	0.37	-0.07	0.64	-0.15
White	0.30	0.35	-0.05	0.22	-0.10
Other	0.36	0.24	0.12	0.04	0.25
Non-English Speaker	0.34	0.36	-0.02	0.92	-0.04

Notes. p-values are based o multilevel regression models in which treatment group status is included as a covariate. Effect size calculated by dividing difference by sample standard deviation.

* Significantly different from zero at the .10 level, two-tailed test.

** Significantly different from zero at the .05 level, two-tailed test.

*** Significantly different from zero at the .01 level, two-tailed test

Table 14. Pre-intervention Characteristics of Students in Treatment and Control Schools – Biology Sample

	Treatment	Control	Difference	p-value	Diff/SD
Longitudinal Test Score Sample (33 Schools)					
Student Characteristics					
Female	0.55	0.54	0.01	0.74	0.02
English Learner	0.33	0.38	-0.05	0.51	-0.10
African American	0.12	0.04	0.08	0.25	0.34
Asian	0.09	0.14	-0.05	0.31	-0.16
Latino	0.40	0.31	0.10	0.45	0.21
Other	0.31	0.38	-0.07	0.80	-0.15
White	0.09	0.14	-0.05	0.48	-0.17
English Language Arts CST 05	0.03	0.04	-0.01	0.88	-0.01
Reading Comprehension 05	0.14	0.20	-0.05	0.91	-0.04
Mathematics CST 05	0.02	0.01	0.00	0.83	0.00
Cross-sectional Test Score Sample (38 Schools)					
Student Characteristics					
Female	0.47	0.50	-0.03	0.64	-0.06
English Learner	0.38	0.32	0.05	0.66	0.11
African American	0.08	0.04	0.04	0.18	0.18
Asian	0.09	0.16	-0.07	0.35	-0.22
Latino	0.46	0.35	0.11	0.31	0.22
Other	0.24	0.33	-0.09	0.68	-0.19
White	0.13	0.12	0.01	0.72	0.03
Baseline ELA CST	0.17	0.05	0.11	0.32	0.11
Baseline Reading Comprehension	0.13	0.17	-0.03	0.95	-0.03
Baseline Biology CST	334.92	335.35	-0.43	0.77	-0.01
Student OTL Survey Sample (46 Schools)					
Student Characteristics					
African American	0.07	0.03	0.04	0.13	0.17
Latino	0.49	0.32	0.17*	0.09	0.34
White	0.17	0.31	-0.14	0.44	-0.32
Other	0.27	0.33	-0.06	0.46	-0.14
Non-English Speaker	0.17	0.18	-0.01	0.71	-0.02

Notes. p-values are based o multilevel regression models in which treatment group status is included as a covariate. Effect size calculated by dividing difference by sample standard deviation.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test

OUTCOME ANALYSES

H1: Teacher Outcomes: Integration of Literacy into Instructional Practice in High School US History and Biology

Teacher Surveys: Analysis of pre- and post- surveys at the end of Year 2 offered evidence that the intervention had produced increased teacher facility in integrating US history and literacy teaching, or biology and literacy teaching, respectively. These results are presented in **Tables 15 and 16.**

For the History sample, we found significant differences favoring the treatment group relative to the control group on 11 of the 14 sub-constructs shown in **Table 15** at the end of the intervention year. These constructs included:

- *Reading Opportunities: Texts,* the range of reading materials used in instruction;
- *Reading Opportunities: Content*, the extent to which history content from reading materials is acquired through student work and meaning making (versus delivered directly by the teacher through lecture);
- *Collaboration, Teacher Modeling:* the extent to which teachers modeled and supported collaboration instructionally;
- *Metacognitive Inquiry, Teacher Modeling:* the extent to which teachers modeled metacognitive inquiry and reading routines;
- *Metacognitive Inquiry, Student Practice:* the extent to which students had opportunities to practice metacognitive inquiry and reading routines;
- *Comprehension Strategies, Teacher Modeling:* the extent to which teachers provided modeling and explicit instruction in comprehension-supporting strategies;
- *Comprehension Strategies, Student Practice:* the extent to which students had opportunities to practice comprehension-supported strategies;
- *Negotiating Success, Instruction:* the extent to which teachers modify instruction on the basis of student need to promote successful engagement and learning;
- *Negotiating Success – Assessment:* the extent to which formative assessment informs instruction; and
- *Teaching Philosophy, Diversity:* the extent to which teachers believe learning differences and varied language and cultural backgrounds can be an asset in the classroom.

A 12th sub-construct, Teaching Philosophy, Learning: the extent to which teachers believe student learning entails constructing new knowledge in relation to prior conceptions, which is facilitated by interaction with others, approached significance. In addition, we looked separately at items within the comprehension strategies item bank that were focused on three high-leverage aspects of historical reading: use of text structures to support comprehension, language learning strategies, and disciplinary thinking processes. Differences between treatment and control teachers on these three sub-constructs were also statistically significant, favoring the treatment group. Effect sizes for survey differences ranged from moderate (.51) to very large (1.64), with the bulk of treatment/control differences hovering between .70 and .96.

Table 15. Treatment/Control differences in Post-surveys – History Sample

	Treatment	Control	Difference	p-val	Diff/SD
Teacher Survey – 2nd Post-Survey					
Reading Opportunities - Texts	3.25	3.02	0.23**	0.05	0.56
Reading Opportunities - Learning Structures	3.13	2.91	0.22	0.17	0.41
Reading Opportunities - Content	3.47	3.01	0.46**	0.01	0.74
Collaboration - Teacher Modeling	3.56	3.04	0.52***	0.00	0.77
Collaboration - Student Practice	3.36	2.93	0.43**	0.01	0.59
Metacognitive Inquiry - Teacher Modeling	3.69	3.25	0.44***	0.00	0.75
Metacognitive Inquiry - Student Practice	3.39	2.42	0.98***	0.00	1.64
Comprehension Strategies -Teacher Modeling	3.68	3.31	0.37***	0.00	0.87
Comprehension Strategies - Student Practice	3.25	2.73	0.53***	0.00	0.96
Negotiating Success - Instruction	3.67	3.21	0.46***	0.00	0.74
Negotiating Success - Assessment	3.72	3.29	0.43**	0.01	0.61
Teaching Philosophy - Reading	4.30	4.13	0.17	0.19	0.46
Teaching Philosophy - Learning	4.45	4.28	0.17*	0.09	0.31
Teaching Philosophy - Diversity	4.72	4.43	0.29**	0.02	0.51

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test

For the **Biology** sample, we found significant differences favoring the treatment group relative to the control group on 11 of the 14 sub-constructs shown in **Table 16** at the end of the intervention year. These constructs included:

- *Reading Opportunities: Learning Structures,* how reading assignments are carried out, with whom, and in what contexts;
- *Reading Opportunities: Content*, the extent to which biology content from reading materials is acquired through student work and meaning making (versus delivered directly by the teacher through lecture);
- *Collaboration, Teacher Modeling:* the extent to which teachers modeled and supported collaboration instructionally;
- *Metacognitive Inquiry, Student Practice:* the extent to which students had opportunities to practice metacognitive inquiry and reading routines;
- *Comprehension Strategies, Teacher Modeling:* the extent to which teachers provided modeling and explicit instruction in comprehension-supporting strategies;
- *Negotiating Success – Assessment:* the extent to which formative assessment informs instruction; and
- *Teaching Philosophy, Reading:* the extent to which teachers view reading as valuable in biology learning.

Two additional constructs approached significance in the biology sample:

- *Teaching Philosophy, Learning:* the extent to which teachers believe the extent to which teachers believe student learning entails constructing new knowledge in relation to prior conceptions, which is facilitated by interaction with others, and
- *Comprehension Strategies, Student Practice:* the extent to which students had opportunities to practice comprehension-supported strategies.

Table 16. Treatment/Control differences in Post-surveys – Biology Sample

	Treatment	Control	Difference	p-val	Diff/SD
Teacher Survey – 2nd Post-Survey					
Reading Opportunities - Texts	3.26	3.08	0.18	0.13	0.44
Reading Opportunities - Learning Structures	3.10	2.77	0.33**	0.00	0.69
Reading Opportunities - Content	3.39	3.03	0.36**	0.03	0.52
Collaboration - Teacher Modeling	3.30	2.97	0.33**	0.03	0.59
Collaboration - Student Practice	2.97	2.82	0.15	0.36	0.22
Metacognitive Inquiry - Teacher Modeling	3.13	2.86	0.26	0.11	0.46
Metacognitive Inquiry - Student Practice	2.75	2.17	0.58**	0.01	0.87
Comprehension Strategies -Teacher Modeling	3.49	2.80	0.69***	0.00	0.66
Comprehension Strategies - Student Practice	3.08	2.85	0.23*	0.08	0.35
Negotiating Success - Instruction	3.72	3.52	0.20	0.13	0.32
Negotiating Success - Assessment	3.54	2.81	0.73***	0.00	0.95
Teaching Philosophy - Reading	4.63	3.89	0.73***	0.00	1.28
Teaching Philosophy - Learning	4.81	4.57	0.24*	0.06	0.48
Teaching Philosophy - Diversity	4.32	4.21	0.11	0.36	0.26

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test

Again, the effect sizes for these differences in survey response on various constructs range from moderate (.52) to strong (1.28), with most in the .59 - .95 range.

In both US history and biology classes, teacher responses to surveys at the end of the study thus show differences between intervention and control teachers in both their knowledge about the role reading plays in learning and in their repertoire of instructional practices. According to teacher reports on the survey, intervention classrooms are distinguished from control classrooms in the degree to which students—rather than teachers—are more frequently doing the work of comprehending. Further, they receive greater teacher support for carrying out this work, and this support frequently takes the form of teacher modeling and metacognitive inquiry into reading and thinking processes.

Teacher Assignments:

As described above, teacher assignments were collected from treatment and control teachers for two instructional units, in each content area – US history and biology. These assignments and accompanying artifacts from classroom instruction were scored separately.

History Sample

As seen in **Table 17**, for both units of instruction, history treatment teachers significantly outscored control teachers on four dimensions: (a) reading comprehension strategies, (b) metacognitive processes, (c) support for reading engagement, and (d) student feedback. On the WWII unit taught later in the school year, US history treatment teachers also significantly outperformed control teachers on three additional dimensions: (e) reading opportunities, (f) disciplinary reading, and (g) collaborative meaning making. The magnitude of difference for these significant dimensions were quite large.

Table 17. Teaching assignment differences – History Sample

	Treatment	Control	Difference	p-val	Diff/SD
Industrialization, Urbanization, Immigration					
Reading opportunities	3.52	3.19	0.33	0.11	0.42
Reading comprehension strategies	3.29	2.11	1.18***	0.00	1.15
Metacognitive processes	2.19	1.32	0.87***	0.00	1.62
Disciplinary reading	2.19	1.97	0.22	0.44	0.20
Collaborative meaning making	2.54	2.20	0.34	0.27	0.32
Teacher instruction: Support for reading engagement	3.43	2.30	1.12**	0.01	1.12
Teacher instruction: accommodations for reading	1.77	1.50	0.27	0.37	0.30
Cognitive challenge	2.68	2.71	-0.03	0.89	-0.04
Teacher instruction: Support for cognitive challenge	2.79	2.87	-0.07	0.76	-0.10
Monitor: Adjusting instruction	1.91	1.69	0.22	0.44	0.24
Assessment: Student feedback	3.17	2.49	0.67***	0.00	0.67
WW II					
Reading opportunities	3.53	3.23	0.30**	0.04	0.50
Reading comprehension strategies	2.95	1.75	1.20***	0.00	1.63
Metacognitive processes	2.06	1.34	0.72**	0.01	1.28
Disciplinary reading	2.72	1.96	0.76**	0.05	0.90
Collaborative meaning making	2.32	1.75	0.57**	0.03	0.65
Teacher instruction: Support for reading engagement	3.12	2.00	1.12***	0.00	1.15
Teacher instruction: accommodations for reading	2.12	1.44	0.68	0.23	0.67
Cognitive challenge	3.22	2.96	0.26	0.33	0.32
Teacher instruction: Support for cognitive challenge	2.78	2.95	-0.18	0.34	-0.24
Monitor: Adjusting instruction	1.50	1.95	-0.45	0.12	-0.38
Assessment: Student feedback	3.18	2.56	0.62**	0.01	0.80

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test

Biology sample

As **Table 18** shows, biology treatment teachers also scored significantly higher than their control counterparts on several dimensions. For the genetics unit, significant differences were found in (a) reading comprehension strategies, (b) metacognitive processes, (c) collaborative meaning making, and (d) support for reading engagement. For the cell biology unit, in addition to these four dimensions, treatment teachers significantly outscored control teachers on (e) reading opportunities, (f) disciplinary reading, and (g) monitoring and adjusting instruction based on student responses to the lesson. As in the case of history, the magnitude of difference was quite large for these dimensions.

Table 18. Teaching assignment differences – Biology Sample

	Treatment	Control	Difference	p-val	Diff/SD
Genetics					
Reading opportunities	3.38	3.08	0.29	0.29	0.33
Reading comprehension strategies	3.17	1.62	1.55***	0.00	1.56
Metacognitive processes	2.04	1.22	0.81***	0.00	1.57
Disciplinary reading	1.55	1.59	-0.03	0.87	-0.05
Collaborative meaning making	2.37	0.99	1.38***	0.00	1.56
Teacher instruction: Support for reading engagement	3.16	1.95	1.21***	0.00	1.33
Teacher instruction: accommodations for reading	2.15	1.82	0.33	0.25	0.32
Cognitive challenge	2.76	3.04	-0.28	0.17	-0.37
Teacher instruction: Support for cognitive challenge	3.27	3.26	0.01	0.96	0.02
Monitor: Adjusting instruction	2.51	1.86	0.65*	0.06	0.54
Assessment: Student feedback	3.23	3.15	0.08	0.72	0.11
Cell Biology					
Reading opportunities	3.42	2.86	0.55**	0.01	0.67
Reading comprehension strategies	3.20	1.64	1.56***	0.00	1.83
Metacognitive processes	2.26	1.17	1.09***	0.00	2.52
Disciplinary reading	1.77	1.35	0.42**	0.03	0.82
Collaborative meaning making	2.03	1.41	0.62**	0.02	0.85
Teacher instruction: Support for reading engagement	3.44	2.04	1.40***	0.00	1.51
Teacher instruction: accommodations for reading	2.22	1.88	0.33	0.30	0.32
Cognitive challenge	2.65	2.83	-0.18	0.32	-0.29
Teacher instruction: Support for cognitive challenge	3.15	3.28	-0.13	0.50	-0.20
Monitor: Adjusting instruction	2.40	1.41	0.99***	0.00	1.17
Assessment: Student feedback	3.27	3.38	-0.11	0.57	-0.16

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test

Teacher assignment differences show treatment teachers, regardless of subject area, attending more to reading in instruction and supporting student reading through comprehension strategy instruction, metacognitive processes, collaborative meaning making, and strategies for engagement. In addition, treatment teachers appear to focus more on the unique disciplinary aspects of reading in US history and biology, and to make adjustments in their lessons based on student responses to instruction. The metacognitive processes and collaborative meaning making appear to make student thinking public and available for formative assessment, which in turn can drive instruction in responsive ways.

Teacher Interviews:

As described above, teacher interviews were recorded and subsequently rated on a 4 point rubric on 5 dimensions: reading opportunities, support for student reading, metacognitive inquiry, reading comprehension strategies, and equity. A sixth dimension, inquiry, was developed after the interviews were conducted and was coded as a dichotomous outcome.

History Sample

Table 19 below shows mean ratings for treatment and control teachers in US history. The results indicate that teachers in the intervention group exhibit substantially higher interview ratings than their counterparts in the control group in the areas of (a) Reading Opportunities, (b) Support for Student Reading, (c) Metacognitive Inquiry, and (d) Reading Comprehension Strategies. Thus, the volume and kinds of reading students are asked to do, the degree and type of support for student engagement with course texts, the explicit teaching and modeling and guided practice using specific comprehension routines, and teachers' attention to equitable participation and support for various students in these classrooms differ significantly. Moreover, these differences are quite large in magnitude, ranging from 0.59 to 2.12 standard deviation units.

Table 19. Teacher interviews: Differences by treatment/control group – History Sample

	Treatment	Control	Difference	p-val	Diff/SD
Reading opportunities	3.08	2.72	0.36**	0.04	0.59
Support for student reading	2.95	2.32	0.62***	0.00	0.99
Metacognitive inquiry	2.41	1.47	0.94***	0.00	2.12
Reading comprehension strategies	2.51	2.01	0.50**	0.01	0.82
Equity	2.74	2.18	0.56**	0.03	0.65
Inquiry	0.13	0.08	0.04	0.63	0.16

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test

Biology Sample

As **Table 20** below indicates, biology treatment teachers exhibited substantially higher interview ratings than biology control teachers in the areas of (a) Support for Student Reading, (b) Metacognitive Inquiry, and (c) Reading Comprehension Strategies. Unlike their history counterparts, biology treatment teachers did not offer substantially different volume or range of reading than teachers in control classrooms. Nevertheless, as with history treatment teachers, the degree and type of biology treatment teacher support for student engagement with course texts, the explicit teaching and modeling and guided practice using specific comprehension routines, and teachers' attention to equitable participation and support for various students in these classrooms differ significantly. As in history, the magnitude of these differences are quite large.

Table 20. Teacher interviews: Differences by treatment/control group – Biology Sample

	Treatment	Control	Difference	p-val	Diff/SD
Reading opportunities	2.52	2.28	0.23	0.19	0.43
Support for student reading	2.69	2.08	0.61***	0.00	0.86
Metacognitive inquiry	2.05	1.37	0.68***	0.00	1.40
Reading comprehension strategies	2.55	1.76	0.78***	0.00	1.18
Equity	2.69	2.43	0.26	0.19	0.32
Inquiry	0.15	0.19	-0.04	0.74	-0.09

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test

Classroom observations.

A separate report details the results of the classroom observations.

In summary, several sources of data indicate that the intervention teachers were more knowledgeable about and more able to integrate the teaching of US history or science reading with curriculum content, to create classrooms characterized by collaborative inquiry and meaning making with course texts, to engage students in the work of text inquiry, and to offer their students tools in the form of comprehension routines and strategies to support their work with disciplinary texts. These outcomes support H1 of the study: Teachers participating in the

Reading Apprenticeship professional development program will exhibit greater increases in knowledge and skills regarding the integration of literacy and biology or U.S. history, and will demonstrate greater integration of literacy into their instructional practice than teachers in control classrooms.

H2: Student Outcomes: Content Area Understanding, Reading Proficiency, and Engagement in Content Learning

Student Opportunity To Learn Survey: To investigate treatment/control group differences on the OTL survey outcomes, we estimated multi-level regression models that included controls for baseline characteristics (randomization strata, race/ethnicity, and whether or not the student reported speaking a non-English language at home).

History

The student OTL survey results presented in **Table 21a** did not corroborate findings from the teacher-reported measures. No statistically significant differences were apparent across treatment and control groups on the OTL measures.

Table 21a. Student Opportunity to Learn Surveys by Treatment/Control Group – History Sample

	Treatment	Control	Difference	p-val	Diff/SD
Reading in History	3.11	3.19	-0.08	0.24	-0.16
Integration of History & Literacy	2.91	2.80	0.11	0.25	0.16
Motivation/Effort in class	2.95	3.02	-0.07	0.34	-0.11
Academic Identity	2.89	2.97	-0.08	0.24	-0.12

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test

We also examined differences in impacts across self-reported racial/ethnic groups and by whether or not the student reported speaking a language other than English at home (see **Table 21b**). For the most part, no differences in impacts across racial/ethnic groups were detected, and differences across racial/ethnic groups were not statistically significant. For home language, however, there was limited evidence of differences in impacts across groups. The program impact on *Integration of History and Literacy* was positive for students whose home language was not English (es = .29, p<.10), suggesting that these students perceived literacy and history instruction to be integrated to a greater degree than did similar students in control classes. However, this impact was not statistically significant at conventional levels.

Table 21b. Student Opportunity to Learn Surveys by Treatment/Control Group, Student Ethnicity, and Primary Home Language – History Sample

	Treatment	Control	Difference	p-val	Diff/SD
Ethnicity					
Reading in History					
African American	3.12	3.17	-0.04#	0.80	-0.09
Latino	3.09	3.21	-0.12	0.16	-0.23
Asian	3.19	2.56	0.63**	0.01	1.26
Other	3.13	3.22	-0.09	0.32	-0.18
White	3.11	3.17	-0.06	0.46	-0.13
Integration of History & Literacy					
African American	2.72	2.84	-0.12	0.60	-0.18
Latino	2.91	2.79	0.11	0.30	0.17
Asian	2.96	2.48	0.48	0.15	0.73
Other	2.99	2.81	0.18	0.14	0.27
White	2.85	2.79	0.06	0.60	0.09
Motivation/Effort in class					
African American	2.80	2.95	-0.15	0.50	-0.24
Latino	2.94	3.05	-0.10	0.27	-0.16
Asian	3.02	2.75	0.28	0.41	0.45
Other	3.02	3.04	-0.02	0.85	-0.03
White	2.91	2.97	-0.06	0.51	-0.10
Academic Identity					
African American	2.81	3.05	-0.24	0.29	-0.36
Latino	2.94	3.05	-0.11	0.22	-0.17
Asian	3.09	2.63	0.46	0.18	0.69
Other	2.99	2.96	0.03	0.79	0.04
White	2.75	2.88	-0.13	0.17	-0.20
Home Language					
Reading in History					
Non-English	3.16	3.18	-0.02	0.85	-0.03
English	3.09	3.21	-0.12	0.11	-0.24
Integration of History & Literacy					
Non-English	2.93	2.74	0.19*	0.08	0.29
English	2.90	2.84	0.06	0.55	0.09
Motivation/Effort in class					
Non-English	2.92	3.02	-0.11	0.24	-0.17
English	2.97	3.01	-0.04	0.57	-0.07
Academic Identity					
Non-English	2.92	2.98	-0.06	0.50	-0.09
English	2.88	2.97	-0.09	0.22	-0.14

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test
# Estimated impacts are significantly different across groups at the .10 level, two-tailed test.
## Estimated impacts are significantly different across groups at the .05 level, two-tailed test.
### Estimated impacts are significantly different across groups at the .01 level, two-tailed test

Biology

For the biology sample, the results presented in **Table 22a** partially corroborated the findings from the Teacher Survey and Teacher Assignment ratings related to integration of biology and literacy. The results favored the treatment group and was statistically significant for one of the four measures: *Student Integration of Biology & Literacy*—a measure of the degree to which students perceived that teachers integrated the learning of content and literacy practices through comprehension supporting routines and strategies.

Table 22a. Student Opportunity to Learn Surveys by Treatment/Control Group – Biology Sample

	Treatment	Control	Difference	p-val	Diff/SD
Reading in Biology	2.99	2.91	0.08	0.27	0.15
Integration of Biology & Literacy	2.75	2.51	0.23***	0.00	0.37
Motivation/Effort in class	2.85	2.75	0.11	0.13	0.16
Academic Identity	2.77	2.70	0.07	0.37	0.10

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test

As with the History sample, for biology classrooms we also examined differences in impacts across self-reported racial/ethnic groups and by whether or not the student reported speaking a language other than English at home. Patterns of impact across racial/ethnic groups was suggestive that there was a greater impact for Latino and White students in their perception of *Integration of Biology & Literacy* in class instruction in treatment vs. control classes (es = .49 and .51, respectively). White students in intervention classes also experienced greater *Motivation/Effort in class* (es = .36) and increased *Academic Identity* (es = .41) compared to their counterparts in control classes. No differences across racial/ethnic groups were found to be statistically significant. For home language there was limited evidence of differences in impacts across groups. The program impact on students' perceptions of the degree of *Integration of Biology & Literacy* instruction was positive and statistically significant for students whose home language was not English (es = .39) as well as for students whose home language was English (es = .37).

Table 22b. Student Opportunity to Learn Surveys by Treatment/Control Group, Student Ethnicity, and Primary Home Language – Biology Sample

	Treatment	Control	Difference	p-val	Diff/SD
Ethnicity					
Reading in Biology					
African American	2.85	2.80	0.05	0.79	0.09
Latino	3.02	2.91	0.11	0.22	0.20
Asian	2.99	2.93	0.06	0.70	0.11
Other	2.95	2.95	0.01	0.96	0.01
White	3.03	2.89	0.14	0.20	0.25
Integration of Biology and Literacy					
African American	2.51	2.61	-0.10	0.63	-0.15
Latino	2.80	2.48	0.31***	0.00	0.49
Asian	2.57	2.64	-0.07	0.71	-0.11
Other	2.71	2.54	0.17	0.12	0.26
White	2.81	2.48	0.33***	0.01	0.51
Motivation/Effort in class					
African American	3.04	2.74	0.30	0.16	0.47
Latino	2.81	2.73	0.08	0.36	0.12
Asian	2.64	2.88	-0.24	0.22	-0.37
Other	2.84	2.75	0.09	0.40	0.14
White	2.98	2.74	0.23**	0.05	0.36
Academic Identity					
African American	2.82	2.89	-0.07	0.75	-0.11
Latino	2.76	2.74	0.02	0.86	0.02
Asian	2.73	2.75	-0.02	0.90	-0.04
Other	2.74	2.69	0.05	0.63	0.08
White	2.87	2.61	0.26**	0.03	0.41
Home Language					
Reading in Biology					
Non-English	3.00	2.88	0.12	0.28	0.22
English	2.99	2.92	0.08	0.32	0.14
Integration of Biology and Literacy					
Non-English	2.73	2.49	0.25***	0.04	0.39
English	2.75	2.52	0.23***	0.00	0.37
Motivation/Effort in class					
Non-English	2.95	2.78	0.17	0.17	0.26
English	2.84	2.74	0.09	0.19	0.15
Academic Identity					
Non-English	2.80	2.63	0.17	0.16	0.27
English	2.76	2.72	0.05	0.54	0.07

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test
# Estimated impacts are significantly different across groups at the .10 level, two-tailed test.
## Estimated impacts are significantly different across groups at the .05 level, two-tailed test.

### Estimated impacts are significantly different across groups at the .01 level, two-tailed test

Integrated Learning Assessment:

History

There were significant differences in the *History Content Knowledge* (es = .29) and *Reading Strategies* (es = .76) scores between treatment and control (see **Table 23a**). The History Content Knowledge score was intended to serve as a prior knowledge measure and to capture how much students knew about the African-American experience related to WWII before completing the ILA. The Reading Strategies score was designed as a process measure, and the effect size of .76 offers evidence that compared to control students, students in treatment classes approached their reading with an array of reading comprehension strategies. As part of Reading Apprenticeship instruction, students are taught to utilize robust reading strategies in order to improve both reading comprehension and content understanding. The results indicate that the history students in treatment classes implemented these strategies to a greater extent than those in control classes.

Table 23a. Student Integrated Learning Assessment – History Sample

	Treatment	Control	Difference	p-val	Diff/SD
Content knowledge	5.97	5.38	0.60**	0.02	0.29
Reading comprehension	6.89	6.93	-0.04	0.88	-0.02
Metacognition	2.32	2.22	0.11	0.22	0.16
Reading strategies	1.71	1.26	0.44**	0.05	0.76
Writing content	1.91	1.82	0.10	0.50	0.14
Writing language	2.07	1.98	0.09	0.53	0.11

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test

We also examined differences in student performance on the History ILA for impacts across self-reported racial/ethnic groups and by whether or not the student reported speaking a language other than English at home. No discernable pattern of differences in impacts across racial/ethnic groups was detected, and differences across racial/ethnic groups were not statistically significant. Asian students in treatment classes had greater *Reading comprehension* scores on this measure than their counterparts in control classes (es = 1.14, p <.10) but this difference was not statistically significant at conventional levels. African American (es = .99, p < .10), Latino (es = .73, p<.10), and White (es = .82) students in intervention classes demonstrated more frequent use of comprehension-supporting *Reading strategies.* Treatment students whose home language was not English (es = 0.65, p < .10) and students whose home language was English (es = 0.82) showed greater evidence of the use of comprehension-supporting *Reading strategies* than their counterparts in control classes.

Table 23b. Student Integrated Learning Assessment by Treatment/Control Group, Student Ethnicity, and Primary Home Language – History Sample

	Treatment	Control	Difference	p-val	Diff/SD
Ethnicity					
Content knowledge					
African American	5.84	5.14	0.70	0.31	0.34
Latino	5.69	5.20	0.49	0.12	0.24
Asian	5.97	3.64	2.33*	0.09	1.14
Other	5.86	5.46	0.40	0.25	0.19
White	6.41	5.60	0.81**	0.02	0.39
Reading comprehension					
African American	6.23	6.51	-0.28	0.73	-0.12
Latino	6.32	6.42	-0.10	0.78	-0.04
Asian	7.01	4.64	2.37*	0.08	1.03
Other	7.02	7.14	-0.12	0.76	-0.05
White	7.45	7.42	0.03	0.94	0.01
Metacognition					
African American	2.57	2.47	0.10	0.71	0.15
Latino	2.33	2.24	0.09	0.41	0.14
Asian	2.36	2.10	0.26	0.55	0.39
Other	2.30	2.17	0.13	0.29	0.20
White	2.30	2.20	0.09	0.43	0.14
Reading strategies					
African American	1.71	1.13	0.58*	0.08	0.99
Latino	1.67	1.24	0.43*	0.07	0.73
Asian	1.80	1.24	0.56	0.19	0.96
Other	1.74	1.35	0.39*	0.10	0.67
White	1.72	1.24	0.48**	0.04	0.82
Writing content					
African American	2.21	2.21	0.00	0.99	0.00
Latino	1.83	1.63	0.20	0.28	0.28
Asian	1.69	2.07	-0.38	0.61	-0.53
Other	2.04	1.81	0.22	0.25	0.31
White	1.86	1.95	-0.09	0.63	-0.12
Writing language					
African American	2.37	2.47	-0.10	0.82	-0.13
Latino	1.93	1.80	0.13	0.48	0.17
Asian	1.88	1.54	0.35	0.66	0.44
Other	2.17	1.96	0.21	0.29	0.26
White	2.09	2.13	-0.04	0.83	-0.05
Home Language					
Content knowledge					
Non-English	5.74	5.37	0.38	0.25	0.18
English	6.10	5.39	0.71**	0.01	0.35
Reading comprehension					
Non-English	6.67	7.06	-0.39	0.28	-0.17
English	7.02	6.87	0.15	0.62	0.07

Table 23b. Student Integrated Learning Assessment by Treatment/Control Group, Student Ethnicity, and Primary Home Language – History Sample

	Treatment	Control	Difference	p-val	Diff/SD
Metacognition					
Non-English	2.29	2.25	0.03	0.76	0.05
English	2.35	2.20	0.14	0.13	0.22
Reading strategies					
Non-English	1.68	1.29	0.38*	0.10	0.65
English	1.73	1.25	0.48**	0.03	0.82
Writing content					
Non-English	1.87	1.88	0.00	0.98	-0.01
English	1.94	1.78	0.16	0.32	0.22
Writing language					
Non-English	2.02	2.07	-0.05	0.77	-0.07
English	2.10	1.93	0.16	0.28	0.21

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test
# Estimated impacts are significantly different across groups at the .10 level, two-tailed test.
## Estimated impacts are significantly different across groups at the .05 level, two-tailed test.
### Estimated impacts are significantly different across groups at the .01 level, two-tailed test

We were also interested in identifying and recording frequencies for types of reading strategies used. Specifically, we focused on identifying annotations that were indicative of discipline-specific reading strategies since these types of strategies may be most useful when reading the history texts in the ILA and completing the tasks that follow. The discipline-specific strategies were counted as present when it was possible to identify them from the text annotations alone.

Students in treatment classrooms in comparison to students in control classrooms more frequently connected to prior knowledge, conducted intertextual reading, identified bias or point of view, placed the document into a historical context, and identified cause and effect (**Table 24**).

Table 24. Frequency of Evidence of Discipline-Specific Reading Strategies- History

	Treatment and Control (N=273)			
	Treatment (N=176)		Control (N=97)	
	N	%	N	%
Connecting to prior knowledge	42	23.9	12	12.4
Conducting intertextual readings	4	2.3	1	1.0
Evaluating the source of the document	6	3.4	5	5.2
Identifying bias or point-of-view	15	8.5	0	0.0
Placing a document into a historical context	9	5.1	1	1.0
Identifying cause and effect	13	7.4	5	5.2

Biology

Unlike the case for the history sample, students in treatment schools scored lower on *Biology Content Knowledge* than their counterparts in control schools (**Table 25a**), but scored higher on *Metacognition* (es = .27). Metacognition is a measure of students' awareness and control of their reading and problem solving processes. Since this is an intentional target of the Reading Apprenticeship instructional model, this difference offers some evidence that students in intervention classes were approaching their reading of science differently than those in control classes, and this difference is in the expected direction.

Table 25a. Student Integrated Learning Assessment – Biology Sample

	Treatment	Control	Difference	p-val	Diff/SD
Content knowledge	3.84	4.29	-0.45*	0.06	-0.26
Reading comprehension	5.14	5.34	-0.20	0.55	-0.09
Metacognition	2.75	2.52	0.23**	0.04	0.27
Reading strategies	2.46	2.15	0.31	0.34	0.30
Writing content	1.75	1.76	-0.01	0.93	-0.01
Writing language	1.98	1.93	0.04	0.78	0.05

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test

We also examined differences in student performance on the Biology ILA for impacts across self-reported racial/ethnic groups and by whether or not the student reported speaking a language other than English at home. For the most part, no discernable pattern of differences in impacts across racial/ethnic groups was detected, and differences across racial/ethnic groups were not statistically significant. However, Latino students in treatment classes had greater *Metacognition* scores on this measure than their counterparts in control classes (es = .33). In addition, White students in intervention classes demonstrated more frequent use of comprehension-supporting *Reading strategies* (es = .86). Treatment students whose home language was not English (es = .38, p < .10) and students whose home language was English (es = .25, p < .10) both demonstrated greater evidence of *Metacognition* than their counterparts in control classes, but these increases were not statistically significant at conventional levels.

Table 25b. Student Integrated Learning Assessment by Treatment/Control Group, Student Ethnicity, and Primary Home Language – Biology Sample

	Treatment	Control	Difference	p-val	Diff/SD
Ethnicity					
Content knowledge					
African American	3.71	4.10	-0.39	0.48	-0.22
Latino	3.54	3.99	-0.45*	0.10	-0.26
Asian	4.14	4.68	-0.54	0.28	-0.31
Other	3.99	4.40	-0.41	0.17	-0.24
White	4.08	4.58	-0.50	0.16	-0.29
Reading comprehension					
African American	5.34	4.55	0.79	0.27	0.37
Latino	4.83	4.94	-0.11	0.78	-0.05
Asian	5.65	6.17	-0.52	0.43	-0.24
Other	5.28	5.43	-0.15	0.70	-0.07
White	5.05	5.72	-0.67	0.16	-0.31
Metacognition					
African American	2.19	2.33	-0.14	0.64	-0.16
Latino	2.75	2.47	0.28**	0.04	0.33
Asian	3.05	2.74	0.30	0.25	0.36
Other	2.74	2.46	0.28*	0.06	0.33
White	2.79	2.66	0.13	0.46	0.16
Reading strategies					
African American	2.13	2.21	-0.08	0.90	-0.09
Latino	2.54	2.28	0.26	0.38	0.28
Asian	2.19	2.16	0.03	0.96	0.03
Other	2.29	2.27	0.02	0.96	0.02
White	2.80	2.02	0.79**	0.01	0.86
Writing content					
African American	1.40	1.51	-0.11	0.82	-0.11
Latino	1.52	1.59	-0.07	0.73	-0.07
Asian	2.09	2.03	0.06	0.89	0.06
Other	1.81	1.75	0.06	0.80	0.06
White	2.04	2.04	0.00	1.00	0.00
Writing language					
African American	1.60	1.64	-0.04	0.93	-0.04
Latino	1.85	1.76	0.09	0.66	0.09
Asian	2.19	2.21	-0.02	0.95	-0.02
Other	2.01	1.99	0.02	0.93	0.02
White	2.17	2.13	0.03	0.89	0.04

→

Table 25b. Student Integrated Learning Assessment by Treatment/Control Group, Student Ethnicity, and Primary Home Language – Biology Sample

	Treatment	Control	Difference	p-val	Diff/SD
Home Language					
Content knowledge					
Non-English	3.63	4.18	-0.55	0.11	-0.32
English	3.88	4.31	-0.43*	0.07	-0.25
Reading comprehension					
Non-English	5.32	5.32	0.00	0.99	0.00
English	5.11	5.35	-0.23	0.48	-0.11
Metacognition					
Non-English	2.75	2.54	0.32*	0.07	0.38
English	2.76	2.43	0.21*	0.06	0.25
Reading strategies					
Non-English	2.31	2.33	-0.03	0.94	-0.03
English	2.47	2.12	0.36	0.13	0.39
Writing content					
Non-English	1.67	1.66	0.01	0.98	0.01
English	1.77	1.79	-0.02	0.91	-0.02
Writing language					
Non-English	1.93	1.77	0.16	0.52	0.17
English	1.99	1.97	0.02	0.91	0.02

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test
# Estimated impacts are significantly different across groups at the .10 level, two-tailed test.
## Estimated impacts are significantly different across groups at the .05 level, two-tailed test.
### Estimated impacts are significantly different across groups at the .01 level, two-tailed test

Further analyses of recorded frequencies for types of reading strategies indicated that students in treatment classrooms more frequently made connections between the text and their prior biology knowledge while control students more frequently considered science implications beyond the scope of the document sections (**Table 26**).

Table 26. Frequency of Discipline-specific Reading Strategies by Status – Biology Sample

	Treatment and Control (N=268)			
	Treatment (N=161)		Control (N=107)	
	N	%	N	%
Connect to prior knowledge	13	8.1	2	1.9
Questioning scientific methods	3	1.9	2	1.9
Attending to and evaluating evidence	1	0.6	0	0.0
Analyzing graphs, diagrams, etc.	1	0.6	0	0.0
Considering science implications beyond the text's scope	10	6.2	11	10.3

A separate report describes the outcomes for this measure in more detail. Of note is the fact that students in treatment classrooms annotated the texts in the ILA far more frequently than those in control classes, and that text annotation was highly and positively correlated with comprehension of these texts.

Degrees of Reading Power test of reading comprehension

As an additional reading comprehension assessment and to validate the ILA, we administered the Degrees of Reading Power test to assess the level of complexity of text that students are able to read. This test was given on the second day of the ILA administration to half of the students in each class, while the other half wrote essays based on the previous day's readings. Analyses of program impacts on DRP test scores revealed no differences between treatment and control schools on DRP scores, either for the history sample (**Table 27a**) or the biology sample (**Table 28a**).

State Standardized Test Scores

To examine potential program impacts on student performance in history, biology, and reading comprehension, we examined treatment/control differences state mandated criterion-referenced test scores. As described above, two types of test score data were collected - linked, longitudinal test score data for students for we had obtained parental consent and anonymous, unlinked, cross-sectional data student for students for whom we did not obtain parental consent. To account for treatment/control group non-equivalence in the sample retained, all analyses include controls for student and teacher characteristics measured prior to the intervention. **Tables 27** and **28** show the results based on both sets of test score data for the history and biology samples, respectively, for the subject area samples as a whole, as well as by demographic groups.

History

For the longitudinal test data (**Table 27a**), history students in treatment schools exhibited higher scores in *Reading comprehension* and *History*, with effect sizes of 0.16 and 0.19, respectively. These students also scored higher on English language arts (p<.13), although the results for English Language Arts were not statistically significant at conventional levels. For the cross-sectional data, which was a more representative sample of the students in the study, students in the treatment schools performed better than their counterparts in control schools in *English language arts*, *Reading Comprehension*, and *History*, with effect sizes ranging from 0.22 to 0.26.

Table 27a. DRP and CST Test Scores– History Sample

	Treatment	Control	Difference	p-val	Diff/SD
Degrees of Reading Power	37.07	38.97	-1.89	0.53	-0.12
Longitudinal Sample					
ELA CST	0.00	-0.08	0.08	0.13	0.08
Reading Comprehension CST	0.01	-0.16	0.17*	0.05	0.16
History CST	349.56	338.69	10.87**	0.02	0.19
Cross-sectional Sample					
ELA CST	0.18	-0.07	0.25**	0.02	0.26
Reading Comprehension CST	0.14	-0.09	0.23**	0.04	0.22
History CST	342.59	327.34	15.26***	0.01	0.25

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test

Demographic Subgroups

We also examined program impacts by student racial/ethnic status, English learner status, and gender. **Table 27b** shows impacts by subgroup for the longitudinal and cross-sectional samples. For the longitudinal test data, students in treatment schools, regardless of subgroup, exhibited similar levels of performance on the state standardized assessments as their counterparts in control schools. The two exceptions are the statistically significant positive impacts on *English language arts* for Asian students (es = 0.41)) and Latino students on (es = 0.28).

For the cross-sectional test data in **Table 27b,** an analysis of scores by demographic group found increases across all demographic groups in *English language arts* for students in intervention classes, with effect sizes of 0.32 for African American students, 0.22 for Latino students, 0.50 for Asian students, and 0.25 for White students, though these comparisons were only statistically significant at conventional levels for African American, Asian, and White students. Latino and Asian students in treatment classes also show increased test scores compared to their counterparts in control classes on *Reading comprehension* and *History* CSTs. Overall, the pattern of results based on the cross-sectional data suggest that the impacts are most consistent and robust for English speaking students than for students who speak languages other than English at home. It is important to recognize that with so many statistical tests, even though these subgroup analyses were planned comparisons, it is possible that these results are due to chance factors alone.

Table 27b. DRP and CST Test Scores by Treatment/Control Group, Student Ethnicity, and Primary Home Language – History Sample

	Treatment	Control	Difference	p-val	Diff/SD
Degrees of Reading Power by Ethnicity					
African American	22.70	31.16	-8.46	0.32	-0.54
Latino	34.34	36.63	-2.29	0.54	-0.15
Asian	42.55	34.48	8.07	0.60	0.51
Other	36.70	41.28	-4.58	0.26	-0.29
White	42.17	40.08	2.09	0.61	0.13
Degrees of Reading Power by Home Language					
Non-English	35.87	37.85	-1.98	0.59	-0.13
English	37.72	39.56	-1.85	0.57	-0.12
Longitudinal Sample by Ethnicity					
ELA CST					
African American	-0.11	-0.20	0.10	0.63	0.09
Latino	-0.12	-0.08	-0.04	0.59	-0.04
Asian	0.11	-0.31	0.42**	0.05	0.41
Other	0.03	-0.13	0.17	0.26	0.17
White	0.11	-0.05	0.15*	0.05	0.15
Reading Comprehension CST					
African American	0.00	-0.16	0.16	0.51	0.14
Latino	0.00	-0.07	0.07	0.19	0.06
Asian	0.00	-0.31	0.32	0.27	0.29
Other	0.00	-0.15	0.15	0.14	0.13
White	0.43	-0.43	0.86*	0.06	0.78
History CST					
African American	335.93	330.85	5.08	0.71	0.09
Latino	350.09	334.51	15.58***	0.01	0.28
Asian	359.50	339.55	19.95	0.16	0.35
Other	359.22	338.71	20.51*	0.05	0.36
White	344.96	341.48	3.48	0.57	0.06
Longitudinal Sample by Home Language					
ELA CST					
Non-English	0.04	-0.04	0.08	0.19	0.08
English	-0.13	-0.22	0.10	0.34	0.09
Reading Comprehension CST					
Non-English	-0.11	-0.32	0.21	0.16	0.19
English	0.04	-0.12	0.16*	0.08	0.15
History CST					
Non-English	349.87	343.04	6.83##	0.14	0.12
English	347.14	321.68	25.46***	0.00	0.45

→

Table 27b. DRP and CST Test Scores by Treatment/Control Group, Student Ethnicity, and Primary Home Language – History Sample

	Treatment	Control	Difference	p-val	Diff/SD
Cross-sectional Sample by Ethnicity					
ELA CST					
African American	-0.02	-0.32	0.30**#	0.03	0.32
Latino	-0.02	-0.22	0.21*	0.06	0.22
Asian	0.43	-0.05	0.48***	0.00	0.50
Other	0.26	0.00	0.26*	0.08	0.27
White	0.30	0.06	0.24**	0.03	0.25
Reading Comprehension CST					
African American	0.10	-0.16	0.26*	0.07	0.25
Latino	0.03	-0.26	0.29***	0.01	0.28
Asian	0.33	-0.10	0.43***	0.01	0.42
Other	0.23	0.03	0.20	0.20	0.19
White	0.22	0.12	0.10	0.33	0.10
History CST					
African American	331.07	314.95	16.12	0.11	0.26
Latino	332.04	318.70	13.35*	0.08	0.22
Asian	355.99	326.44	29.55***	0.00	0.48
Other	352.64	338.95	13.69	0.19	0.22
White	349.70	336.04	13.66*	0.07	0.22
Cross-sectional Sample by Home Language					
ELA CST					
Non-English	-0.16	-0.32	0.16##	0.15	0.17
English	0.29	0.00	0.29***	0.01	0.31
Reading Comprehension CST					
Non-English	0.20	-0.01	0.22**	0.02	0.21
English	-0.04	-0.30	0.26**	0.01	0.25
History CST					
Non-English	327.92	316.19	11.73	0.13	0.19
English	347.21	330.74	16.47**	0.02	0.27

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test
# Estimated impacts are significantly different across groups at the .10 level, two-tailed test.
## Estimated impacts are significantly different across groups at the .05 level, two-tailed test.
### Estimated impacts are significantly different across groups at the .01 level, two-tailed test

Biology

For the biology-sample longitudinal test data (**Table 28a**), students in treatment schools exhibited similar levels of performance on the state standardized test scores as their counterparts in control schools. For the cross-sectional data, students in the treatment schools performed better in *Biology* (p < .07) than their counterparts in control schools, although the estimated impacts are not statistically significant at conventional levels. The estimated effect size was 0.29.

Table 28a. DRP and CST Test Scores– Biology Sample

	Treatment	Control	Difference	p-val	Diff/SD
Degrees of Reading Power	41.61	42.35	-0.73	0.83	-0.04
Longitudinal Sample					
ELA CST	-0.03	-0.04	0.01	0.92	0.01
Reading Comprehension CST	0.11	-0.02	0.13	0.13	0.13
Biology CST	355.58	355.87	-0.28	0.96	-0.01
Cross-sectional Sample					
ELA CST	0.25	0.08	0.17	0.13	0.18
Reading Comprehension CST	0.17	0.08	0.09	0.39	0.09
Biology CST	355.29	339.88	15.41*	0.07	0.29

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test

Demographic Subgroups

As with the History sample, we also examined program impacts by student racial/ethnic status, English learner status, and gender for biology. **Table 28b** shows impacts by subgroup for the longitudinal and cross-sectional samples. For the longitudinal test data, students in treatment schools, regardless of subgroup, exhibited similar levels of performance on the state standardized assessments as their counterparts in control schools. Overall, the pattern of results for the cross-sectional test data suggests that the impacts are most consistent and robust for Latino students (es = 0.26 for *English language arts* and 0.35 for *Biology*) and for students who speak languages other than English at home (es = 0.29, 0.28, and 0.36 for *English language arts*, *Reading Comprehension*, and *Biology*, respectively). Again, it is important to recognize that with so many statistical tests, even though these subgroup analyses were planned comparisons, it is possible that these results are due to chance factors alone.

Table 28b. DRP and CST Test Scores by Treatment/Control Group, Student Ethnicity, and Primary Home Language – Biology Sample

	Treatment	Control	Difference	p-val	Diff/SD
Degrees of Reading Power by Ethnicity					
African American	43.65	36.85	6.79	0.32	0.40
Latino	37.97	38.91	-0.94	0.81	-0.06
Asian	36.13	43.23	-7.10	0.33	-0.42
Other	42.21	43.61	-1.40	0.73	-0.08
White	48.52	47.80	0.73	0.89	0.04
Degrees of Reading Power by Home Language					
Non-English	37.28	40.52	-3.24	0.51	-0.19
English	42.43	42.71	-0.28	0.94	-0.02
Longitudinal Sample by Ethnicity					
ELA CST					
African American	-0.12	-0.42	0.30	0.15	0.30
Latino	-0.04	-0.12	0.08	0.42	0.08
Asian	-0.13	0.05	-0.19	0.31	-0.18
Other	0.12	-0.13	0.24	0.32	0.24
White	-0.06	0.06	-0.12	0.33	-0.11
Reading Comprehension CST					
African American	0.00	-0.35	0.35	0.16	0.34
Latino	0.10	-0.06	0.16	0.19	0.16
Asian	0.10	0.05	0.05	0.81	0.05
Other	0.13	0.03	0.10	0.46	0.09
White	0.46	0.22	0.24	0.79	0.24
Biology CST					
African American	352.44	344.52	7.91##	0.51	0.14
Latino	354.75	346.74	8.01	0.21	0.14
Asian	353.77	373.62	-19.85**	0.05	-0.36
Other	358.94	347.53	11.41	0.39	0.20
White	350.90	361.07	-10.17	0.15	-0.18
Longitudinal Sample by Home Language					
ELA CST					
Non-English	0.01	-0.01	0.02	0.86	0.02
English	-0.10	-0.13	0.03	0.81	0.03
Reading Comprehension CST					
Non-English	0.13	-0.08	0.21	0.12	0.20
English	0.12	0.01	0.11	0.33	0.10
Biology CST					
Non-English	356.47	357.41	-0.94	0.87	-0.02
English	352.61	352.78	-0.18	0.98	0.00

→

Table 28b. DRP and CST Test Scores by Treatment/Control Group, Student Ethnicity, and Primary Home Language – Biology Sample

	Treatment	Control	Difference	p-val	Diff/SD
Cross-sectional Sample by Ethnicity					
ELA CST					
African American	-0.05	-0.15	0.10	0.52	0.11
Latino	0.18	-0.07	0.25**	0.04	0.26
Asian	0.53	0.37	0.16	0.24	0.17
Other	0.10	0.13	-0.03	0.86	-0.03
White	0.30	0.17	0.13	0.30	0.13
Reading Comprehension CST					
African American	-0.21	-0.13	-0.08	0.61	-0.08
Latino	0.10	-0.05	0.15	0.18	0.15
Asian	0.42	0.33	0.09	0.54	0.09
Other	0.05	0.17	-0.11	0.48	-0.12
White	0.24	0.16	0.08	0.48	0.08
Biology CST					
African American	339.54	331.65	7.88	0.46	0.15
Latino	348.97	330.51	18.46**	0.04	0.35
Asian	371.63	358.19	13.45	0.17	0.25
Other	350.89	342.43	8.46	0.44	0.16
White	359.65	344.54	15.11*	0.10	0.28
Cross-sectional Sample by Home Language					
ELA CST					
Non-English	0.13	-0.15	0.28**##	0.03	0.29
English	0.31	0.18	0.13	0.27	0.13
Reading Comprehension CST					
Non-English	0.14	-0.15	0.28**#	0.02	0.28
English	0.19	0.18	0.01	0.90	0.01
Biology CST					
Non-English	350.44	331.18	19.26**	0.03	0.36
English	357.65	343.86	13.79	0.11	0.26

Notes. Data are regression-adjusted using multilevel regression models to account for differences in baseline characteristics and study design characteristics. Effect sizes were calculated by dividing impact estimates by the control group standard deviation of the outcome variable.
* Significantly different from zero at the .10 level, two-tailed test.
** Significantly different from zero at the .05 level, two-tailed test.
*** Significantly different from zero at the .01 level, two-tailed test
# Estimated impacts are significantly different across groups at the .10 level, two-tailed test.
## Estimated impacts are significantly different across groups at the .05 level, two-tailed test.
### Estimated impacts are significantly different across groups at the .01 level, two-tailed test

In summary. Thus, state standardized assessments provide some evidence that the intervention – Reading Apprenticeship in biology and history – is associated with improved performance on state standardized test scores in English language arts, reading comprehension, biology, and history, with effect sizes ranging from 0.14 to 0.29. These effect sizes give an estimate of the magnitude of the difference between student test scores in the intervention and control groups. A year of reading growth at the high school level has been estimated to produce a magnitude of change of approximately .19 (Hill, Bloom, Black, & Lipsey, 2008). This indicates that students in the intervention classes were approximately one year ahead of their counterparts in control classes at the end of the study. Thus, there is some evidence that the intervention—professional development to support implementation of the Reading Apprenticeship instructional framework in high school biology and history classes—is associated with increases in performance on the state standardized assessments examined.

SUMMARY

The study reported here has made significant progress in building tools and processes for linking teacher professional development to meaningful classroom change, and from there, to student engagement and achievement, within a scientifically rigorous experimental study design. Multiple measures of teacher implementation reveal a robust corroboration of teacher level outcomes. Across these measures, teachers in the experimental group demonstrated increased support for science literacy learning, increased use of metacognitive inquiry routines, increased reading comprehension instruction, and increased use of collaborative learning structures. In short, they were more able to integrate science and science literacy learning in classroom instruction than their counterparts in the control group.

Although the results for comprehension measures – the Degrees of Reading Power (DRP) test and comprehension questions from the Interactive Learning Assessments – did not provide evidence that these differences in teaching resulted in learning differences for students, Interactive Learning Assessments show evidence that students in treatment classes approached reading differently than their counterparts in control classes. In particular, they annotated text more often, showing evidence of reading strategies and discipline-specific reasoning, compared to their counterparts in control classes. Study impacts for targeted groups of students (English learners, Latino students, African American students) were found for *Reading strategies* in history treatment classes. In biology treatment classes, only the impact on *Metacognition* was statistically significant for Latino and second language learners in biology. Analyses of student annotations indicated that students in treatment classrooms annotated the texts in the ILA far more frequently than those in control classes, and that text annotation was highly and positively correlated with comprehension of these texts. Student Opportunity to Learn (OTL) surveys partially corroborated teacher reports of increased integration of literacy and content instruction and some sub-groups reported increased levels of motivation and effort in class as well as an increased sense of academic identity. Program impacts were found on *Integration of History and Literacy* for students whose home language was not English.

Further, the results for state-mandated criterion-referenced test scores offer some evidence that differences in teacher practice resulted in improvements in student academic performance. Two types of state standardized test score data were collected — linked, longitudinal test score data for students for whom we had obtained parental consent; and anonymous, unlinked, cross-sectional data for all students, regardless of parental consent status. To enhance the precision of the impact estimates and to account for potential differences in pre-intervention characteristics between groups, the test score analyses controlled for student and teacher characteristics.

For the longitudinal test data, history students in treatment schools exhibited higher levels of performance on state standardized assessments in reading comprehension (es=0.16, p < .06) and history (es=0.19) than students in control schools. Biology students in treatment schools exhibited similar levels of performance on state standardized assessments as their counterparts in control schools, based on analyses of the longitudinal data. For the cross-sectional data, which was a more representative sample of the students in the study, both history and biology students in the treatment schools performed better than their counterparts in control schools on the state standardized assessments. For the history sample, students in treatment schools exhibited higher scores in history (es=0.25), reading comprehension (es=0.22), and English language arts (es=0.26). An analysis of scores by demographic group found increases across all demographic groups in *English language arts* for students in history intervention classes, with effect sizes of 0.32 for African American students, 0.22 for Latino students, 0.50 for Asian students, and 0.25 for White students, though these comparisons were only statistically significant at conventional levels for African American, Asian, and White students. Latino and Asian students in treatment classes also show increased test scores compared to their counterparts in control classes on *Reading comprehension* and *History* CSTs. For the biology sample, students in treatment schools exhibited higher scores in biology (es=0.29, p<.10). Overall, the pattern of results for the cross-sectional test data suggests that the impacts are most consistent and robust for Latino students (es = 0.26 for *English language arts* and 0.35 for *Biology*) and for students who speak languages other than English at home (es = 0.29, 0.28, and 0.36 for *English language arts*, *Reading Comprehension*, and *Biology*, respectively) for the groups targeted by the study.

Thus, state standardized assessments provide some evidence that the intervention – Reading Apprenticeship in biology and history – is associated with improved performance on state standardized test scores in English language arts, reading comprehension, biology, and history, with effect sizes ranging from 0.14 to 0.29. At the high school level, a years' growth is approximated at an effect size of 0.19 (Hill, Bloom, Black & Lipsey, 2008). The results of the study thus present a positive picture with regards to the effectiveness of the Reading Apprenticeship framework for integrating academic literacy content with biology and history coursework and instructional practices.

However, several cautions should be raised in interpreting the results. The study utilized a design in which teachers were recruited and randomly assigned to treatment and control groups fully two years prior to final data collection. Participating teachers in the treatment group had the opportunity to teach students utilizing what they had learned in the professional development for *two* consecutive academic years. The implication of this sequencing is that treatment teachers had one academic year to *practice* using the framework, which helped ensure that program impacts were assessed after teachers had adequate experience with the framework. However, this aspect of the design required that participating teachers be retained in the study for a lengthy period. Retaining teachers in the study for such an extended period was a challenge. Such designs, while acknowledging the importance of practice for teachers, expose studies to greater risk of attrition. Many teachers were reassigned by administrators, dropped out of the study due to changes in districts or schools, or were lost due to changing life circumstances, such as health or even in one case, death.

Of the 219 teachers randomly assigned to experimental condition, 87 (40%) provided survey, interview, or student test score data at the final data collection point. Analyses of characteristics of schools, teachers, and students who were retained in the sample indicated that treatment schools served students with similar characteristics as those served by control schools, suggesting that sample selectivity is unlikely to be responsible for the differences.

In a previous experimental study, professional development in Reading Apprenticeship was shown to improve student test scores in biology, reading comprehension, and English language arts in biology intervention classes, compared to students in control classes. This study found similar outcomes for history and biology, with the exception that the professional development did not have a pronounced impact on reading comprehension in biology classrooms. This differential impact on reading comprehension across studies raises questions for further analysis.

The experimental impact estimates reported here are based on models that control for group differences in pre-intervention student and teacher characteristics. As such, the estimates represent the best estimates of program impacts given data limitations and provide evidence that the Reading Apprenticeship program of professional development can impact teaching and learning outcomes in high school biology and history. At the outset of this study, we posited that professional development would lead to greater teacher knowledge and practice integrating literacy and content teaching, and that these changes would result in increased student engagement and achievement in both literacy and science. This study demonstrates that professional development focused on literacy teaching in an academic content area such as science or history can substantially impact teachers' classroom practices and the resulting opportunities students experience to learn to read and reason with complex materials and texts. Further, these outcomes indicate that focusing on developing teachers' capacity to provide literacy instruction to support active, intellectual inquiry with content texts can support students' achievement in both reading comprehension skills and in content learning. At a time when strategies for improving educational outcomes for underachieving students increasingly focus on making structural changes, increasing accountability, and redistributing effective teachers through incentives for working in the most challenging schools, this is highly significant, demonstrating that it is possible to improve outcomes for students by building existing teachers' capacity through well designed professional development interventions of this kind.

This study indicates the promise of taking a disciplinary approach to literacy instruction, showing through a rigorous, scientific study design that it is possible to improve the instructional quality of content teaching at the high school level through professional development focused on literacy in learning, and that these changes can result in improved engagement and learning for students. Further, at a time when secondary students are increasingly removed from content area learning to remediate their literacy skills, this study makes a contribution of great potential practical import, since not only would integrating literacy and content instruction mitigate these unintended consequences of restricting students' access to vital content area learning, but would result in substantial cost savings to districts and schools. The results of this study indicate that integrated literacy instruction can support, rather than supplant, content learning for students, and conversely, that an instructional focus on developing students' reading proficiencies in specific disciplines like science and history can meaningfully improve students' reading comprehension and literacy, more generally.

REFERENCES

Allington, R. L., & McGill-Franzen, A. (1989). School response to reading failure: Chapter 1 and special education students in grades 2, 4, and 8. *Elementary School Journal, 89*, 529-542.

Alvermann, D., & Moore, D. (1991). Secondary school reading. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (vol. 2, pp. 951 - 983). New York: Longman.

Aschbacher, P. R. (1999). Developing indicators of classroom practice to monitor and support school reform (CSE Tech. Rep. No. 513). Los Angeles: UCLA/CRESST.

Baker, L. (1991). Metacognition, reading, and science education. In C. Santa & D. Alvermann (Eds.) *Science learning: Processes and applications*. Newark, DE: International Reading Association, pp. 2-13.

Ball, D. & Cohen, D. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In L. Darling-Hammond & D. Sykes, (Eds.) *Teaching as the Learning Profession: Handbook of Policy and Practice*. San Francisco: Jossey-Bass, Inc.

Barton, P. (2003). Parsing the achievement gap: Baselines for tracking progress. Policy Information Report of the Educational Testing Service. Princeton, NJ: ETS. Downloaded from the web in January, 2004 at www.ets.org/research/pic

Baumann, J. F., & Duffy, A. M. (1997). Engaged reading for pleasure and learning: A report from the National Reading Research Center. Athens, GA: National Reading Research Center.

Bayer, A. S. (1990). Collaborative apprenticeship learning: Language and thinking across the curriculum, K-12.Mountain View, CA: Mayfield Publishing.

Beck, I. L., McKeown, M. G., Hamilton, R. L., & Kucan, L. (1997). *Questioning the author: An approach for enhancing student engagement with text*. Newark, DE: International Reading Association.

Borasi, R. & Seigel, M. (2000). *Reading counts*. NY: Teachers College Columbia University.

Brown, J. S., Collins, A., & Newman, S. (1989). The new cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning and instruction: Essays in honor of Robert Glaser* (pp. 453 - 494). Hillsdale, NJ: Erlbaum.

Calfee, R. & Miller, R. G. (2004). Separate no more: A theoretical and practical basis for using embedded reading and writing instruction for expository text. Paper distributed at the Annual Meetings of the American Educational Research Association, San Diego.

Cervetti, G., Pearson, P. D., Bravo, M. A., & Barber, J. (2006). Reading and writing in the service of inquiry-based science. In R. Douglas, M. Klentschy, & K. Worth (Eds.), *Linking science and literacy in the K–8 classroom* (pp. 221–244). Arlington, VA: National Science Teachers Association Press.

California Department of Education. (2000). *Construction of California's 1999 School Characteristics Index and Similar Schools Ranks.* California Department of Education, Office of Policy and Evaluation.

Clare, L. (2000). Using teachers' assignments as an indicator of classroom practice (CSE Tech. Rep. No. 532). Los Angeles: UCLA/CRESST.

Cook, T.D. & Campbell, D.T. (1979) Quasi-Experimentation: Design and Analysis Issues in Field Settings. Chicago: Rand-McNally, 1979.

Delpit, L. D. (1995). Other people's children: Cultural conflict in the classroom. New York: New Press.

Donahue, P. L., Voelkl, K. E., Campbell, J. R., & Mazzeo, J. (1999). *The NAEP 1998 reading report card for the nation and the states*. Washington, DC: National Center for Education Statistics.

Duffy, G. G., Roehler, L. R., Meloth, M. S., Vavrus, L. G., Book, C., Putnam, J. & Wesselman, R. (1986). The relationship between explicit verbal explanations during reading skill instruction and student awareness and achievement: A study of reading teacher effects. *Reading Research Quarterly, 21*(3), 237-252).

Duke, N. (2000). 3.6 minutes per day: The scarcity of informational texts in first grade. Reading Research Quarterly, 35 (2), pp. 202 – 224.Durkin, 1984

Fielding, L. G., & Pearson, D. P. (1994). Reading comprehension: What works. *Educational Leadership*, 51, 62-68.

Freedman, S. W., Flower, L., Hull, G., & Hayes, J. R. (1995). *Ten years of research: Achievements of the National Center for the Study of Writing and Literacy* (Technical Report No. 1-C). Berkeley, CA: National Center for the Study of Writing.

Gee, J. (1999). Critical issues: Reading and the new literacy studies: Reframing the National Academy of Science report on reading. *JLR, 31*, (3), 355-374.

Greenleaf, C., Schoenbach, R., Cziko, C., & Mueller, F. (2001). Apprenticing adolescents to academic literacy. *Harvard Educational Review, 71*(1), 79-129. www.wested.org/stratlit/pubsPres/HER/p01green.htm

Greenleaf, C., Brown, W., & Litman, C. (2004). Apprenticing urban youth to science literacy. In D. Strickland & D. Alvermann (Eds.), *Bridging the gap: Improving literacy learning for preadolescent and adolescent learners in grades 4–12*. Newark, NJ: International Reading Association.

Greenleaf, C., Hull, G. & Reilly, B. (1994). Learning from our diverse students: Helping teachers rethink problematic teaching and learning situations. *Teaching & Teacher Education, 10* (5), 521-541.

Greenleaf, C.L. & Katz, M. (2004). Ever newer ways to mean: Authoring pedagogical change in secondary subject-area classrooms. In S.W. Freedman & A. F. Ball (Eds.), *New literacies for new times: Bakhtinian perspectives on language literacy and learning for the 21st century*. Cambridge: Cambridge University Press.

Greenleaf, C. & Schoenbach, R. (2001). Close Readings: A Study of Key Issues in the Use of Literacy Learning Cases for the Professional Development of Secondary Teachers. Final Report to the Spencer and MacArthur Foundations Professional Development Research and Documentation Program.

Greenleaf, C. & Schoenbach, R. (2004). Building capacity for the responsive teaching of reading in the academic disciplines: Strategic inquiry designs for middle and high school teachers' professional development. In Dorothy Strickland and Michael L. Kamil, (Eds.), Improving Reading Achievement through Professional Development, Temple University: Laboratory for Student Success.

Guskey, T. R. & Huberman, M. (1996). *Professional Development in Education: New Paradigms and Practices*. Columbia, NY: Teachers College Press.

Guthrie, J. T., McGough, K., Bennett, L., & Rice, M. E. (1996). Concept-oriented reading instruction: An integrated curriculum to develop motivations and strategies for reading. In L. Baker, P. Afflerbach, & D. Reinking (Eds.), *Developing engaged readers in school and home communities* (pp. 165-190). Mahwah, NJ: Lawrence Erlbaum.

Haycock, K. (2001). Closing the achievement gap. *Educational Leadership*, *58(6),* 6-11. www.ascd.org/readingroom/edlead/0103/haycock.html

Heller, R. & Greenleaf, C. (2007). Literacy instruction in the content areas: Getting to the core of middle and high school improvement. Washington, DC: Alliance for Excellent Education.

Hiebert, E. (1991). *Literacy for a diverse society: Perspectives, policies, and practices*. New York: Teachers College Press.

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

Hillocks, G. Jr. (1995). *Teaching writing as reflective practice.* NY: Teachers College Press.

Hull, G. A., & Rose, M. (1989). Rethinking remediation: Toward a social-cognitive understanding of problematic reading and writing. *Written Communication, 8*, 139-154.

Hynd, C. (1998). *Learning from text across conceptual domains*. Mahwah, NJ: Lawrence Erlbaum.

Jencks, C., & Phillips, M. (Eds.). (1998). *The Black-White test score gap*. Washington, DC: Brookings Institution.

Knapp, M. S., & Turnbull, B. (1991). Better schools for the children in poverty: Alternatives to conventional wisdom. Berkeley, CA: McCutchan.

Lave, J., & Wenger, E. (1991). Situated learning: Legitimate peripheral participation. Cambridge, Eng.: Cambridge University Press.

Lee, C. (1995). A culturally based cognitive apprenticeship: Teaching African American high school students skills in literary interpretation. *Reading Research Quarterly, 30,* 608-630.

Lee, C.D., Spratley, A. (2010). *Reading in the disciplines: The challenges of adolescent literacy*. New York, NY: Carnegie Corporation of New York.

Lemke, J. L. (1990). *Talking science: Language, learning, and values*. Norwood, NJ: Ablex.

Lemke, J. L. “Teaching All the Languages of Science: Words, Symbols, Images, and Actions.” http://www-personal.umich.edu/~jaylemke/papers/barcelon.htm. This is an invited conference address that may be published in whole or in part in *Metatemas*, a journal edited in Barcelona.

Little, J. (2001). Inside Teacher Community: Representations of Classroom Practice. Paper presented as an invited address at the bi-annual conference of the International Study Association on Teachers and Teaching, Faro, Portugal, September 21-25.

Matsumura, L. C. (2003). Teachers, assignments and student work: Opening a window on classroom practice (CSE Tech. Rep. No. 602). Los Angeles: UCLA/CRESST.

McMahon, M., & McCormich, B. (1998). “To think and act like a scientist: Learning disciplinary knowledge. In C. Hynd (Ed.), *Learning from text across conceptual domains*. Mahwah, NJ: Erlbaum.

McMurrer, J. (2007) *Choices, Changes, and Challenges: Curriculum and Instruction in the NCLB Era* Washington, DC: CEP

Miller, R. G. (2004). Creating thoughtful and expressive teachers and students using Read-Write activities: Final Report to NSF: IERI.

Moje, E. B., Ciechanowski, K. M., Kramer, K., Ellis, L., Carrilo, R., & Collazo, T. (2004). Working toward third space in content area literacy : An examination of everyday funds of knowledge and Discourse. *Reading Research Quarterly, 39* (1), pp. 38 – 70.

Moje, E. B., Dillon, D. R., & O'Brien, D. G. (2000). Re-examining the roles of the learner, the text, and the context in secondary literacy. *Journal of Educational Research, 93,* 165-180.

Mullis, I.V., Dossey, A., Campbell, J. R., Gentile, C. A., O'Sullivan, C., & Latham, A. S. (1994). *NAEP 1992 trends in academic progress* (Report No. 23-TR01). Washington, DC: U.S. Government Printing Office.

Murray, D. M. (1998) *Design and Analysis of Group Randomized Trials*. New York: Oxford University Press.

National Assessment of Educational Progress (2006). *The Nation's Report Card: Science 2005*. Washington, DC: NCES Retrieved 7/18/08 http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2006466

National Assessment of Educational Progress (2007). *The Nation's Report Card: Reading 2007.* Washington, DC: NCES Retrieved 7/18/08 http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007496

New London Group. (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review, 66*, 60-92.

Norris, S.P. and Phillips. L.M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, *87*, 224-240.

Osbourne, J. (2002). Science without literacy: A ship without a sail? *Cambridge Journal of Education 32(2),* 203-218.

Pearson, P.D. (1996). Reclaiming the center. In M. Graves, P. van den Broek, & B. M. Taylor (Eds.) *The first R: Every child's right to read* (pp. 259 - 274). NY: Teacher's College Press.

Pressley, M. (1998). *Reading instruction that works: The case for balanced teaching*. New York: Guilford Press.

Richardson, V., (Ed.) (1994). Teacher change and the staff development process. NY: Teachers College Press.

Roth, K. (1991). Reading science texts for conceptual change. In C. Santa & D. Alvermann (Eds.) *Science learning: Processes and applications*. Newark, DE: International Reading Association, pp. 48 - 63.

Rutherford, F. J., & Ahlgren, A. (1990). Science for all Americans. New York: Oxford University Press.

Rycik, J. & Irvin, J. (Eds.). (2001). *What adolescents deserve: A commitment to students' literacy learning*. Newark, DE: International Reading Association.

Schoenbach, R., Greenleaf, C., Cziko, C., & Hurwitz, L. (1999). Reading for understanding: A guide to improving reading in middle and high school classrooms. San Francisco: Jossey-Bass.

Scott, J. (1993). Science and language links: Classroom implications. Portsmouth, NH: Heinemann.

Scribner, S., & Cole, M. (1981). *The psychology of literacy*. Cambridge, MA: Harvard University Press.

Snow, C. (2002). Reading for understanding: Toward a research and development program in reading comprehension. Arlington: RAND. http://www.rand.org/publications/MR/MR1465/

Snow, C., Burns, S., & Griffin, P. (1998). *Preventing reading difficulties in young children.* Washington, D.C.: National Academy Press.

Strickland, D. S. & Kamil, M. L., (Eds.), (2004). *Improving reading achievement through professional development*. Norwood, MA: Christopher-Gordon Publishers, Inc.

Their, M. with Davis, B. (2002). *The new science literacy: Using language skills to help students learn science*. Portsmouth, NH: Heinemann.

Wineburg, S. S. (1991). Historical Problem Solving: A Study of Cognitive Processes Used in the Evaluation of Documentary and Pictorial Evidence. *Journal of Educational Psychology* 83, pp. 73-87.

APPENDICES

Appendix A – Professional Development Overviews for History and Biology

WEEK-AT-A-GLANCE AGENDAS
Literacy in History Summer Opening:
Cohort I 2006, Cohort II 2007

Monday	Tuesday	Wednesday	Thursday	Friday
Introduction to the study **Model Lessons** Personal Science Reading Histories, Capturing Reading Processes and Reading Strategies Lists **Reading Process Analysis** Think Aloud with *Sula*	**Assessing and Re-Teaching** **Model Lesson** Building schema and Talking to the Text: *Internment primary sources text set* **Reading Process Analysis** Text and Task Analysis **Reading Process Analysis** Summarizing	**Assessing and Re-Teaching** **Professional Reading** What do Reading Apprenticeship Classrooms look like? **Classroom Case** What do supports for challenging academic reading look like? Acids and Bases case	**Assessing and Re-Teaching** **Reading Process Analysis** Four Interacting Areas of Reading with Totalitarianism **Literacy Learning Case** Rosa' Question **Teaching Toolbox:** Questioning. CERA	**Professional Reading** **Assessing and Re-Teaching** Got and Need **Supporting Classroom Application** Teachers draft an implementation plan for the first six weeks of school
Literacy Learning Case LaKeisha *Sula* **Reading Process Analysis** Think Aloud with a primary source **Student Literacy Learning Case** LaKeisha Reads a primary source **Teaching Toolbox** Scaffolding metacognitive conversation **Assessing and Re-Teaching** Got and Need	**Classroom Case** Reading Koramatsu *v* United States in Honors History **Teaching Toolbox** Talking to the Text, building on the reading strategies list, Planning responsive instruction, Summarizing. **Assessing and Re-Teaching** Got and Need	**Supporting Classroom Application:** Plan supports for talk and reading **Assessing and Re-Teaching** Got and Need	**Supporting Classroom Application** Collaborative planning and mentoring Prepare a draft scope and sequence plan to share in the morning. **Assessing and Re-Teaching** Got and Need	**Supporting Classroom Application** Small groups meet, share plans, offer receive feedback Leave taking and logistics: Bring back a CERA sample! See you in January! **Assessing and Re-Teaching** Got and Need

Literacy in History Winter follow up: Cohort I 2007, Cohort II 2008

Thursday	Friday
Assessing and Re-teaching Got and Need ***REFLECTING ON PRACTICE*** Assessing practice in the Dimensions of Reading Apprenticeship **Classroom Case** Acids and Bases Epilogue: Designing responsive instruction ***PROFESSIONAL READING*** Building vocabulary and background knowledge	**Assessing and Re-teaching** Got and Need ***REFLECTING ON PRACTICE*** Analyzing student work and setting instructional goals **Professional Reading** Supporting Extensive Reading ***TEACHING TOOLBOX*** Classroom Libraries based on Blueprints and power standards
MODEL LESSON Word learning strategies in history **Assessing and Re-teaching** Got and Need	**Supporting Classroom Application** Planning classroom libraries and next steps in metacognitive conversation **Assessing and Re-teaching** Got and Need

Literacy in History Summer Final Institute
Cohort I 2007, Cohort II 2008

Monday	Tuesday	Wednesday
Reflecting on practice Share lessons in trio's, reflect on Dimensions of Reading Apprenticeship **Reflecting on practice** Analyzing student work with the CERA rubric **Reading Process Analysis** **T**ext and Task Analysis of the CERA texts	**Assessing and Re-teaching** Got and Need **Reading Process Analysis** What kinds of questions do readers of science ask to make sense of science reading? **Model Lesson** Teaching students to be active questioners	**Assessing and Re-teaching** Got and Need **Reading Process Analysis** How do experienced readers of science clarify scientific text/language? **Model Lesson** Teaching word learning and clarifying strategies
Reflecting on practice Analyzing pre- and post instructional samples of student work and using CERA data to identify instructional needs and design responsive instruction. **Assessing and Re-teaching** Got and Need	**Teaching toolbox** ReQuest Question Answer Relationships Connecting to prior knowledge Monitoring conceptual change **Assessing and Re-teaching** Got and Need	**Supporting Classroom Application** Scope and Sequence: power standards as an equity tool Classroom Libraries Collaborative planning and conferring **Assessing and Re-teaching** Got and Need

Literacy in Science, Summer 2007

Monday	Tuesday	Wednesday	Thursday	Friday
Introduction to the study **Model Lesson** Personal Science Reading Histories, Capturing Reading Processes and Reading Strategies Lists **Reading Process Analysis** Think Aloud with Acids and Bases	**Assessing and Re-Teaching** **Model Lesson** Extensive Reading building schema and Talking to the Text. Evolution Text set **Reading Process Analysis** Text and Task Analysis	**Assessing and Re-Teaching** **Professional Reading** What do Reading Apprenticeship Classrooms look like? **Student Literacy Case** The Pedagogy of Equity. What supports do struggling students need?	**Assessing and Re-Teaching** **Reading Process Analysis** Cell Theory text set. Thinking Aloud with visual models **Model lesson** Reading visuals, in science text. **Teaching Toolbox** Chunking, accessing prior knowledge. Repeated Readings	**Professional Reading** **Assessing and Re-Teaching** Got and Need **Supporting Classroom Application** Teachers draft an implementation plan for the first six weeks of school
Model Lesson KWL and Think Aloud: acid /base text set **Classroom video case** Introducing Think Aloud in Intro to Chemistry. **Teaching Toolbox** Scaffolding the metacognitive conversation **Assessing and Re-Teaching** Got and Need	**Student Literacy Case:** "Patterns of Evolution" in Modern Biology **Teaching Toolbox** Talking to the Text, building on the reading strategies list, Planning responsive instruction, Previewing and predicting from text structure. **Assessing and Re-Teaching** Got and Need	**Classroom video Case** What does a Reading Apprenticeship class look like in science? **Teaching Toolbox:** Supports for talk and inquiry **Supporting Classroom Application:** Plan supports for talk and reading **Assessing and Re-Teaching** Got and Need	**Supporting Classroom Application** Collaborative planning and mentoring Prepare a draft scope and sequence plan to share in the morning. **Assessing and Re-Teaching** Got and Need	**Supporting Classroom Application** Small groups meet, share plans, offer receive feedback Leave taking and logistics: Bring back a CERA sample! See you in January! **Assessing and Re-Teaching** Got and Need

Literacy in Science, Winter 2008

Thursday	Friday
Assessing and Re-teaching Got and Need ***REFLECTING ON PRACTICE*** Assessing practice in the Dimensions of Reading Apprenticeship **Classroom Video Case** Acids and Bases Epilogue ***PROFESSIONAL READING*** Building vocabulary and background knowledge	**Assessing and Re-teaching** Got and Need ***REFLECTING ON PRACTICE*** Analyzing student work and setting instructional goals **Professional Reading** Supporting Extensive Reading ***TEACHING TOOLBOX*** Classroom Libraries
TEACHING TOOLBOX Word learning strategies in science ***MODEL LESSON*** Test as genre ***TEACHING TOOLBOX*** Building schema: power standards, testing blueprints, test takers strategy list **Assessing and Re-teaching** Got and Need	**Supporting Classroom Application** Planning classroom libraries and next steps in metacognitive conversation **Assessing and Re-teaching** Got and Need

Literacy in Science, Summer 2008

Monday	Tuesday	Wednesday
Reflecting on practice Share lessons in trio's, reflect on Dimensions of Reading Apprenticeship **Reflecting on practice** Analyzing student work with the CERA rubric **Reading Process Analysis** **T**ext and Task Analysis of the CERA texts	**Assessing and Re-teaching** Got and Need **Reading Process Analysis** What kinds of questions do readers of science ask to make sense of science reading? **Model Lesson** Teaching students to be active questioners	**Assessing and Re-teaching** Got and Need **Reading Process Analysis** How do experienced readers of science clarify scientific text/language? **Model Lesson** Teaching word learning and clarifying strategies
Reflecting on practice Analyzing pre- and post instructional samples of student work and using CERA data to identify instructional needs and design responsive instruction. **Assessing and Re-teaching** Got and Need	**Teaching toolbox** ReQuest Question Answer Relationships Connecting to prior knowledge Monitoring conceptual change **Assessing and Re-teaching** Got and Need	**Supporting Classroom Application** Scope and Sequence: power standards as an equity tool Classroom Libraries Collaborative planning and conferring **Assessing and Re-teaching** Got and Need

Appendix B –Item Map for Student Opportunity to Learn Survey

OTL Constructs – History OTL Survey – Pilot Data

		Alpha
(1) Class Emphasis on Reading in History		
1a	Reading a wide variety of history materials (textbooks, newspapers, … etc.)	
1b	Learning from one another's different ways of reading and thinking about history	.84
1c	Working together to figure out the meaning of the readings.	
1d	Listening and responding to one another's ideas.	
1e	Learning to read, write, listen and talk about history.	
2a	Taught ways to make history reading interesting and motivating for students.	
2b	Taught different strategies to help students understand history reading better ()	
2c	Taught ways to read charts, graphs, and maps.	
2d	Talked about what is going on…teacher's mind … teacher reads history material.	
2f	Encouraged students to use each other's ideas.	
(2) Frequency of Student Integration of History & Literacy Activity		
3a	Spent class time reading.	.82
3b	Worked with partners or groups on reading assignments in class.	
3c	Practiced reading comprehension strategies with history materials.	
3d	Shared difficulties and ways you solved reading comprehension problems.	
3f	Analyzed the way history materials are written and organized (e.g., headings,..).	
(3) Motivation/Effort in Focus Class		
4a	Completed reading assignments.	
4b	Enjoyed completing a reading assignment … that required a lot of thinking...	.81
4c	Put forth a great deal of effort when doing your history reading.	
4e	Tried to really understand history reading assignments in this class.	
4f	Felt motivated to work harder than usual on reading assignments in this class.	
4g	Wanted to do a good job on reading assignments.	
(4) Academic Identity		
4h	Became really interested in the history reading assigned in this class	.96
5a	Understanding yourself better as a reader and learner.	
5b	Making you curious to read about other things in history.	
5c	Seeing yourself as a reader.	
5d	Being a more serious student.	
5e	Thinking about your future educational goals.	
5f	Making you interested in taking more history classes.	
5g	Understanding history materials better when you read.	
5h	Given you more confidence that that you can read and do history.	
5i	Being willing to tackle challenging reading materials.	
5j	Thinking of yourself as a capable student	
5k	Feeling like you can succeed in more challenging classes.	
5l	Seeing your education as important.	
6a	Learning history better.	
6b	Understanding history concepts better.	
6c	Feeling like you can be more successful reading in other history classes.	
6d	Feeling more positive about reading history.	
6e	Having a more positive attitude about reading in general.	