JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 49, NO. 5, PP. 631–658 (2012)

Research Article

The Impact of an Integrated Approach to Science and Literacy in Elementary School Classrooms

Gina N. Cervetti,[1] Jacqueline Barber,[2] Rena Dorph,[2] P. David Pearson,[2] and Pete G. Goldschmidt[3]

[1]*University of Michigan, Ann Arbor, Michigan*
[2]*University of California, Berkeley, California*
[3]*California State University, Northridge, California*

Received 30 August 2011; Accepted 2 March 2012

Abstract: This study investigates the efficacy of an integrated science and literacy approach at the upper-elementary level. Teachers in 94 fourth grade classrooms in one Southern state participated. Half of the teachers taught the treatment unit, an integrated science–literacy unit on light and energy designed using a curriculum model that engages students in reading text, writing notes and reports, conducting firsthand investigations, and frequent discussion of key concepts and processes to acquire inquiry skills and knowledge about science concepts, while the other half of the teachers taught a content-comparable science-only unit on light and energy (using materials provided by their districts) and provided their regular literacy instruction. Students in the treatment group made significantly greater gains on measures of science understanding, science vocabulary, and science writing. Students in both groups made comparable gains in science reading comprehension. © 2012 Wiley Periodicals, Inc. J Res Sci Teach 49: 631–658, 2012

Keywords: science education; language and literacy; curriculum development

It is widely acknowledged that robust science learning occurs most effectively through firsthand experience combined with ample opportunities for reflection and rich talk (Bransford, Brown, & Cocking, 2000; Brown & Campione, 1994; Metz, 2000; National Research Council, 1996, 2000, 2011). While there has historically been tension between text-dominated and experience-dominated science instruction (Cervetti & Barber, 2008), the National Resource Council's (2011) conceptual framework for the new science education standards warns of the misrepresentation and marginalization of science and engineering in an approach to instruction that:

> Focuses predominantly on the detailed products of scientific labor—the facts of science—without developing an understanding of how those facts were established or that ignores the many important applications of science in the world. (p. 42)

Further, the NRC framework calls for the cultivation of science practices that reflect those engaged in by professional scientists and engineers, including reading science text, and

Additional Supporting Information may be found in the online version of this article.

Correspondence to: G. N. Cervetti; E-mail: cervetti@umich.edu

DOI 10.1002/tea.21015
Published online 3 April 2012 in Wiley Online Library (wileyonlinelibrary.com).

© 2012 Wiley Periodicals, Inc.

engaging in the specialized ways of talking and writing (p. 43). This represents a culmination of more than a decade of interest in the roles of reading and writing in inquiry-based science education (e.g., Glynn & Muth, 1994; Yore et al., 2004) and general growing interest in how language and literacy can be used to support the goals of inquiry-based science[1]. In recent years, many well-established inquiry-based science programs that began as predominantly ''hands-on'' approaches to inquiry have introduced texts in the form of science ''readers'' and science ''notebooking,'' including the series of NSF-funded inquiry-based curriculum programs initiated in the mid-1980s—programs that were born of the inquiry movement in science education (e.g., Lawrence Hall of Science, n.d.; National Sciences Resource Center, n.d.).

In a parallel but independent movement, reading educators have started to reconsider the role of content knowledge and genre in learning to read. They have become increasingly interested in how reading is shaped by genre and have started to question the wisdom of anchoring early literacy instruction almost exclusively in fictional narrative texts when expository texts constitute much of school reading beyond the primary grades (Duke & Bennett-Armistead, 2003; Palincsar, 2005) and when so many children find great interest in using reading to understand the natural world around them (Duke & Bennett-Armistead, 2003). Consensus is growing around the idea that genre makes a difference, that is, that students need guidance in learning to negotiate the structural and rhetorical terrains of different genres (RAND, 2002).

At the same time, knowledge—especially that which can be acquired through content-area instruction—has come to be viewed by many reading educators as both a foundation and a motive for reading. Palincsar (2005) describes the relationship between reading and knowledge as follows:

> Children who are exposed to text have opportunities to acquire knowledge of vocabulary, background knowledge, and knowledge regarding how reading material is structured that children who are not exposed to text do not have. Students with this richer knowledge base experience a bootstrapping of further vocabulary, real-world knowledge, and knowledge of and comfort with the structure of texts. (p. 3)

Reading educators such as Neuman (2006) increasingly suggest that acquiring information about the world is important in its own right and that it is a foundation for reading and writing in the future. Moreover, reading as part of scientific investigations has been shown to invite more engaged reading by providing both a context of growing expertise and a motive for reading (e.g., Guthrie & Cox, 2001; Guthrie & Wigfield, 2000).

Research on Integration

The parallel movements in science and literacy education have given rise to impressive lines of research that investigate what is possible when inquiry-based science goes beyond the obligatory use of textbooks or science readers to embrace a more authentic and synergistic relationship with learning to read and write as part of scientific inquiry (e.g., Guthrie, Anderson, Alao, & Rinehart, 1999; Palincsar & Magnusson, 2001; Romance & Vitale, 1992, 2001). Two programs of science-literacy integration—the CORI program and In-depth Expanded Applications of Science (IDEAS)—are particularly notable for their longevity and for the impacts on student learning that they have demonstrated across both domains.

In CORI, instructional activities, including instruction in reading comprehension and writing, focus on disciplinary conceptual understanding (Guthrie et al., 1999). Hands-on science activities are used to support students' concept development, to create situational interest

to energize their reading, and to provide opportunities for students to pose conceptual questions that can become the basis for reading activity and instruction. The goal of the CORI program is to increase reading engagement and improve reading achievement through integration with content-area learning (Guthrie et al., 1999). Across a series of small-scale studies in the middle elementary grades, the CORI intervention has consistently been associated with positive outcomes for reading comprehension, reading strategy use, and conceptual understanding using a variety of performance tasks and paper-and-pencil measures; often the positive effects extend to measures of reading motivation (Guthrie et al., 1999; Guthrie et al., 2009).

In IDEAS classrooms (Romance & Vitale, 1992, 2001), the time that is traditionally allocated to language arts instruction is replaced by a joint science-reading program. The IDEAS model includes concept-focused teaching, firsthand activities, reading, writing, and the construction of propositional concept maps based on reading. The IDEAS model places the in-depth study of core science concepts at the center of instruction both as a way of improving science teaching and creating coherence among literacy and science activities. Romance and Vitale (1992, 2001) have demonstrated through a long series of studies that IDEAS students outpace comparison students on standardized measures of science and reading achievement and display more positive attitudes toward and self-confidence in reading and science. The researchers attribute these results to the focus on conceptually meaningful structured knowledge development as an organizing mechanism that links the various firsthand science activities and literacy activities. The alliance between science and literacy was initiated as a way of reclaiming time in the school day for science instruction, but the researchers have demonstrated in a series of studies that reading achievement also benefits from integration.

While CORI and IDEAS are the longest-standing programs of science–literacy integration at the elementary level, there are a number of other efforts to integrate science and literacy instruction that have demonstrated similarly impressive effects on student learning. Notably, Varelas, Pappas, and their colleagues (Varelas, Pappas, & Rife, 2006; Varelas & Pappas, 2006) have conducted a series of studies that examine the relationship between dialog, text, hands-on experiences, and science understanding. In one of these studies, Varelas and Pappas (2006) examined the intertextual connections made by first and second grade students as they engaged in dialogically oriented read alouds in science. The researchers found that students made more intertextual connections among texts, discussions, and experiences as the unit progressed and that this intertextual connection-making provided opportunities for students to develop more conventional forms of scientific ideas and scientific language.

More recently, Fang and Wei (2010) examined the efficacy of a year-long inquiry-based science program for 6th graders with and without reading infusion. The reading infusion included reading strategy instruction and out-of-school reading of science trade books. Students who participated in the reading infusion condition made greater gains on a standardized assessment of reading ability and a curriculum-referenced science test, and received higher science grades compared with students who participated in the inquiry-based program alone. The researchers conclude that strategy instruction enabled students in the reading infusion condition to better cope with content-area texts, improving their learning in both reading and science.

Our efforts build on these programs, particularly CORI and IDEAS. Most notably, we have appropriated the goal of developing knowledge of important science concepts, both as an end unto itself and as a means of engendering purposes for and engagement in reading. In

addition, we readily accept Guthrie and Alao's (1997) idea that the resulting synergy provokes students to process text deeply. Like both of these programs, the experimental program in this study embeds cognitive strategy instruction into the integrated curriculum and rely on firsthand science activities both to advance conceptual understanding and to build knowledge in support of reading. Finally, we share an underlying conviction that science can provide an authentic context for applying reading and writing skills and strategies.

The study described in this article investigates the efficacy of integrated science and literacy instruction at the upper-elementary level. This investigation employs a curriculum model, instantiated in a unit about light that engages students in reading, writing, investigating, and discussing to acquire knowledge about important science concepts, inquiry skills, and literacy skills that students need to be successful in science. The program of research and development that includes the present study was initiated as an attempt to reconceptualize inquiry-based science instruction as inseparable from scientific literacy practices.

The Treatment Model of Science–Literacy Integration

According to Stoddart, Pinal, Latzke, and Canaday (2002), there are three principal approaches to the integration of content areas: (i) a thematic approach characterized by the use of overarching themes to create connections among domains; (ii) an interdisciplinary approach in which content or processes in one domain are used to support learning in another; and (iii) an integrated approach, in which emphasis on two or more domains is balanced. The model of science–literacy integration examined here is situated squarely within the integrated tradition. The model was developed by a team of science and literacy educators and researchers (including some of the authors of this article). The broader program of research and curriculum development began eight years ago as an attempt to infuse authentic scientific literacy practices into existing inquiry-based science units in order to enhance learning in both domains. We viewed integration as an opportunity to enhance science learning by making it better contextualized, more informed, and better documented. We also viewed integration with science as an opportunity to situate literacy instruction in a powerful knowledge-building domain.

The conceptual framework underlying this work rests on the idea that an integrated model that honors two disciplinary traditions can build mutually reinforcing relationships between literacy and science inquiry. One example of principle of mutual benefit through integration rests in the model's approach to the use of text. While many inquiry-oriented science educators have expressed concerns about text undermining the process of scientific investigation (e.g., Yore, 2000), we developed a series of roles for text that directly support students' involvement in inquiry, that provide meaningful (investigation-driven) experiences with nonfiction text genres, and that mirror the ways that scientists use text in investigations. Cervetti and Barber (2008) provide a detailed discussion of these roles. Briefly, reading provides opportunities for students to revisit concepts about physical phenomena experienced directly or through models in the classroom, to view these phenomena in the wider context of the world outside of the classroom, and to learn about how these phenomena are studied by professional scientists. For example, in the light unit that served as the treatment in this study, students consult findings from their own investigations of light as well as a reference book that provides data about the interactions of light with different materials. Students use data from the book to extend and challenge their firsthand observations of light interacting with different materials. Texts are also used in the light unit to model the processes and dispositions that real scientists bring to their work and to learn about the difficult-to-observe role of

light in vision. The model of science–literacy integration under study here relies on the understanding that these kinds of text-based experiences deepen students' involvement in firsthand investigations, enhance students' conceptual understandings, and support their ability to navigate science text.

The conceptual framework for our model of science–literacy integration also relies on the idea that inquiry-based science and literacy share skills, strategies, and goals that can be capitalized upon as the central features of integrated instruction. For example, both literacy education and science education share an interest in using discussion to support sense-making and comprehension of complex ideas. The development of the experimental model was informed by studies over the previous two decades that had started to shed light on the role of discussion in supporting students' development of conceptual understandings in science, meaningful participation in scientific inquiry, and ability to communicate their scientific understandings (Mercer, Dawes, Wegerif, & Sams, 2004; Richmond & Striley, 1996; Syh-Jong, 2007).Taken together, this research suggests that talk supports students' science learning in at least three ways. First, discussions provide opportunities for knowledge sharing and co-reasoning (e.g., Rivard & Straw, 2000). Students not only share information that, taken together, contributes to more coherent understandings and more successful problem-solving than could be accomplished alone, they also use dialog to talk through, clarify, and negotiate their growing understandings with others—forms of processing ideas that ultimately support understanding (Gee, 2002; Rivard & Straw, 2000). Second, talk can help make models of scientific thinking and reasoning available to students (Duschl & Osborne, 2002). Elusive scientific practices, such as weighing evidence and evaluating claims, can be made visible in talk. Third, talk creates opportunities for students to experience science as a process of revision and, therefore, to learn about the nature of scientific knowledge in addition to science concepts (Driver, Newton, & Osborne, 2000; Herrenkohl, Palincsar, DeWater, & Kawasaki, 1999). The model of integration investigated here involves a central role for student-to-student talk as a way of making sense of science investigations and gaining insight into the nature of scientific reasoning.

The model of science–literacy integration under study further relies on the understanding reading and scientific investigation are both acts of inquiry—driven by an interest in understanding and by a motive of gathering and making sense of evidence in order to learn about the natural world— and there is some evidence that inquiry and comprehension share goals, functions, and strategies (Padilla, Muth, & Lund Padilla, 1991). In our work on science–literacy integration, we have often used the word, ''synergistic'' to describe this relationship (Cervetti, Pearson, Bravo, & Barber, 2006). For example, predicting, inferring, and questioning are as central to ''inquiry'' in the discipline of science as they are the core of the ''comprehension'' in the literacy domain. In our development of integrated instruction, we take advantage of this synergy by targeting pairs of highly related inquiry/comprehension strategies. In the light unit used in this study, for example, students predict which materials will block light before their first-hand investigations, and they make predictions before and during reading. In each situation, students revise their predictions based on additional evidence they gather. Students are asked to reflect on the similarities and differences of prediction in the two contexts.

An additional synergy between science and literacy rests upon the understanding that mature word knowledge in science can be described as conceptual knowledge. In science, word knowledge involves understanding of words as they are situated within a semantic network of other words and ideas. From this perspective, word learning in science can be approached as conceptual learning—that is, words can be thought of *as* the surface-level

instantiations of underlying concepts that can be connected to other concepts to form rich conceptual networks (Armbruster, 1992; Bravo & Cervetti, 2008). For example, within the light unit, we treat word learning as conceptual learning by (i) providing students with repeated, multimodal (during inquiry activities, conversations, text reading, and writing) exposure to conceptually important, related words in the context of ongoing investigations; (ii) engaging students in activities that help them see the deep relationships between concepts/words; and (iii) providing opportunities for students to build an active understanding of the concepts/words through talk and writing. In addition, we believed that topically narrow reading would support students' conceptual vocabulary learning by providing opportunities to encounter the words repeatedly in meaningful contexts, building knowledge that would increasingly support their comprehension of subsequent content-relevant texts (Krashen, 2004).

The experimental model distinguishes itself from earlier efforts in several ways. First, we have sought to be more explicit in attending not only to the ways in which science supports literacy development, but also to the ways in which literacy supports science development, that is, the authentic roles that language and literacy can play in supporting knowledge acquisition and involvement in inquiry. In particular, we set out to develop a curriculum-based model of science–literacy integration that would serve both domains equally by (a) capitalizing on the knowledge-building context of science for supporting students' reading comprehension, informational writing, and academic language, and (b) enriching students' experiences in inquiry science by including reading, writing, and discussion. Second, we have made language a more central focus of our curriculum model. Building on Yore's notion that language is a ''critical communication and thinking tool'' that is ''integral to science'' (Yore, 2004, p. 71), we designed a model of integration that attends to students' development of academic and technical vocabulary, their involvement in sense-making talk around their investigations, and their ability to communicate their growing understandings in talk and writing. Third, in a direct attempt to respond to the critique by science educators of text-based approaches to teaching science (e.g., National Research Council, 2000), we have actively avoided the supplanting of investigations with reading by developing and implementing an elaborated model of how text can and should support inquiry-based science (Cervetti & Barber, 2008).

Research Design and Method

Design

In this article, we investigate the efficacy of an integrated science and literacy approach to instruction as compared to separate, content-comparable science instruction and literacy instruction. We test a model of integration as instantiated in a curriculum unit for fourth grade students. This study asks, *How does an integrated approach to science–literacy curriculum compare to business-as-usual approaches in terms of outcomes of science understanding, reading comprehension, science vocabulary and science writing?* During the 2007–2008 school year, we conducted an experimental study to address this question. The study examines the efficacy of a light unit as one instantiation of our model of science–literacy integration.

Participants

The study took place in one Southern state, selected as the study site because of the close correlation between that state's fourth grade science standards addressing the topic of light and the content of the treatment unit, enabling us to readily locate content-comparable comparison sites. In addition, the state has a strong system of standards

compliance, including regular site visits to verify that teachers at each school are engaged in on-topic standards-based instruction. This increased the predictability of content-comparable business-as usual instruction. Teachers were recruited with the assistance of district-level science coordinators. We engaged in two rounds of recruitment—one for Fall 2007 and one for Spring 2008. Several meetings were held at sites across the regions to inform principals and teachers about the project. In order to participate in the study, a teacher had to be teaching fourth grade and have available a minimum of three hours per week for a minimum of eight weeks for unit implementation. This time constraint applied to both the treatment and comparison groups.

A total of 100 applications were received. Three applicants were denied participation; one of these applicants did not meet the eligibility requirement of being able to teach science for a minimum of three hours per week, and the other two submitted incomplete applications. Three teachers dropped out of the study after being accepted but before beginning to teach. One of these teachers declined to participate for personal reasons, and the other two were unable to secure administrative approval to participate. Thus the study population consisted of ninety-four teachers (60 in Fall 2007, 34 in Spring 2008) from 16 school districts. The districts were located in urban, suburban, and rural areas in the same southern state.

After being accepted into the study, each teacher was randomly assigned to either a treatment or comparison group using a random number generating algorithm that is part of the Microsoft Excel software. Results indicate that treatment teachers were less experienced both in total years of experience (9.2 vs. 11.6 years among comparison group teachers) and years teaching at grade four (4.4 vs. 5.6 years). Comparison teachers were also more educated, with 51% versus 34% having an advanced degree. Neither the differences in experience or in advanced degrees was statistically significant. Advanced degrees held by teachers in both the treatment and comparison groups were almost entirely in education (including early childhood education, special education, educational leadership, educational technology, etc.) and/or psychology. Only four out of the 94 teachers reported having advance degrees in another area. (fisheries, biology, sociology, social work). Salary (as indexed by the natural log of salary) was roughly equal across groups. Class size was roughly equal across conditions.

The student demographic characteristics presented in Table 1 are based on the means for fourth grade students at participating schools during the 2007–2008 school year. We estimated the student demographic characteristics for the treatment and comparison groups by assuming that the distribution in a given classroom would reflect the distribution within that grade for the entire school. The mean percentage of students receiving free or reduced price lunch in the treatment schools was 57.6; in comparison schools, it was 52.9.

Table 1
Student participant characteristics

Characteristic	Comparison classroom (%)	Treatment classroom (%)
Gender		
Female	49.7	49.3
Ethnicity		
Asian/Pacific Islander	2.7	2.8
African-American	35.7	38.5
Hispanic	5.7	7.1
Native American	0.5	0.9
Multiracial	2.4	2.2
White	53.1	48.5

Treatment and Comparison Interventions

Treatment Intervention. The unit that served as the treatment intervention in this study is designed to engage students in reading, conducting firsthand investigations, discussing, and writing in the interest of developing their conceptual understandings about the characteristics of light, light interactions, and light as energy. We were guided in the development of the curriculum intervention by a set of principles for science–literacy integration.

Teachers in the treatment group were given a set of researcher-designed integrated science–literacy materials on the topic of light. The treatment unit focused on the characteristics of light, light interactions, and light as energy. While we believe that the most robust approach to supporting change in instruction will come about through providing teachers with curriculum *and* associated professional development, the purpose of the study reported here was to learn more about the curriculum portion of the equation. Accordingly, treatment teachers conducted an ''out-of-the-box'' implementation, receiving only the curriculum materials. Teachers in the treatment group received a teacher's guide with step-by-step instructions on each left hand page and text designed to provide related information and support on each right hand page. The right hand page teacher support included content and pedagogical information, instructional rationale, and suggestions for modifying the instructional sequence according to the characteristics and needs of different implementation contexts. In addition to the teacher's guide, teachers in the treatment group received 18 copies of nine different researcher-designed nonfiction science books (most ranging from 18 to 24 pages each), an investigation notebook for each student, and a kit of materials for students to use in the firsthand activities of the unit, including materials such as flashlights and hand lenses.

The treatment unit was forty sessions in length, comprised of four investigations—each with 10 sessions. Sessions were designed to be taught in 45–60 minutes. In each 10-session investigation the equivalent of approximately four sessions are devoted to firsthand (hands-on) activities, two sessions to reading, two sessions to writing, and two sessions to discourse, review of important concepts, or assessment activities. While this framework guided the development of the unit, and the treatment unit held the proportion of learning modalities constant across each investigation, the flow of the instruction was designed to be seamless and coherent. The following description of the first of the four investigations in the light and energy unit provides an example of the how the instruction flows.

In Investigation 1: Characteristics of Light, students begin by reflecting on what they know and wonder about light. They make the first of many predictions before reading *Can You See in the Dark?,* a book that invites students to wonder about whether or not people need light to see. This book introduces the idea that all light comes from a source and enables students to identify sources of light in the text and illustrations. Next, students go on to investigate with flashlights— their own light sources—to see what they can observe themselves. Their investigations involve students in making predictions, gathering evidence, and then revising their predictions to reflect new evidence. They then make light tubes and use them to gather evidence that light travels in a straight line. The class begins the creation of the Class Concept Wall (a giant concept map on the wall of the classroom). After making predictions, they read *The Speed of Light,* a book that presents data comparing the speed of light to other fast things through descriptive examples and in tables. By reflecting on the data, the students are better able to understand how fast light travels—a characteristic that is impossible for them to observe firsthand. Using key words, students construct main idea statements about passages in the book. They summarize what they have learned about the characteristics of light by writing details to support a topic sentence.

Approximately, 40% of the treatment unit (four sessions for each of the four investigations) involved students in using firsthand investigations to gather evidence about the characteristics of light, light interactions, and light as energy. Investigations were driven by guiding questions such as, ''What are the characteristics of light?'' ''What materials transmit light?'' ''What materials reflect light?'' and ''How do lenses interact with light?'' and involved students in using manipulatives. While the unit involves students in engaging in a variety of inquiry practices, teachers provided explicit instructions and systematic opportunities for practice related to the strategies of: making predictions, summarizing, evaluating claims and evidence, and making explanations from evidence. These focus strategies in inquiry were also selected as the focus strategies in reading (making predictions) and writing (summarizing, evaluating claims and evidence, and making explanations from evidence). Once students had been introduced to a cognitive strategy, such as predicting, in the context of reading, the teacher would re-instruct the strategy in the context of firsthand investigations, first reminding students about the utility of the strategy for reading and then discussing its application in firsthand investigations. Students then practiced using the strategy in their investigations—for example, using the strategy of prediction to leverage what they already know in order to think ahead about the process and possible outcomes of a firsthand investigation.

Approximately, 20% of the treatment unit (two sessions for each of the four investigations) involved students in engaging with student science books. The student science books were designed to serve specific roles in supporting students' involvement in the unit's firsthand science investigations (Cervetti & Barber, 2008). The books were also designed to support students' development of fluency, vocabulary, comprehension, and understanding of informational text features and structures. The model of text and its role in supporting students' literacy development has been described at length by Hiebert (2006). The texts offered students repeated opportunities to encounter specific, targeted vocabulary words related to the topic of light (e.g. absorb, block, characteristic, emit) and ''academic'' words (words that are commonly used in science and other school disciplines) that are highly useful as they communicate about their investigations (e.g., analyze, claim, data). The books also featured an array of common nonfiction text features (e.g. headings, diagrams, captions, table of contents, index) that were used to teach students to navigate and comprehend science text. The nine books in the unit were designed to build in their conceptual complexity over the course of the unit as a way of building knowledge to support students' comprehension of subsequent texts.

Within each unit, eight sessions were devoted to reading the science books. Instruction in the reading sessions was organized using a before, during, and after reading framework. Before reading, students were typically engaged in setting goals for their reading related to their ongoing investigations, and they were often introduced to a selected comprehension strategy, such as predicting and summarizing. Each strategy was instructed using a gradual release of responsibility approach (Pearson & Gallagher, 1983) in which the teacher: (i) directly explained and modeled the strategy, (ii) provided guided practice with the strategy in the context of reading, (iii) provided opportunities for independent practice during reading over the course of the unit, and (iv) regularly discussed the utility of the strategy and application to other situations. Students read the books in partners, and then engaged in a whole class discussion of their learning from the text and, if applicable, the utility of the comprehension strategy introduced before reading. Students often used these eight books and a ninth reference book at different points throughout 40 sessions in order to obtain information related to their investigations. The treatment unit also offered explicit instruction in the use of targeted inquiry skills and literacy strategies and ongoing opportunities for the application of these

skills and strategies in firsthand situations and in text. These skills and strategies included making predictions and summarizing. An additional 20% of the treatment unit focused on writing. In the light unit, students received explicit instruction and repeated opportunities for practice in writing summaries of what they were learning (main idea and supporting details) as well as writing scientific explanations (claim and evidence). As with reading comprehension, writing instruction was structured using a gradual release of responsibility approach. For writing summaries, teachers instructed, modeled, and provided practice opportunities for students related to creating main ideas and supporting details. Constructing main ideas was instructed using a key word strategy, in which students identify the most important words in a reading passage and use those words to create a main idea statement. Adding supporting ideas was instructed using a concept map approach to organize ideas for supporting details around the main idea statement. For writing scientific explanations, teachers instructed how to use data tables to summarize results of investigations, how to use that summary to write a claim, and then how to support the claim with evidence from the data table. In addition, there were multiple opportunities for students to engage in reflective writing through periodic reflections.

Approximately, 20% of the plan for the unit was devoted to a collection of activities related to reflection and formative assessment, including regular opportunities for informal student-to-student talk, reflection on word relationships, and structured small group discussions (discourse circles) of science content and processes. Each of these reflection and discourse activities were repeated instructional routines. One repeated full-session routine called Discourse Circles, involved students working in small groups to analyze a statement, collect evidence that supports the statement and evidence that refutes it, and then engage in a discussion about whether the evidence leads them to agree or disagree with the statement. Teachers were guided in facilitating whole class discussions using several class reflection routines including the Debrief Discussion in which the teacher uses a series of questions to invite student to share the evidence they gathered during an investigation and their conclusions based on the evidence.

Comparison Group Intervention. Teachers in the comparison group were asked to present the content of their state science standards related to the topic of light, using curriculum materials they regularly use for the same amount of time each week and for the same duration. The treatment light unit included content related to two of the three major subtopics of the state's light standards: characteristics of light and light interactions. A third subtopic (light and color) appears in the state standards, but is not addressed in the treatment light unit. Conversely, a third subtopic that is addressed in the treatment light unit (light as energy) does not appear in the state standards. Thus, while teachers used a variety of different materials to teach these standards, the standards were well-aligned with the treatment light unit, thus creating a content-comparable comparison. See Table 2 for a comparison of the light treatment unit and the state standards. As we describe below, the analysis of science learning

Table 2
Comparison of light content state standards and the treatment light unit

	Characteristics of light	Interactions of light	Light and color	Light as energy
Light-related state standards	×	×	×	
Treatment light unit	×	×		×

focused on items that address the two overlapping topics, characteristics of light and light interactions.

Data Collection

Data collection took place in two waves—Fall 2007 and Spring 2008—so that teachers could participate according to the science content coverage timelines for their respective districts. A student pretest assessing student knowledge of science vocabulary, reading comprehension, and science understanding related to the topic of light, was administered by research staff in all treatment and comparison classrooms. Teachers then taught either an integrated science–literacy unit on light (treatment) or their regular science lessons on light (comparison) and their regular literacy instruction. Immediately following the completion of the unit, researchers administered a post-test (identical to the pretest) to the students as part of their classroom activities. In addition, an assessment of student writing was administered by teachers before and after they taught their light unit and returned to researchers via U.S. Mail. Teacher measures included a background survey and an end-of-unit survey regarding their teaching practices during the study.

Student Learning Measures

The student learning outcomes were measured through a pre–post assessment developed through a process coordinated by the Center for Research, Evaluation, and Assessment at the Lawrence Hall of Science in partnership with the curriculum research and development team. Items included in the assessment were selected based on three criteria: (i) centrality to the domain and (ii) alignment with state science content standards, and (iii) alignment with the content covered in the treatment unit. An independent scientist review panel evaluated the items for fit with criterion (i), an affiliated assessment specialist evaluated the items for fit with criterion (ii), and the R&D team evaluated the items for fit with criterion (iii). The assessment validation process included the external reviews related to criterion (i) above as well as studies of multiple iterations of pilot assessments. Wherever possible items were drawn from existing validated assessments. In other instances, items were developed and reviewed by a team of assessment developers, curriculum developers, and classroom teachers.

The first set of items (Version 1) were developed in classrooms alongside the treatment curriculum as it was being developed and field tested in a study that took place approximately one year prior to the randomized experiment reported herein. Student responses to the field test items were used to refine the items, and the revised instrument was tested in a validation study ($n = 166$). The validation study version of the assessment (Version 2) included 76-items designed to assess students' mastery of science vocabulary, reading comprehension, science writing, and science understanding. Each scale was examined for dimensionality and reliability as well as for item fit, difficulty and distribution. The science understanding scale was also examined for correlations with each teacher's ratings of their individual students' science abilities on a three-point scale (low-medium-high).

Rasch analysis revealed a good range of difficulty and alignment to student ability. At the same time, reliabilities for the scales were modest—ranging from $\alpha = 0.47$–0.67. Further principal components analysis suggested problems with dimensionality, with some items loading more strongly on a second component within a given scale. As an index of concurrent validity, we found that the correlations between students' science understanding scores and teacher ratings of their level of science ability were statistically significant, but with more overlap in scores across levels than desired. This suggested that the scale was well aligned with teachers' science content goals assessments but also that there was a need to improve

scale reliability. The results from the validation study led the team to modify and/or replace items so that there were a sufficient number of items addressing the construct of interest for each scale.

Science Understanding Measure. The final makeup of the science understanding measure consisted of forty-two multiple choice items; 15 had two answer choices, while the remaining 27 items had four response options. Research staff categorized 19 items as factual measures of declarative knowledge. The remaining 23 items were categorized as schematic or strategic measures of conceptual understanding.

For the purpose of the study described herein, we analyzed a subset of these items. This subset includes 23 items that were determined by external science experts to fairly assess both the content of the treatment unit and the state science standards that guided instruction for the comparison group. By utilizing this subset of data, we even further increase the likelihood that the content of the assessment was included in the comparison classrooms as it was in the treatment classrooms.[2]

Given the iterative process of measure development, reliabilities were calculated on the final versions of these assessments using the data collected for the study itself. The reliability for the science understanding measure is based on the 23 items included in the analysis for this article. The alpha reliability of the final measure was 0.84 at pretest and 0.81 at post-test.

Science Writing Measure. The writing assessment asked students to respond to an open-ended prompt, "How does light interact with materials? Give three examples." It was expected that students who had *not* learned about light (at the pretest) would still be able to address this prompt in some way, while students who *had* studied light would be able to describe at least three interactions, give evidence of those interactions, and/or explain what was happening in each type of interaction.

Rubrics were developed by project staff for six dimensions: use of evidence, introduction, clarity, conclusion, vocabulary definition, and science content. The dimensions are described in Appendix SA[3]. A count was also done for the number of times students used science vocabulary words in their writing (vocabulary count). This count was included in the analysis as a seventh dimension of science writing. A subset of randomly selected student papers was used to refine the scoring rubrics. The science writing assessments were scored by project staff and trained undergraduate students at University of California, Berkeley. The writing tests were blind-scored as to condition and time point (pre-or-post).

A subset of matched pretest and post-test science writing assessments were randomly selected from each of the classrooms for which both pretest and post-test science writing data was returned by teachers. In all, the work of 467 students was scored. Each scorer achieved a 90% or higher inter-judge reliability score with an assessment specialist on the six rubric-based Science Writing dimensions and the seventh dimension, vocabulary count. Twenty percent of the 467 articles were used to initially calculate inter-rater reliability among scorers. The remaining 80% of tests were to be scored after the external evaluator concurred that we had achieved a sufficient level of inter-rater reliability. After the scorer training sessions, the scoring process began with double-scoring of each dimension on the science writing prompt rubric for a 20% sample of the remaining papers. Differences were resolved by project staff, and scorers were retrained as needed. The overall inter-rater reliability for the scoring of the seven writing dimensions was 0.85 at pretest and 0.79 at post-test.

Science Vocabulary Measure. Items were developed for the science vocabulary measure using two item formats: a definition matching format designed to test whether students can

identify the meaning of the words from a list of possible definitions, and cloze item formats that were designed to test students' ability to apply knowledge of the words in context. Words for the assessment were selected based on expert assessments of their centrality to the topic of light. All of the words were instructed in the treatment unit as part of a much larger set and appeared in the student books, and all but one appeared in the standards that served as a guide for the comparison teachers. Sixteen items were included in the draft for the validation study. The reliability for the 16-item measure was 0.61. Following the validation study, we replaced three poor-performing items (that were very easy or had poor discrimination values).

The final science vocabulary measure included 15 items. Eight of these were multiple-choice definition matching items for the words reflect, transmit, emit, refract, shadow, light, interact, and transform. Students were given the target word and asked to select the best definition from four options. The remaining seven items were multiple-choice cloze sentences for the words reflect, refract, transmit, transform, shadow, interact, and absorb. Students were given sentence with blank and asked to select a word to fill in the blank. The same target words were used across the two item-types, because the item formats were designed to elicit different kinds of understandings about the words. While the definition matching items were designed to test whether students can identify the meaning of the words, the cloze items were designed to test students' ability to apply knowledge of the words in context. The reliability of the final measure was 0.43 at pretest and 0.69 at post-test.

Reading Comprehension Measure. The reading comprehension measure is a researcher-designed set of expository passages and multiple-choice questions. The multiple-choice items for each passage were developed using a framework of item types—recall, main idea, inferences, predictions, and questioning. Item developers were given descriptions of these types, as well as examples and sentence frames. Once a pool of items had been developed that included each type, items were selected based on a review by the development team. This measure was tested as part of the initial validation study described above. The reliability for the instrument was 0.67. We determined that many of the items were easy (with p-values above 0.85) and that there were too few items overall.

Given these results, we developed a second version of the measure, adding more challenging items and increasing the number of items to 16. We replaced one passage—a passage about bats—with a more difficulty passage about hummingbirds, which had also been validated as part of a different research study. We then conducted a second validation study of this revised reading comprehension measure. For the purpose of the second study, we administered the revised reading comprehension instrument to 202 students from diverse backgrounds. The reliability improved (0.70). Based on these results, we modified the instrument slightly—changing the distracters on one item and replacing one item—before administering as part of this study. The reliability of the final measure was 0.77 at pretest and 0.76 at post-test.

Data Analysis Procedures

Given that students were assigned to treatments by teacher and given that teachers were nested within schools, a multilevel modeling framework was used to take advantage of the data structure by examining the potential impact of context on treatment effects. By using a three-level random effects model, we are able to divide the variation in achievement into between-student, between-teacher, and error components. This is particularly important because data containing multiple levels of aggregation can lead to errors in interpretation when

these multiple levels are ignored (Aitkin & Longford, 1986; Burstein, 1980). A basic two level unconditional model is akin to the traditional ANOVA model, although more flexible with respect to assumptions and modeling options (Raudenbush & Bryk, 2002).

In this study, we utilized various multilevel models (MLM) that are derivative of the three level model that we present below. We utilized a three-level MLM to examine the research hypothesis that the treatment intervention has a significant impact on student performance on measures of science understanding, writing, science vocabulary, and reading comprehension. We briefly present this model below. In general, the model consists of three levels and allows for a flexible specification of the covariance structure at every level of the analysis (Snijders and Bosker, 1993).

The level one model is:

$$Y_{ijk} = r_{0jk} + e_{ijk}, \tag{1a}$$

where Y_{ijk} is the outcome (e.g., science understanding assessment) for student i in class[4] j in school k, and represents the unconditional, or base, level one model. Where π_{0jk} represents mean outcome of classroom j in school k and e_{ijk} is a random student effect.

At level two (between teachers, within schools) we model the impact of the treatment, given that treatment assignment was by teacher (teacher level).

$$\pi_{0jk} = \beta_{00k} + \lambda_{01k}\mathrm{TRT}_{jk} + r_{0j} \tag{2}$$

In (2) β_{00k} represents the school mean performance while λ_{01k} represents the treatment effect, r_{0jk} is a random teacher effect. Using (2) alters the interpretation of π_{0jk}. Now π_{0jk} is the mean class performance of comparison classrooms and $\pi_{0jk} + \lambda_{01k}$ is the mean performance of treatment classrooms.

$$\begin{aligned} \beta_{00k} &= \gamma_{000} + u_{00k} \\ \lambda_{01k} &= \gamma_{010} \end{aligned} \tag{3}$$

In (3) γ_{000} is the grand mean of student performance. γ_{010} is the overall treatment effect.

The level one model represented in (1a) can be further specified to account for differences in classroom intake characteristics, for example, pretest performance or student background characteristics. The level 1 model becomes:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(Y_{ijk}Y_{\cdot\cdot k}) + e_{ijk}, \tag{1b}$$

Hence, $\pi_{0\mathrm{jk}}$ becomes the adjusted mean outcome of comparison[5] classroom j in school k.

$$\pi_{1jk} = \beta_{10k} + \gamma_{11k}\mathrm{TRT}_{jk} + r_{1jk} \tag{2b}$$

Given the extension (or possible extension) in (1b), the level two model can be respecified to include treatment indicators. Hence, β_{10k} represents the mean class relationship between the pretest and the post-test in comparison classrooms. γ_{11k} represents the cross-level interaction between the treatment and pretests scores. Whereas γ_{01k} represents the main effect of the treatment, that is, did treatment classrooms outperform comparison classrooms, given pretest performance, γ_{11k} estimates whether the treatment is differentially effective for students with different levels of preparedness, that is, pretest scores. This cross-level interaction can be used to test whether the treatment is differentially more effective for low achievers or

more effective for high achievers. Additional student characteristics can be added to (1b) and tested by expanding (2b) (e.g., including ELL status in model 1b and adding a γ_{11k} TRT_{jk} into 2b).

At level three we account for the fact that classrooms are nested within schools. Using an average pretest for the classroom tests the impact of the classroom average achievement, or context, on individual student post-test performance. Interactions between school level variables test whether school context impacts student performance after accounting for student and teacher inputs.

Results

Teachers

Table 3 includes descriptive results for teacher practices as derived from teacher surveys. Table 3 also presents indicators of teacher practices prior to the treatment period. This includes both time and instructional mix. A key element of pre-existing teacher practice is experience with and disposition toward inquiry-based teaching practices. Employing a standard of reporting the use of hands-on practices at least 25% of the time, 73% of comparison teachers as compared to 61% of treatment teachers would be considered inquiry-based at the start of the study. Also important is the fact that comparison teachers reported teaching the topic light more often than treatment teachers prior to this study.

Students

Table 4 presents descriptive results for the Student Assessments. These results indicate that mean scores on the science vocabulary, reading, and science tests were higher on the post-test than on the pretest in both the comparison classrooms and treatment classrooms. For each (treatment and comparison) group, attrition from pretest to post-test varied by subject and was generally around 5%. For science understanding and science vocabulary the null

Table 3
Descriptive results for teacher practices

	Comparison classrooms			Treatment classrooms		
	Mean	*n*	*SD*	Mean	*n*	*SD*
Pre-study teacher practices						
Hrs Sci. Instruct	3.74	47	1.19	3.59	47	1.04
Hrs Lit Instr.	9.57	47	4.39	9.84	47	5
Inquiry-based 0–24%	0.28	47	0.45	0.38	47	0.49
Inquiry-based 25–49%	0.38	47	0.49	0.38	47	0.49
Inquiry-based 50–74%	0.26	47	0.44	0.21	47	0.41
Inquiry-based 75–100%	0.09	47	0.28	0.02	47	0.15
Number times taught light	3.69	47	4.16	2.84	47	2.56
During study teacher practices						
Hours teaching science/wk	3.03	47	1.18	3.66[a]	47	102.6
Percent of science time students spent						
Doing hands on inquiry	26.85	47	18	24.7	47	10.5
Reading	22.55	47	15.6	19.92	47	8.88
Class discussions	24.68	47	12.5	24.89	47	7.26
Science writing	11.49	47	6.09	15.61[a]	47	7.57
Science vocabulary	14.74	47	9.2	14.78	47	5.71

[a]Difference between treatment and control significant at $P < .05$.

Table 4
Descriptive results for student assessment

	Comparison classrooms			Treatment classrooms		
	Mean	*n*	*SD*	Mean	*n*	*SD*
Vocabulary pretests	11.67	992	2.55	11.33	1027	2.62
Vocabulary post-test	12.89	939	2.79	13.72	974	3.51
Reading pretests	10.21	992	3.28	9.59	1026	3.46
Reading post-test	10.72	936	3.06	10.30	969	3.29
Science pretests	12.59	937	2.15	12.42	976	2.12
Science post-test	14.05	937	2.58	15.41	976	3.45

hypothesis that pretest scores differed based on the missing status of post-test scores could not be rejected. However, in reading there was evidence that students missing post-test scores had somewhat higher (0.16 *SD*) scores on the pretest. Comparison students' reading pretest scores were about 0.30 *SD* higher for non-missing post-tests, while treatment students' reading pretest scores were about 0.06 *SD* higher. Results based on alternative specifications[6] were robust and consistent with results presented below.

Pretest Scores. Preliminary multilevel models using pretests as outcomes indicated that pretest science understanding scores did not vary significantly among classrooms, and there was no difference in mean pretest performance between treatment and comparison classrooms. However, both science vocabulary and reading comprehension pretests results indicated significant between-teacher variability in scores and also significant differences between treatment and comparison classrooms favoring the comparison group. The comparison classrooms, as a group, scored about 0.10 *SD* higher in reading and about 0.30 *SD* higher in science vocabulary at pretest intake. Given the fact that pretests are related to post results, it is important to account for intake differences when evaluating treatment outcomes.

The results in Table 5 indicate students in both treatment and comparison classrooms demonstrated statistically significant gains ($p < 0.01$) from pretest to post-test on the science understanding measure, science vocabulary measure, and reading comprehension measure.

Science Understanding. Table 6 summarizes the results that speak most directly to the major question of interest, comparing the differential impact of the treatment and comparison

Table 5
Student gains

	Gain	*SE*
Science understanding		
Treatment	2.99	0.12***
Comparison	1.46	0.10***
Science vocabulary		
Treatment	0.69	0.086***
Comparison	0.39	0.079***
Reading comprehension		
Treatment	2.38	0.104***
Comparison	1.18	0.090***

***$p < 0.01$.

Table 6
Estimated treatment effects on student science understanding post-test results (raw scores)

			1		2	
Model[a]	Base	*SE*	Base	*SE*	Base	*SE*
Fixed effects						
Mean post-test	14.07	0.214				
Comparison classroom			14.07	0.159***	14.06	0.158***
Treatment[b]			1.47	0.289***	1.51	0.285***
Treatment effect size[c]			0.65		0.65	
Treatment interaction						
Treatment effect size[d]					0.05	0.072
		SD		*SD*		*SD*
Random effects						
Post-tests						
Student		2.68		2.68		2.63
Classroom		1.33***		1.09***		1.07***
School		0.93***		0.99***		1.00***
Deviance		8990.0; df = 4		8969.8; df = 5		8929.6; df = 9
χ^2 for model improvement				19.8***		59.9***

***$p < 0.01$.
[a]Odd numbered models include only unconditional treatment effects. Even numbered models estimate conditional treatment effects, conditioned on pretests and pretests by treatment joint.
[b]Treatment effect represents marginal treatment effect.
[c]Effect size estimated as δ, (Treatment − Comparison)/*SD*(Outcome).
[d]Effect size estimated comparing effect at (± 1 *SD* mean of pretests)/*SD*(outcome).

treatment. The results in Table 6 summarize the two models examining the treatment science understanding results. Model one tests the main effect of the treatment and answers the question whether students in treatment classrooms scored higher on the post-test, accounting for the fact that the treatment was assigned at the classroom level and classrooms were nested within schools. The results indicate that treatment classrooms scored about 1.5 points higher on the science understanding post-test, which is an effect size of about 0.65. Model two tests whether there is a joint effect between the pretests and the treatment, that is, whether the relationship between the pre-and the post-test is different in treatment and comparison classrooms. If this effect is significant, it provides evidence that the treatment is not equally effective for students with different pretest scores. The results for model two imply that there is no joint effect (essentially the pre–post slopes are parallel in treatment and comparison classrooms and hence there is no differential change in the performance gap between high and low achievers due to the treatment). Thus, the treatment effect on science understanding is robust for students at all points along the distribution of pretest scores.[7]

Science Vocabulary and Reading Comprehension. Table 7 presents results for both the science vocabulary and the reading comprehension measures. Models three and five present results testing only the treatment condition and the comparison condition without accounting for pretest scores. It is important to note the results based on models three and five are not conditioned on pretests performance and hence are not affected by the relatively low pretests reliability of the science vocabulary measure. The results indicate that students in the treatment condition scored significantly higher than students in the comparison condition on the

Table 7
Estimated treatment effects on student science vocabulary and reading comprehension post-test results (raw scores)

	Science vocabulary						Reading comprehension					
Model[a]	Base	*SE*	3	*SE*	4	*SE*	Base	*SE*	5	*SE*	6	*SE*
Fixed effects												
Mean post-test	13.35	0.22					10.51	0.13				
Comparison classroom			12.97	0.21	12.97	0.17			10.69	0.17	10.46	0.17
Treatment[b]			0.75	0.23***	0.91	0.22***			−0.36	0.26	0.12	0.26
Treatment effect size[c]			0.23		0.22				−0.11		0.09	
Treatment interaction												
Treatment effect size[d]				0.13	0.08						−0.03	0.45
	SD		*SD*		*SD*		*SD*		*SD*		*SD*	
Random effects												
Post-tests												
Student	2.90		2.90		2.60		3.00		3.00		2.30	
Classroom	0.93***		0.85***		0.63		1.05***		1.03***		0.45***	
School	1.18***		1.15***		1.05		0.13*		0.16*		0.02	
Deviance	9,245; df = 4		9,245; df = 5		8,853; df = 9		9,334; df = 4		9,332; df = 5		8,295; df = 9	
χ^2 for model improvement			8.5***		391.6***				1.5		1,038***	

***$p < 0.01$.
[a]Odd numbered models include only unconditional treatment effects. Even numbered models estimate conditional treatment effects, conditioned on pretests and pretests by treatment joint effects.
[b]Treatment reflects marginal treatment effect.
[c]Effect size estimated as δ, (Treatment − Comparison)/*SD*(Outcome).
[d]Effect size estimated comparing effect at (±1 *SD* mean of pretests)/*SD*(Outcome).

science vocabulary measure at post-test. The effect size is approximately 0.23. The results for reading comprehension indicate that treatment and comparison students did equally well on the post-test. Models four and six include pretest as a covariate in the analysis and test whether there are joint (i.e., interaction) effects between pretest and treatment. Neither model four nor model six indicate that treatment operated differentially for students with differing pretest scores. Thus, the results for science vocabulary and reading comprehension are the same whether pretest is entered into the model: a significant effect favoring the treatment for science vocabulary and no differences between groups for reading comprehension.

Science Writing. We also examined the impact of the treatment on differences between groups in performance on the science writing measure. Our analyses were based on a randomly selected subset of students who participated in the study ($n = 467$)[8]. Table 8 presents the correlations among the seven writing dimensions assessed in each essay. It is important to reiterate that ratings were subject to a generalizability analysis that determined that there is sufficient precision in scores to use them for additional analyses. The results in Table 8 indicate that the correlations among assessed domains are moderate at best indicating that, in general, they tap into different aspects of student writing.

Table 8
Correlations among writing dimensions

	Vocab. use	Vocab. count	Evidence	Introduction	Conclusion	Clarity
Pretest						
Science concepts	0.54	0.48	0.68	0.48	0.35	0.31
Vocabulary definition	1.00	0.46	0.64	0.42	0.28	0.33
Vocabulary count		1.00	0.31	0.45	0.38	0.31
Evidence			1.00	0.38	0.29	0.36
Introduction				1.00	0.62	0.24
Conclusion					1.00	0.28
Clarity						1.00
Post-test						
Science concepts	0.55	0.66	0.63	0.56	0.30	0.52
Vocabulary definition	1.00	0.37	0.67	0.37	0.25	0.36
Vocabulary count		1.00	0.33	0.57	0.31	0.46
Evidence			1.00	0.33	0.17	0.39
Introduction				1.00	0.46	0.44
Conclusion					1.00	0.28
Clarity						1.00

In order to determine whether the treatment had a significant impact on student science writing, we first examined overall writing achievement based on the observed scores on the seven dimensions. The results are presented in Table 9. The results indicate that at the pre-tests, there was marginally suggestive evidence ($p < 0.10$) that comparison students had higher writing achievement. The results in Table 9 also indicate that at the post-test students in the treatment group had higher latent writing achievement ($p < 0.05$). The treatment effect size is 0.40.

We conducted additional exploratory analyses to examine separate results on the seven individual writing dimensions (Table 10). Overall, the results in Table 10 refine the results

Table 9
Estimated treatment on latent student science writing results

	Science writing	
	Estimate	*SE*
Fixed effects		
Mean pretests		
Comparison classroom	1.8	0.04*
Treatment[a]	−0.095	0.057
Mean post-test		
Comparison classroom gain	0.22	0.026***
Treatment[a]	0.60	0.038***
Treatment effect size[b]	0.4	
Random effects		
Heterogeneous random effects		
Model fit-from null model		
Change in deviance χ^2	442***	
Change in df	3	

***$p < 0.01$, *$p < 0.10$.
[a]Estimate reflects marginal treatment effect.
[b]Effect size estimated as δ, (Treatment − Comparison)/*SD*(Outcome).

Table 10
Estimated treatment on student writing by dimension

Writing Dimension	Unconditional proportion of var. btwn. teachers	Gain Scores	*SE*	Effect size[a]
Fixed effects				
Science concepts	0.19			
Comparison classroom		2.1	0.07***	
Treatment[b]		0.59	0.12***	0.63
Vocabulary definition	0.11			
Comparison classroom		2.01	0.10***	
Treatment[b]		0.22	0.14	
Vocabulary count	0.30			
Comparison classroom		2.74	0.13***	
Treatment[b]		1.49	0.20***	0.8
Pre–post science GAIN		0.08	0.02***	0.85
Evidence	0.13			
Comparison classroom		1.84	0.09***	
Treatment[b]		0.38	0.15**	0.33
Post science score		0.03	0.01***	0.66
Introduction	<0.01			
Comparison classroom		2.36	0.07***	
Treatment[b]		0.35	0.10***	0.38
Conclusion				
Comparison classroom	<0.01	1.97	0.04***	
Treatment[b]		0.06	0.05	
Post science score		0.01	0.01	0.41
Clarity	0.13			
Comparison classroom		1.81	0.05***	
Treatment[b]		0.33	0.09***	0.43
Pre–post science GAIN		0.02	0.01***	0.77

***$p < 0.01$, **$p < 0.05$.
[a]Treatment effect sizes as in Table 9, note (3); GAIN and score effect sizes as in Table 9, note (4).
[b]Treatment effect represents marginal effect of treatment.

presented in Table 9 by suggesting that among the seven writing dimensions, only vocabulary definition and conclusion demonstrated no treatment effect. The remaining five dimensions demonstrated statistically significant treatment effects with effect sizes ranging from 0.33 (evidence) to 0.80 (vocabulary count). The models in Table 10 also examined whether there were any effects on writing associated with science understanding, under the research hypothesis that science understanding has a positive effect on writing scores. The results indicate that in some instances the pre–post gain[9] in science understanding that was related to better writing scores, and other instances it was overall science understanding as measured at post-test that was associated with higher writing scores. Both vocabulary count and clarity were associated with gains on the science understanding measure. The effect sizes were quite large, 0.85 and 0.77 for vocabulary count and clarity, respectively. Both the evidence and conclusion dimensions were impacted by overall science understanding, as represented by post-test scores (but not pretests or gains). The effect sizes were moderate, 0.66 and 0.44, for evidence and conclusion, respectively. These results are consistent with the hypothesis that science understanding is positively associated with writing performance. It is interesting to note that the effects of science understanding were independent effects of the treatment; and in one case (conclusion) occurred without a significant treatment effect. It is important to note that

the subset of students for whom we have both writing, pre and post-test science understanding scores ($n = 458$) scored similarly on the treatment science pretest and post-test to the entire sample (23.3 vs. 23.3 on the pretests and 27.4 vs. 27.04 on the post-test, for the writing sample students and the entire sample, respectively). Hence, these results are not attributable to performance of students who were exceptional on science performance.

Implementation Results

We gathered information about implementation as a part of the teacher self-report post-survey data for both treatment and comparison teachers. These data reveal several insights about the ways in which teachers implemented the treatment and comparison units.

Treatment teachers reported primarily using the treatment unit for science, whereas comparison teachers used whatever they typically use to instruct about the topic of light. Because the comparison teachers were located in multiple school districts around the state, there was little commonality to what to the materials used as a comparison treatment. The comparison group teachers reported using a wide range of approaches, though many relied at least in part on a textbook. Seventy-seven percent of the comparison group teachers reported that they used a textbook in response to the open-ended survey question, ''What materials did you use to teach your science unit?'' Several mentioned the Harcourt Brace Science (11%) or McGraw Hill Science (4%) textbooks. Sixty percent of comparison group teachers reported using specific hand-on materials, such as flashlights, prisms, and lenses, and another 15% reported use of hands-on materials generically or use of lab experiments. Twenty-eight percent reported using technologies, such as video or internet sites. Similarly, treatment and comparison teachers report using a wide range of materials for literacy instruction, including basal textbooks, trade books, magazines, and teacher gathered materials from the Internet and elsewhere.

Treatment teachers reported spending more time on science instruction than comparison teachers (219.8 vs. 182.3 minutes/week). Treatment teachers, however, reported that more of their science instruction also included attention to literacy (reading and writing in the context of the science unit) compared with comparison teachers (77 vs. 61 minutes/week). Even though we do not know the nature of the reading and writing included in the range of comparison science units, we do know that the reading and writing included in the treatment science unit includes explicit instruction in the use of literacy strategies, rather than just practice reading and writing. And while treatment and comparison teachers estimate approximately the same amount of time spent reading in the context of their science unit (43 vs. 41 minutes/week), treatment teachers report more time spent writing in the context of their science unit (34 vs. 20 minutes/week). Teacher reported time on (respective) task was used as a covariate for all three outcomes; it proved to be marginally significant ($p \leq 0.10$) in science understanding but not reading nor writing. Thus, although treatment teachers spent more time on science instruction than comparison teachers, time on task does not appear to account for treatment effects.

Discussion

Summary of Findings and Discussion

In this work, we have evaluated the efficacy of a curriculum-based approach to science–literacy integration by assessing its impact on learning in comparison with a ''business-as-usual'' approach (teachers addressing the same unit content, light, with their normal curriculum materials in fourth grade classrooms). We asked, *How does an integrated approach to science–literacy curriculum compare to business-as-usual approaches in terms of outcomes of*

science understanding, reading comprehension, science vocabulary and science writing? Using conservative approaches to statistical analysis—a three level HLM with students nested within classrooms (teachers) and classrooms nested within schools—we found moderate effect sizes in favor of the treatment for science learning (ES = 0.65). For science vocabulary, we observed a small effect (ES = 0.22) in favor of the experimental treatment. We have additional evidence of vocabulary learning from the writing measure: while the treatment students did not more often provide definitions of science words in their writing (the vocabulary definition dimension), they did more often use science words in their writing (ES = 0.80; the word count dimension). For the science writing measure, we found a moderate multivariate effect for the seven dimensions of the rubric (ES = 0.40). The univariate analyses of writing demonstrated separate effects favoring the treatment for five of the seven dimensions: science concepts, vocabulary count (but not vocabulary definition), evidence, introduction (but not conclusion), and clarity. It is important to note that treatment students actually included more science concepts than comparison students in their responses to the writing prompt on their post-test assessments.

Overall, the results show promising, occasionally robust, trends on science and literacy outcomes, thus contributing to the growing body of evidence (e.g., Guthrie et al., 1999; Palincsar & Magnusson, 2000; Romance & Vitale, 2001) suggesting that integrated approaches not only benefit student science learning outcomes, but also support student literacy development. In addition to providing additional evidence that both science outcomes and literacy outcomes are supported by integration, the study adds to understandings about the impact of science–literacy integration. Compared with previous studies, this study included a wider range of literacy measures, including measures of vocabulary and writing. The use of a range of measures is most significant as it is related to another important aspect of this study that distinguishes it from previous work: the model of integration that underlies the curriculum tested in this study provided balanced attention to science and literacy growth, rather than using one domain to achieve growth in the other or to reclaim time in the school day. In Stoddart et al.'s (2002) framework, this approach is integrated, rather than merely interdisciplinary. The curriculum under study here rests on the understanding that engaging students in reading, writing, and discussing directly linked to their firsthand investigations would not only support their conceptual understandings, but also their ability to communicate these understandings in talk and writing, and that doing so would support students' development of important dimensions of informational literacy—including reading comprehension, vocabulary knowledge, and expository writing. While we cannot confirm every aspect of this underlying conceptual framework, given the non-significant results for reading comprehension, the results do suggest that explicit attention to literacy in the context of science is supportive of students' conceptual growth and at least some dimensions of their informational literacy skills.

Underlying the broad conceptual framework regarding the efficacy of integration was a set of specific theoretical principles regarding the relationship between literacy learning and science learning. These principles included the understanding that many cognitive strategies, such as predicting, are shared across domains, that vocabulary knowledge and conceptual knowledge are closely related in content-area learning, and that reading can enhance conceptual learning from science investigations. This study evaluated the efficacy of a complex model of science–literacy integration governed by these principles. As such, we cannot attribute outcomes to specific elements of the framework. At the same time, the results do suggest that another intervention that relies on the same understandings should provide similar results.

Unlike our earlier studies and unlike other studies of science–literacy integration, and unlike the CORI and IDEAS studies, we found no effect on reading comprehension of science passages (Wang & Herman, 2005). Several explanations seem plausible, but our data do not allow us to distinguish definitively among them. First, it could be that the reading comprehension component of the treatment was not robust enough to support more growth in reading comprehension than the array of regular reading curricula that characterized the various comparison classrooms. While other studies (e.g., Guthrie et al., 2009; McNamara, O'Reilly, Best, & Ozuru, 2006) have demonstrated gains in reading comprehension in interventions of similar or lesser duration, the comprehension component of the treatment in this study is only one element in a complex array of instructional features and goals. The volume of reading (nine books over eight weeks) and instruction (approximately 20 minutes/week of discussion and explicit teaching associated with the book reading) is small relative to the amount of literacy instruction that students, both treatment and comparison, received in their regular English language arts instruction. With science in general and light in particular, students began the study as relative novices—and at similar levels of understanding as measured by the science understanding pretests—and the treatment (intervention or comparison) represented a major change in dosage in exposure to science concepts. However, given what we know about instruction with the current crop of commercial reading programs, it is likely that students in both groups were exposed to a great deal of comprehension instruction by fourth grade and continued to receive relatively large doses during the study. Although we share many commitments and instructional principles with programs that have demonstrated effects on comprehension (e.g., Guthrie et al., 2009; McNamara et al., 2006), attention to comprehension in the treatment unit was modest by comparison to these programs. We were, however, able to show a broad array of other effects for literacy including science vocabulary knowledge, productive vocabulary use, and writing. Writing, like science, is receiving little attention in many classrooms, so it is especially important treatment groups made major advances there.

A second plausible explanation for the lack of an effect on reading comprehension outcomes focuses on important differences in design and analysis between the current study in comparison to the Guthrie et al. (2009) and McNamara et al. (2006) studies that show the comprehension effect. Both of these predecessors were smaller in size and scope than the study reported herein. They also used comprehension measures that were closely related to the intervention and the treatment texts (and were highly labor-intensive to score). Further, they both used conventional analysis of variance rather than HLM as the primary statistical tool. Consistent with the way in which the treatment was implemented in the present study (classrooms were randomly assigned to treatment), we used classroom means rather than individual students as the unit of analysis.

A third explanation turns on the extreme difficulty teachers experience in attempting to teach reading comprehension strategies (Hacker & Tenant, 2002; Wilkinson & Son, 2011). It may be that teachers experienced greater difficulties in implementing the reading comprehension piece of curriculum we provided without substantial professional development. Recall that the treatment teachers received only a kit of materials with a teacher's guide. It is possible that a teacher's guide provides sufficient support for implementing science and vocabulary instruction the first time through, but that more time and additional support in the form of professional development is required to help teachers implement comprehension strategy instruction around science texts. It is important to note, however, that a curriculum-based approach has the potential for much wider implementation than one involving intensive professional development. This study is true-to-life in the sense that teachers were asked to

perform an out-of-the-box implementation much like what they would experience in receiving a new set of materials from a publisher.

Finally, the most plausible explanation is that we have yet to create the valid index of the construct of science reading comprehension for which we (and others) strive. We used a format in which students read topically more and less related science passages and answered multiple-choice questions ranging from the factual to the inferential. The passages and questions do not align well with aspects of reading or characteristics of text that are unique to science, let alone our particular approach; in short, we have yet to develop a tool that would exhibit more precise estimates for reading in the context of science. It is interesting to note that in our earlier work, we obtained significant treatment effects with younger children using a different kind of measure—a MAZE test, which is a multiple choice (pick the missing word from a set of three) version of a cloze (fill in the blank) test. It might be that the MAZE format is more sensitive to the application of key vocabulary knowledge from the unit in completing the ''choose-the-word-to-fill-in-the-blank'' task. All these speculations aside, it is clear that further work is needed in order to find a way to impact and measure reading comprehension within the treatment.

Nonetheless, the thrust of the findings across science knowledge, vocabulary, and writing provide compelling evidence for great uptake and application of key ideas within the integrated curriculum. Students in the integrated curriculum acquired more vocabulary related to the key concepts in the unit, they learned more of the big ideas related to light, and, perhaps most significantly, they used both the ideas and the words when they wrote about what they had learned. Acquiring knowledge, the words we use to name that knowledge, and using the ideas to convey what you have learned to others—those seem like the very goals to which any literacy or science educator would aspire.

In addition to the issues highlighted above, we recognize a number of limitations to the analytic efforts presented within this artcile. The most significant is related to the amount of information we could gather about the substance of our ''business-as-usual'' comparison given the scope and budget of this study. Although we gathered self-report post-surveys from teachers regarding the substance and duration of the teaching and learning that took place in these comparison classrooms, we still have an incomplete picture. Moreover, self-reports can be unreliable (Cook & Campbell, 1979). Future research efforts are being designed to gather additional implementation data in order to gain a more complete picture.

Additionally, while we took a broad perspective on the aspects of literacy that can support and be supported by science in the design of the intervention, we assessed only vocabulary knowledge, reading comprehension, and science writing. While the intervention also included attention to participation in science discussions, reading fluency, and the use of text features and structures, we were limited in the amount of time we had available for student assessment. While the limited measures risk a simplistic interpretation of the role of language and literacy in science, they do mirror those aspects of literacy instruction that are most central to educational standards in fourth grade.

Even with these limitations, this study addresses an important issue in science education and literacy education and contributes additional important evidence to the studies on science–literacy integration, by employing a broader range of measures and utilizing a comparison group that features a real and familiar alternative—teachers using content-comparable science units. Unlike most other studies on science–literacy integration, ours focuses on the impact of curriculum use only. It is noteworthy that treatment teachers taught the treatment unit ''cold,'' that is, without the benefit of any professional development. The results presented herein provide critical evidence for supporting integrated science–literacy approaches

such as the one described here. The implications of these findings are especially significant given that the current school accountability context has privileged time for literacy and mathematics over time for science and other subjects. Further, these findings are timely given the current convergence of the vision spelled out in the *Common Core State Standards for English Arts and Literacy in Science and Technical Subjects* (Achieve, 2010) and the *Framework for K-12 Science Education* (NRC, 2011), that students must be able to respond to the communication demands of science.

While a recent research synthesis by Minner, Levy, and Century (2010) provided strong evidence of the effectiveness of an inquiry-based approach to science fueled by firsthand experiences, and while inquiry-based science is the indisputable standard of high-quality science teaching, this study and those that preceded it suggest that the integration of experiences with language and literacy can support science learning while maintaining an inquiry orientation. The current study also suggests that doing so supports important learning outcomes in both domains.

The research literature is beginning to provide images of the integration of text- and inquiry-based approaches to science (e.g., Howes, Lim, & Campos, 2009). While this study and others have supported the integration of literacy and inquiry-based science in principle, it is also becoming clear that not all forms of integration are equal. For example, Howes et al. followed three elementary teachers as they integrated science and literacy. The researchers found that not all approaches to integration were equally supportive of students' involvement in inquiry, and that integration can result in the privileging of literacy learning over science learning.

In closing, we want to be clear that we believe the foregrounding of inquiry is central to the efficacy of our model of science–literacy integration. In our opinion, one of the most important aspects of the model tested in this study is the development of roles for text that support students' involvement in extended, multi-modal, and question-driven investigations. While more work is needed to refine some aspects of this model, we believe that there is sufficient evidence to suggest that literacy and firsthand experiences in science are best positioned as tools for inquiring about the natural world. In practice, this means foregrounding compelling science questions, following an inquiry cycle and using tools of text and firsthand experience to investigate those questions, and supporting access to science concepts, language, and text along the way.

Notes

[1]For example, an entire section of the April 23, 2010 issue of *Science* was devoted to this very issue (Alberts, 2010).

[2]With respect to the use of a subset of items for the science understanding measure: This was done in order to ensure that the measure was a good test of the content being taught in comparison classrooms, as well as treatment classrooms. In the end, there were only small changes in performance based on the use of the shorter assessment. For example, the treatment group's pre-test performance on the two versions (in terms of percent correct) was 0.8% on the pretest and 0.4% on the post-test. The differences for the comparison students were 1.5% on the pretest and 1.1% point on the post-test. The absolute difference (between treatment and comparison students) in scores is larger for the shorter test, but the standard deviation is larger for the 42-item test. The larger standard deviation associated with the longer test is consistent with the notion of construct irrelevant variance (specifically on the post-test, where comparison classroom students were assessed on content that they did not have opportunities to learn). This is corroborated by the correlations between the 42 and 23 item tests on the post-test, which is 0.91 for treatment students and 0.70 for control students. Treatment effects based on the 23-item assessment are slightly larger (due to increased precision realized by utilizing the subset of the item

most closely aligned with the state content standards) than treatment effects based on the 42-item assessment. Overall, the two assessments minimally changed student performance, and empirically allow for the same inferences about results.

[3]Additional materials are available as Supporting Information accompanying the online article.

[4]We use the term class and teacher interchangeably. It is natural to consider a group of students sitting in a classroom, but each classroom is taught by a single teacher. Moreover, student performance is considered to be impacted by the teacher.

[5]Comparison classroom given (2).

[6]We used simple mean replacement and indicators for cases with missing values to test whether students with replaced scores demonstrated significantly different results than students with complete data. The indicator was not significant; and given the relatively small attrition, treatment effect estimates were unchanged.

[7]We also analyzed the Science Understanding results using IRT rather than raw scores. While the effect size was somewhat smaller, neither the overall results nor the inferences drawn about effectiveness of the treatment changed.

[8]All students were administered the writing assessment, but we limited the analysis to a random subset due to costs associated with raters scoring each of the essays.

[9]Testing the impact of Gains tests whether there was transfer in student learning–implying complimentary science and writing performance. Testing the impact of science understanding levels tests whether content knowledge impacts writing related to the content.

References

Aitkin, M., & Longford, N. T. (1986). Statistical modeling issues in school effectiveness studies. Journal of the Royal Statistical Society, A, 149(1), 1–43.

Alberts, B. (2010). Prioritizing science education. Science, 328, 405.

Armbruster, B. B. (1992). Vocabulary in content area lessons. The Reading Teacher, 45, 550–551.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.) (2000). How people learn: Brain, mind, experience, and school. Washington, D.C: National Academy Press.

Bravo, M. A., & Cervetti, G. N. (2008). Teaching vocabulary through text and experience. In: A. E. Farstrup & S. J. Samuels (Eds.), What research has to say about vocabulary instruction (pp. 130–149). Newark, DE: International Reading Association.

Brown, A. L., & Campione, J. C. (1994). Guided discovery in a community of learners. In: K. McGilly (Ed.), Classroom lessons: Integrating cognitive theory and classroom practice (pp. 229–272). Cambridge: MIT Press.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. In: E. Berliner (Ed.), Review of research in education, Volume 8 (pp. 158–233). Washington, DC: American Educational Research Association.

Cervetti, G. N., & Barber, J. (2008). Text in hands-on science. In: E. H. Hiebert & M. Sailors (Eds.), Finding the right texts: What works for beginning and struggling readers (pp. 89–108). New York: Guilford.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues. Boston, MA: Houghton Mifflin Company.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. Science Education, 84, 287–312.

Duke, N. K., & Bennett-Armistead, V. S. (2003). Reading & writing informational text in the primary grades. New York: Scholastic Teaching Resources.

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. Studies in Science Education, 38, 39–71.

Fang, Z., & Wei, Y. (2010). Improving middle school students' science literacy through reading infusion. The Journal of Educational Research, 103, 262–273.

Gee, J. P. (2002). Identity as an analytic lens for research in education. Review of Research in Education, 25, 99–125.

Glynn, S., & Muth, K. D. (1994). Reading and writing to learn science: Achieving scientific literacy. Journal of Research in Science Teaching, 31(9), 1057–1073.

Guthrie, J. T., & Alao, S. (1997). Designing contexts to increase motivation for reading. Educational Psychologist, 32, 95–105.

Guthrie, J. T., Anderson, E., Alao, S., & Rinehart, J. (1999). Influences of concept-oriented reading instruction on strategy use and conceptual learning from text. Elementary School Journal, 99, 343–366.

Guthrie, J. T., & Cox, K. E. (2001). Classroom conditions for motivation and engagement in reading. Educational Psychology Review, 13, 283–302.

Guthrie, J. T., McRae, A., Coddington, C. S., Klauda, S. L., Wigfield, A., & Barbosa, P. (2009). Impacts of comprehensive reading instruction on diverse outcomes of low- and high-achieving readers. Journal of Learning Disabilities, 42, 195–214.

Guthrie, J. T., & Wigfield, A., (2000). Engagement and motivation in reading. In: M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), Handbook of reading research: Volume III (pp. 403–422). New York: Erlbaum.

Hacker, D. J., & Tenant, A. (2002). Implementing reciprocal teaching in the classroom: Overcoming obstacles and making modifications. Journal of Educational Psychology, 94, 699–718.

Herrenkohl, L. R., Palincsar, A. S., DeWater, L. S., & Kawaski, K. (1999). Developing scientific communities in classrooms: A sociocognitive approach. The Journal of the Learning Sciences, 8(3/4), 451–493.

Hiebert, E. H. (2006). Becoming fluent: What difference do texts make? In: S. J. Samuels & A. E. Farstrup (Eds.), What research has to say about reading fluency (pp. 204–226). Newark, DE: International Reading Association.

Howes, E. V., Lim, M., & Campos, J. (2009). Journeys into inquiry-based elementary science: Literacy practices, questioning, and empirical study. Science Education, 93, 189–217.

Krashen, S. D. (2004). The case for narrow reading. Language Magazine, 3(5), 17–19.

Lawrence Hall of Science. (n.d.). Foss science stories. Nashua, NH: Delta Education.

McNamara, D. S., O'Reilly, T. P., Best, R. M., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. Journal of Educational Computing Research, 34, 147–171.

Mercer, N., Dawes, L., Wegerif, R., & Sams, C. (2004). Reasoning as a scientist: Ways of helping children to use language to learn science. British Educational Research Journal, 30, 359–372.

Metz, K. E. (2000). Young children's inquiry in biology: Building the knowledge bases to empower independent inquiry. In: J. Minstrell & E. van Zee (Eds.), Inquiring into inquiry in science learning and teaching (pp. 371–404). Washington, D.C: AAAS.

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction—What is it and does it matter? Results from a research synthesis years 1984 to 2002. Journal of Research in Science Teaching, 47, 474–496.

National Research Council. (1996). National science education standards. Washington, DC: National Academy Press.

National Research Council. (2000). Inquiry and the national science education standards: A guide for teaching and learning. Washington, DC: National Academy Press.

National Research Council. (2011). A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Washington, DC: National Academy Press.

National Sciences Resource Center. (n.d.). Science and Technology Concepts Program. Washington, DC: National Academies.

Neuman, S. (2006). How we neglect knowledge—And why. American Educator, 30, 24–27.

Padilla, M. J., Muth, K. D., & Lund Padilla, R. K. (1991). Science and reading: Many process skills in common?. In: C. M. Santa & D. E. Alvermann (Eds.), Science learning—Processes and applications (pp. 14–19). Newark, DE: International Reading Association.

Palincsar, A. S. (2005). Reading in science: Why, what, and how (Brief). Washington, DC: National Science Resources Center. Retrieved from http://nsrconline.org/pdf/ReadingInScienceEssay.pdf

Palincsar, A. S., & Magnusson, S. J. (2000). The interplay of firsthand and text-based investigations in science class. Ann Arbor, MI: CIERA Report 2-007. Retrieved from http://www.ciera.org/library/reports/inquiry-2/2-007/2-007.pdf

Palincsar, A. S., & Magnusson, S. J. (2001). The interplay of firsthand and text-based investigations to model and support the development of scientific knowledge and reasoning. In: S. Carver & D. Klahr (Eds.), Cognition and instruction: Twenty five years of progress (pp. 151–194). Mahwah, NJ: Lawrence Erlbaum.

Pearson, P. D., & Gallagher, M. C. (1983). The instruction of reading comprehension. Contemporary Educational Psychology, 8, 317–344.

RAND Reading Study Group. (2002). Reading for understanding. Santa Monica, CA: RAND.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods. Newbury Park, CA: Sage Publications.

Richmond, G., & Striley, J. (1996). Making meaning in classrooms: Social processes in small-group discourse and scientific knowledge building. Journal of Research in Science Teaching, 33, 839–858.

Rivard, L., & Straw, S. (2000). The effect of talk and writing on learning science: An exploratory study. Science Education, 84, 566–593.

Romance, N. R., & Vitale, M. R. (1992). A curriculum strategy that expands time for in-depth elementary science instruction by using science-based reading strategies: Effects of a year-long study in grade four. Journal of Research in Science Teaching, 29, 545–554.

Romance, N. R., & Vitale, M. R. (2001). Implementing an in-depth expanded science model in elementary schools: Multi-year findings, research issues, and policy implications. International Journal of Science Education, 23, 272–304.

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. Journal of Educational Statistics, 18, 237–259.

Stoddart, T., Pinal, A., Latzke, M., & Canaday, D. (2002). Integrating inquiry science and language development for English language learners. Journal of Research in Science Teaching, 39(8), 664–687.

Syh-Jong, J. (2007). A study of students' construction of science knowledge: Talk and writing in a collaborative group. Educational Research, 49, 65–81.

Varelas, M., & Pappas, C. C. (2006). Intertextuality in read-alouds of integrated science–literacy units in urban primary classrooms: Opportunities for the development of thought and language. Cognition and Instruction, 42, 211–259.

Varelas, M., Pappas, C. C., & Rife, A. (2006). Exploring the role of intertextuality in concept construction: Urban second-graders make sense of evaporation, boiling, and condensation. Journal of Research in Science Teaching, 43, 637–666.

Wang, J., & Herman, J. (2005). Evaluation of Seeds of Science/Roots of Reading Project: Shoreline Science and Terrarium Investigations. Los Angeles, CA: CRESST, UCLA (http://scienceandliteracy.org/research/efficacy_studies).

Wilkinson, I. A. G., & Son, E. H. (2011). A dialogic turn in research on learning and teaching to comprehend. In: M. L. Kamil, P. D. Pearson, E. Moje, & P. Afflerbach (Eds.), Handbook of reading research: Volume IV (pp. 359–387). New York: Erlbaum.

Yore, L. D. (2000). Enhancing science literacy for all students with embedded reading instruction and writing-to-learn activities. Journal of Deaf Studies and Deaf Education, 5, 105–122.

Yore, L. D. (2004). Why do future scientists need to study the language arts? In: E. Saul (Ed.), Crossing borders in literacy and science instruction (pp. 71–94). Arlington, VA: NSAT Press.

Yore, L. D., Hand, B., Goldman, S. R., Hildebrand, G. M., Osborne, J. F., Treagust, D. F., & Wallace, C. S. (2004). New directions in language and science education research. Reading Research Quarterly, 39, 347–352.