Explanatory Modeling in Science Through Text-based Investigation: Testing the Efficacy of the READI Intervention Approach

Project READI Technical Report #27

Susan R. Goldman[1], Cynthia Greenleaf[2], Mariya Yukhymenko-Lescroart[1] with Willard Brown[2], Monica Ko[1], Julia Emig[1], MariAnne George[1], Patricia Wallace[3], Dylan Blum[3], M.A. Britt[3], and Project READI.

PROJECT READi

inquirium

Northern Illinois University

NORTHWESTERN UNIVERSITY

UIC UNIVERSITY OF ILLINOIS AT CHICAGO

WestEd

Citation for the Report: Goldman, S. R., Greenleaf, C., and Yukhymenko-Lescroart, M., with Brown, W., Ko, M., Emig, J., George, M., Wallace, P., Blum, D., Britt, M.A. & Project READI. (2016). *Explanatory modeling in science through text-based investigation: Testing the efficacy of the READI intervention approach.* Project READI Technical Report #27. Retrieved from URL: www.projectreadi.org

Project READI: Kimberly Lawless[1], James Pellegrino[1], Gayle Cribb[2], Katie James[1], Candice Burkett[1], Angela Fortune[1], Cindy Litman[2], Stacy Marple[2], and Ashley Ballard[1].
We are indebted to the teachers, students, and administrators of the participating schools and districts without whose willing participation this work would not have been possible. Finally, we acknowledge the assistance with data collection and scoring of Michael Bolz[1], Allison Hall[1], Karina Perez[3], Jacqueline Popp[1], Francis Reade[2], and Kathryn Rupp[3].

Please send us comments, questions, etc.: info.projectreadi@gmail.com

Project READI was supported by the *Reading for Understanding (RFU)* initiative of the Institute for Education Sciences, U. S. Department of Education through Grant R305F100007 to the University of Illinois at Chicago from July 1, 2010 – June 30, 2016. The opinions expressed are those of the authors and do not represent views of the Institute or the U. S. Department of Education.

Project READI operated as a multi-institution collaboration among the Learning Sciences Research Institute, University of Illinois at Chicago; Northern Illinois University; Northwestern University; WestEd's Strategic Literacy Initiative; and Inquirium, LLC. Project READI developed and researched interventions in collaboration with classroom teachers that were designed to improve reading comprehension through argumentation from multiple sources in literature, history, and the sciences appropriate for adolescent learners. Curriculum materials in the READI modules were developed based on enacted instruction and are intended as case examples of the READI approach to deep and meaningful disciplinary literacy and learning.

[1] University of Illinois at Chicago (UIC)
[2] WestEd, Strategic Literacy Initiative
[3] Northern Illinois University (NIU)

Explanatory Modeling in Science through Text-based Investigation: Testing the Efficacy of the READI Intervention Approach

Introduction

National and international trends indicate that current reading comprehension instruction is not preparing citizens for full participation in 21st century societies (National Assessment of Educational Progress 2009a, b; Organization of Economic and Cultural Development, 2013). The accessibility of unprecedented amounts of information, much of it unfiltered and often contradictory, means that readers need to analyze, synthesize, and evaluate information within and across sources (e.g. print-based texts, audio and video "texts," images). Further, adolescents are expected to build knowledge and perform discipline-specific tasks requiring specialized ways of reading, thinking, and conveying information (Alvermann & Moore, 1991; Bazerman, 1985; Bromme & Goldman, 2014; Lee & Spratley, 2010; Moje & O'Brien, 2001; Shanahan & Shanahan, 2008; Snow & Biancarosa, 2003). The need is particularly acute for science because of public participation in decision making about quality of life issues (e.g., global climate change or genetically modified foods). Yet the evidence suggests that the public is ill-equipped to deal with the science underlying such issues (National Center for Educational Statistics, 2012). The Common Core State Standards (Council of Chief State School Officers, 2010) and the Next Generation Science Standards (NGSS) (Next Generation Science Standards Lead States, 2013) speak to these needs. For the diverse students in our nation's middle and high schools many of whom are profoundly ill prepared for the CCSS and NGSS standards, educators must simultaneously support literacy and science learning. A critical aspect of the challenge for adolescents is that they are expected to read to understand in multiple content areas. They are presented with discipline specific tasks and texts that require specialized ways of reading, thinking, and conveying information to others (Lee & Spratley, 2010; Moje & O'Brien, 2001; Shanahan & Shanahan, 2008). Yet, the disciplinary literacies - the oral and written communication practices of disciplines (Moje, 2008) - are rarely the focus of instruction, either in content area classes or in reading or English language arts classes.

Motivated in part by these gaps between the literacies citizens need in the 21st century and what they are graduating from high school with, various countries have undertaken different initiatives to redress the gap. This paper reports on the development and results of one such effort undertaken in the United States under the auspices of Project READI (Reading, Evidence, and Argumentation in Disciplinary Instruction). READI is a multi-institution collaboration of researchers, professional development designers and facilitators, and practitioners, funded in 2010 by the Institute for Education Sciences' Reading for Understanding initiative. Project READI conceptualizes adolescent reading comprehension as evidence-based argumentation from multiple sources in the academic disciplines and subject areas. *Multiple sources* reflects an expanded definition beyond traditional verbal text, similar to that adopted by Kress (Kress, 1989; Kress & Van Leeuwen, 2001) and the New London Group (1996) to include multiple media and forms of information representation. In evidence-based argumentation, claims are asserted and supported by evidence that has principled connections to the claim, but the nature of claims, evidence, and principles differ depending on the discipline (Goldman, et al., 2016).

Project READI took on the challenge of developing adolescents' capacity to engage in evidence-based argumentation from multiple sources in three content areas: literary reading, history, and science. To address this challenge, we pursued a set of overarching questions about forms and types of tasks, texts, instructional strategies and tools that would enable students to engage in evidence-based argumentation from multiple texts through iterative design–based research (Cobb, Confey, diSessa, Lehrer & Schauble, 2003; Reinking & Bradley, 2008) and shorter-term quasi-experiments. Design teams for each disciplinary area consisting of researchers, teacher educators, professional development and subject matter specialists, and classroom teachers collaboratively developed, implemented, and revised instructional designs for evidence-based argument instructional modules (E-B AIMs). Quasi–experimental studies tested features of tasks, texts, and supports. A third work strand focused on developing assessments that would provide evidence of student learning relative to the learning objectives.

The fourth work strand focused on teachers' opportunities to learn and followed directly from the READI theory of action. Simply put, teachers mediate students' opportunities to learn. However, many teachers have themselves had little opportunity to engage in inquiry-based approaches to literary reading, history, or science. Meeting in Teacher Inquiry Networks over the course of the project, teachers worked within their own disciplines to explore a variety of constructs and rethink their instructional practices. Constructs explored included argumentation, close reading, and disciplinary reasoning principles. Instructional practices included tasks they were assigning, texts they were using, opportunities for students to interact and engage in collaborative as well as individual sense-making, and how they orchestrated small group but especially whole class discussions. Overall, there was a strong emphasis on teachers learning how to move the intellectual work, including reading texts, from themselves to the students.

This paper reports an efficacy study of the READI science intervention design that was conducted in 9th grade biological sciences classes over the course of a semester (intended for Fall of 2014 but extended through February, 2015). Specifically, this paper examines the impact of READI instruction compared to business as usual 9th grade biology instruction. "Impact" was assessed in several areas including comprehension and use of information from multiple sources to support practices of understanding and producing models of science phenomena. In concert with the question of impact on students and consistent with the READI theory of action, a second research question addressed impact on teachers' attitudes, beliefs, and practices of their participation in the professional development and implementation of the READI intervention. We first provide a summary of the theoretical framework of the READI approach, especially regarding the functionalities of text-based investigation in science. We then describe the design of the intervention, including its teacher professional development, and the assessment strategy used to evaluate its impact on teaching and student learning. We then turn to the specifics of the methods and the results for teachers and students. We end with a discussion of the implications of this study for efforts to move teaching and learning practices toward critical reading, reasoning, and sense-making among adolescent learners. These implications relate to opportunities to learn for teachers as well as their students. Furthermore, in contrast to the typical positioning of literacy as relevant only to obtaining, evaluating, and communicating information in science, these findings also

suggest that disciplinary reading (and writing) practices have a more pervasive role in science learning.

Theoretical Framework: Reading for Understanding as Evidence-Based Argumentation in the Disciplines

The READI conceptualization of reading to understand built on conceptions of reading comprehension as involving the construction of mental representations of text that capture the surface input, the presented information, and inferences that integrate information within and across texts and with prior knowledge (e.g., Goldman, 2004; Kintsch, 1994; Rand, 2002; Rouet & Britt, 2011). We joined this perspective with a disciplinary literacies perspective on argumentation from multiple sources, thus integrating disciplinary reasoning practices with supporting literacy practices. As a general construct or discourse scheme, argumentation refers to the assertion of claims that are supported by evidence that has principled connections to the claim (Toulmin, 1958; Toulmin, Rieke, & Janik, 1984). However, what claims are about, criteria that define what counts as evidence relative to some claim, and the principles that warrant or legitimize why particular evidence does indeed support a particular claim depend on the disciplinary content area. As applied to traditional academic disciplines, what constitutes valid argument depends on the epistemology of a discipline (Goldman, et al., 2016) in conjunction with the discourse norms that the members of the disciplinary community have negotiated and agreed upon (Gee, 1992; Lave & Wenger, 1991). That is, members of a discipline constitute a discourse community and share a set of understandings about valid forms of argument and communication among members of the community (Goldman & Bisanz, 2002). These norms reflect the epistemology of the discipline – the nature of disciplinary knowledge and how new knowledge claims in that discipline are legitimized and established.

In the context of READI, we adopted a discipline-specific approach to defining what students needed to know *about* a discipline to support comprehension and production of argumentation in that discipline (Goldman, et al., 2016). This approach emerged from an extensive examination of theoretical and empirical literature on the reading practices of disciplinary experts, empirical examinations of adolescents' disciplinary reasoning, and close examination of the types of representations and discourse in which members of disciplinary communities engage. Table 1 summarizes the five categories of knowledge about a discipline that emerged from our conceptual meta-analysis of literary reading, history and science. Learning goals for each disciplinary content area reflect the intersection of reading and reasoning processes important in text-based inquiry from multiple sources with knowledge about the discipline specified in the core knowledge categories. Table 2 shows these learning goals for science (See for literary reading and history goals, Goldman et al., 2016). In other words, learning goals in literary reading and history specified the same processes as those in science (close reading, synthesis across multiple information sources, explanation, justification, critique, and demonstration of disciplinary epistemology) but the specifics reflect the epistemic orientations of literary reading or history, the nature of claims, evidence, and reasoning principles appropriate to each, and the kinds of texts that are read and produced. Thus, reading and reasoning processes of argumentation had similar labels in each of the three

disciplines, but what students were closely reading, what they were trying to bring together – the patterns they were attempting to discern, the kinds of explanations they were seeking to construct, justify, and critique - were specific to each discipline (Goldman, Ko, Greenleaf, & Brown, in press).

Text-Based Investigations to Support Scientific Inquiry and Literacy Practices

The reasoning practices of science foreground evidence-based argumentation around the development of models that explain phenomena of the natural world[1] (Cavagnetto, 2010; Osborne & Patterson, 2011; Windschitl & Braaten, 2008). Projects focused on supporting students to develop explanatory models have provided students with frameworks for explanation, modeling, and argumentation, using datasets or hands-on investigations as stimuli for modeling and explanation tasks (Berland & Reiser, 2009; Chin & Osborne, 2010; McNeill & Krajcik, 2011; Passmore & Svoboda, 2012;). Very little of the work on modeling and explanation has been carried out in the context of science reading. Yet students need discipline-specific skills to navigate the complex and varied representations, including models, that convey science information. Data are tabulated, displayed, summarized, and reported in graphs, tables, and schematics and there are conventional linguistic frames that constitute the rhetoric of argument in science (Lemke, 1998; Norris & Phillips, 2003; Osborne, 2002; Pearson, Moje, & Greenleaf, 2010). Thus, there are abundant natural synergies between science and literacy, providing ample opportunities for teaching and learning key academic literacies as well as science inquiry practices (National Research Council, 2012; Pearson, et al., 2010).

The focus of the READI work on text-based investigations centrally involved the use of authentic science texts to construct knowledge, draw on information and evidence, and develop explantions and arguments that fit the data. These are essential skills of science yet are not ones in which students are typically instructed or engaged (Yore, Bisanz & Hand, 2003; Yore, 2004; Osborne, 2002). Indeed, for some of the subdisciplines of science, data come in the form of extant and often longitudinal data sets, such as huge databases on global climate measurements made over centuries, ice core sampling, and similar types of data that analysts themselves did not collect. To learn to practice science, students need to build the literacies required in such an enterprise. Ultimately, science instruction needs to reflect the authentic mix of text-based inquiry and hands-on investigation practices in which scientists engage (NGSS, 2014).

There is ample evidence that current science instruction provides few if any opportunities for students to conduct text-based investigations or engage with text (Vaughan, et al., 2013). Indeed, analyses of classroom observations of 13 science lessons in a variety of middle and high school classrooms found no examples of lessons with close reading, argumentation, nor cross-textual analysis (Litman, et al. 2017). In part, the absence of reading texts in science is related to the kinds of texts typically found in classrooms: textbooks that portray science as a set of known facts rather than a knowledge building process (Chiappetta & Fillman, 2007; Penney, Norris, Phillip & Clark, 2003). Moreover, science information is necessarily communicated in complex sentences that contain technical terminology and mathematical expressions, as well as everyday vocabulary used in highly specific ways. Visual texts of varied kinds including diagrams, graphs, data tables and models are used to communicate science information but students are rarely taught how to comprehend these texts (Fang & Schleppegrel,

2010; Lee & Sprately 2010.) In the face of such seemingly intractable texts that portray science as a known body of facts, teachers transmit orally and "power point" what they are responsible for teaching students. The result is that students have neither opportunities to engage in the reading practices of science nor use information found in texts to construct, justify, or critique explanations and models of science phenomena.

Thus the READI approach to instruction encompassed pedagogies and curricular materials to support students engaging in text-based investigations of science phenomena. Given our action theory that teachers mediate students' opportunities to learn, READI ongoing professional development was designed to create opportunities for teachers to practice and reflect on the reading, reasoning and text-based investigations at the heart of the READI approach to science. In the context of the present paper we summarize the professional development experiences of the teachers in the treatment condition and assessments of its impact as relevant to interpreting the effects on students of the READI intervention. (See Greenleaf, Brown, Litman, et al., (2016) for in depth discussion of the professional development model and its implementation.)

Intervention Design for Text-Based Investigations in Biology

The instructional activities that constituted the intervention reflect a set of design principles related to (1) selecting and sequencing science texts that reflect a range of complexity (van den Broek, 2010); (2) instructional supports to foster reading for inquiry purposes (Schoenbach, Greenleaf & Murphy, 2012); (3) instructional supports to develop and evaluate causal explanations for phenomena (Chin & Osborne, 2010; Passmore & Svoboda, 2012); and (4) discourse rich participation structures (e.g., individual reading, peer to peer text discussion, whole class discussion) to support grappling with difficult text and ideas, knowledge building and evidence-based argumentation (Von Aufschnaiter, Erduran, Osborne & Simon, 2008; Ford, 2012; Osborne, 2010).

The activities were intentionally sequenced to build reading, reasoning, and modeling practices specified in the READI Science Learning Goals (Table 1) in a systematic sequence that would build from simpler to more complex versions of the practices, progressively building skills and dispositions for student science learners to tackle complex intellectual work. The progressive sequence was based on observations from the iterative design-based research strand of READI work as well as the available research literature regarding development of the various kinds of knowledge and skills identified in the core constructs and areas of challenge for students (Greenleaf, Brown, Goldman & Ko, 2014; Greenleaf, Brown, Ko et al., 2016). For example, one consistent observation in the design work was that students needed to learn discourse norms and routines for text based, metacognitive conversations that could support sense-making, building knowledge of science, and building meta-knowledge for science reading and modeling. As well, students needed to learn about the warrants for argument in science. The instructional progression built in these threads as aspects of science literacy practice that would build over time. Figure 1 shows the progression of the intervention over the course of the single semester time frame of this efficacy study, along with the sequence of science topics and materials. The science topics closely paralleled the topics covered in the business as usual (BAU) comparison classrooms over the same time period. Note that the dates shown in the figure were rough approximations of time frames for each phase. (We emphasized that how much time on each would fluctuate for teachers and classes.)[2]

The progression was organized as four learning phases with labels that reflected the instructional and intellectual focus. For each learning phase, careful consideration was given to how each of the six READI science learning goals was being addressed. Generally speaking, a subset of goals were focal in each phase, with goals introduced in earlier phases (e.g., close reading, synthesis) used in service of goals that became the focus in later phases (e.g., construct explanatory model, justify and critique models). The progression in "goal deepening" is described in Table 2 and summarized in the discussion of each of the phases.

1. Building Classroom Routines To Support Science Literacy and Meaning Making. The purpose in this phase was to establish a classroom culture in which students engaged in and began to see close reading of text along with class-wide and critical discussion of it as accepted processes for building science knowledge. This culture emphasized the value of student thinking and ideas in the knowledge building process. Materials that were provided included a text set addressing "What is biology?," and science reading and talking stems for students, teacher talk stems for metacognitive conversations adapted from earlier work (e.g., Schoenbach, Greenleaf, & Murphy 2012). Teachers modeled the use of these stems and then encouraged their use by students so that they became more routine.
2. Building a Repertoire of Science Literacy and Discourse Processes. The emphases during this phase included the functional value of text in deepening understanding of science phenomena by attending to the kind of evidence embedded in different types of texts and the interpretations that could be drawn from them. Students considered how the evidence and interpretations were relevant to constructing explanations of particular phenomena. A particular focus was on how scientists, specifically biologists, represent entities and processes (e.g., cells and biochemical processes). Materials included teacher supplied texts on cell biology, as well as the READI *Reading Models* module (Sela, Brown, Jaureguy, Childers & Ko, 2016). The *Reading Models* module began with a text excerpted from the IQWST materials that discussed why and how scientists use models (Krajcik, Reiser, Sutherland, & Fortus 2011). The remainder of the module used an adapted version of an elicitation task (Pluta, Chinn, & Duncan, 2011) applied to representational models of biology content. Routines for reading and discussing science texts (e.g., think aloud, annotation, and metacognitive conversations) were incorporated into this module to support sense-making, evidence, and interpretation processes. The *Reading Modules* module was followed by a second READI module, *Homeostasis* (Ko, et al., 2016). This module continued into the third phase (see Fig. 1). Summaries of these modules are provided in Appendix A. Complete modules can be accessed and downloaded at www.projectreadi.org
3. Deepening Scientific Literacy And Discourse Practices For Reasoned Sensemaking. The intention of this phase is for students to use the knowledge, skills, and dispositions introduced and practiced in the first two phases for purposes of constructing a causal explanation of science phenomena. The READI *Homeostasis* module provided a sequenced set of texts and tasks that built toward a complete and coherent causal explanation. These include use of tools introduced

earlier (e.g., evidence/interpretation charts, sentence stems, metacognitive discourse routines). As well, students created and revised models as they proceeded through the module and read more deeply about the phenomena. As students actively engage in these activities there are frequent opportunities to reflect on how their thinking has changed, what questions they have, and so forth. During this phase they also moved on to the last set of science topics for the semester, genetics, heredity and evolution. The READI *MRSA* module (Brown, et al., 2016) dealing with antibiotic resistant bacteria supported inquiry into these topics. (See Appendix A for summary and www.projectreadi.org for complete module.)

4. Utilizing Scientific Literacy And Discourse Practices For Disciplinary Knowledge Building. This fourth phase was intended to deepen text-based inquiry goals by introducing justification and critique of causal explanatory models, still within the READI *MRSA* module. Instructional activities of peer to peer and class-wide discussions continued but there was also additional emphasis on independent reading to support small group model construction and peer review of the models constructed by other small groups.

In summary, the READI science progression is a framework for 'on-boarding' novice science readers into science reading practices, culminating in reading multiple science texts for evidenced based argumentation. The instructional progression attempts to reflect an iterative instructional cycle for practices of reading, reasoning, and argumentation during text-based investigations. Practices are introduced, often through modeling and explicit instruction, followed by scaffolded practice with opportunities for feedback, and ultimately to fluent grasp of the concepts and practices that reflect core constructs in the discipline.

Professional Development Design

The READI approach asks teachers to make significant shifts in their current instructional practices. Although some pedagogical shifts are amenable to highly structured, scripted materials and practices, the READI approach is not. Research on professional learning and our past work to support the type of deep instructional change called for by the READI approach underscored the need for teachers to inquire into teaching and learning (Kennedy, 2016; Greenleaf & Schoenbach, 2004), learn in ways that model targeted pedagogical approaches (Schoenbach, Greenleaf & Murphy, 2016; Loucks-Horsley, Hewson, Love & Stiles, 1998), engage in ongoing reflection (Moon, 2013), work with colleagues to translate ideas into their specific contexts (Raphael, Au, & Goldman, 2009), and have ongoing support for their learning (Cognition &Technology Group at Vanderbilt, 1997; Goldman, 2005; Greenleaf, Schoenbach, Cziko, & Mueller, 2001; Zech, Gause-Vega, Bray, Secules, & Goldman,,2000). Thus for students to have opportunities to learn to engage in evidence-based argument from multiple texts in any of the three disciplines targeted by Project READI, teachers needed opportunities to engage with the practices and ways of reading, reasoning, and arguing that they would then support in their students.

Therefore, we invested in a strand of work to develop, study, and refine inquiry designs for evidence-based argumentation through ongoing Teacher Inquiry Network activities, beginning in year one of the project in the California Teacher Inquiry Network.

These opportunities for teacher learning--enactments of evidence-based argumentation tasks in the disciplines—built on the Reading Apprenticeship professional learning model adapted to reflect Project READI learning objectives (Greenleaf, et al., 2011; Greenleaf, Brown, Litman, et al., 2016). The inquiry learning modules were iterated in the Chicago Network starting in year two, as well as in proof of concept studies carried out in year four of the project. The designed and refined teacher learning opportunities thus constituted resources to draw upon in the design and implementation of the science RCT.

The work in the first four years of the project confirmed two important principles regarding professional learning to prepare teachers to engage in the kind of complex instruction called for in READI. First, repositioning the role of the teacher as envisioned in Project READI was a gradual process that took several iterations of implementation and reflection before teachers' adaptations to suggested READI protocols and materials reflected the deep structure of the approach. Typically, the first time teachers tried many of the instructional processes they were tentative and unsure of their success. Many initial adaptations retained the form but not the substance of the principles. Debriefing with colleagues in the design teams and networks was a crucial opportunity for feedback and subsequent tuning of additional implementation efforts. Second, teachers took up the READI approach in different ways, over different time frames, and to different degrees. However, we saw evidence of change toward the envisioned target in approximately 90% of the almost 100 teachers with whom we worked over the first 4 years of the project.

These two principles were in tension with design requirements of typical randomized control trial studies and the pragmatics of the time frame and resources that could be allocated to the RCT. We refer to the design requirement of RCTs that participants have no prior history with the treatment being tested prior to random assignment to treatment or control group. This meant that we would be testing the READI approach during "first time" implementations. Pragmatically, this was dictated by the reality of the time frame of the project, including the 4 years of design and development work. In addition, given the variation that we expected in how teachers would take up and implement the READI approach, especially on a first time implementation, we needed to develop ways to look at "fidelity of implementation" that took variation in teachers prior attitudes and practices into account and that were able to capture change toward the READI approach as well as more mature implementations of the type we had seen emerge after three and four years of participation in professional learning activities associated with the Inquiry Network experiences READI provided. Given these considerations, our focus in the professional development for the RCT READI Intervention teachers was to engage them in (1) text-based investigation practices of reading, reasoning, and arguing that would be challenging for them as adult biological sciences teachers; (2) in-depth study and exploration of EBA instructional modules as educative curriculum. These two foci were intended to prepare them to understand the deep structure of the READI science approach sufficiently to achieve reasonable progress on at least the READI Learning Goals of close reading, synthesis across multiple information sources, and construction of explanatory arguments (see Table 2).

Design of the READI Science RCT professional development. We built on the professional development approach that had been developed by the Strategic Literacy Initiative in the context of a prior RCT focused on integrating literacy into high school biology (Greenleaf, et al., 2011; Schoenbach, Greenleaf, & Murphy, 2012). We adapted

that design to reflect the Project READI learning objectives and emphasis on evidence-based argumentation. Furthermore, we planned it to occur over 11 days that were distributed over a 10 - month period of time: Four days distributed over three months (February – April, 2014); 5 successive days during July; and 2 days during the Fall, 2014 implementation (6 weeks and 10 weeks into the semester). The distribution of the professional development sessions was intended to provide teachers with opportunities to enact and reflect on the pedagogical practices of the READI approach.

An overview of the professional development sessions is provided in Table 3. Over the course of the Spring and Summer sessions, teachers read various sections of *Reading for Understanding* (Schoenbach, et al., 2012) and engaged in adapted forms of many of the activities typically used in *Reading Apprenticeship* professional development. For example, Day 1 focused on engaging the teachers with close reading of science texts; in particular, in participating in the routines that they would enact to lay the ground work for and foster student engagement in science reading and learning. Participants explored how literacy has shaped their engagement with text, how the social conditions of the learning environment affected them, how they read and how they thought as scientists. They were asked to try out these routines in their classrooms in preparation for Day 2. They brought artifacts from these efforts to Day 2, discussed them and engaged in inquiry focused on engaging students in reading to construct explanations of science phenomena. Again teachers were expected to try out these routines in their classrooms and debrief at the next meeting. Similarly, on Days 3 and 4, the emphasis was on pedagogical practices for supporting text-based inquiry in science.

The five days during the summer were devoted more specifically to organizing the work of the semester–long intervention. Teachers were provided with an overview of the intervention semester that mapped out a progression of READI science reading and learning goals, READI materials appropriate to those goals, and science topics (See Figure 1.) As indicated in the figure, three READI modules – *Reading Models, Homeostasis,* and *MRSA* - were to be implemented. Corresponding to the progression shown in Figure 1, over the five days, the professional development focus moved across Building Classroom routines to support science literacy and meaning making, to Building a Repertoire of Science literacy and discourse processes, including an emphasis on science models, to Deepening scientific literacy and discourse practices for sensemaking. Thus, over the five days of professional development, they re-enacted the pedagogical practices that had been introduced during the Spring and worked through more "science-specific" inquiry activities, including opportunities to engage in argumentation in science learning and the nature of scientific models. Consistent with our view of the modules serving a dual role as student materials and educative curricula for teachers, the teachers worked through and discussed the READI candidate texts (see Figure 1) as well as the *Reading Models* module. They were introduced to the *Homeostasis* and *MRSA* modules and carried out some of the tasks embedded in these modules, engaged in formative assessment activities, and planned for instruction. Deeper examination of these READI modules occurred when teachers returned for professional development 6 (*Homeostasis)* and 10 (*MRSA)* weeks into the Fall, 2014 semester.

Throughout the PD, teachers engaged in an iterative cycle of activities in which they participated as learners and then reflected on those experiences to gain insights into the pedagogical issues they could anticipate in their classrooms. As learners, teachers

explored and analyzed their own personal processes for reading multiple texts, constructing models and explanations, and argumentation in the modeling and explanation tasks. They analyzed the pedagogy and design of the READI intervention text dependent investigations, and drawing on lessons learned from these, planned how that might use the READI pedagogies and materials with their students. As well, when they returned during the intervention semester, they engaged in cycles of reflection on practice, formative assessment of student work, and planning for how to continue to use READI pedagogies and materials with their students.

During the five-day summer professional development, teachers were also provided time to plan for their implementations. In the course of doing so as well as throughout the five days, they freely raised questions about how they would implement particular pedagogical practices and obstacles they anticipated with "their students." These obstacles ranged from limited English language skills to motivational issues to differences in achievement levels. (Our teacher sample reflected schools with different demographic characteristics and different percentages of students who met or exceeded passing scores on the Prairie State Achievement Exam.) These discussions confirmed our expectations that we were likely to see a wide range of variability in what and how the intervention would be implemented.

We also developed a six-day version of the professional learning experience for the BAU teachers. It covered all of the same topics but we reduced the time for planning how they would implement in their classrooms and did not hold the two days of PD during implementation. BAU teachers were provided with all of the same instructional resources as the READI intervention teachers. This "delayed" professional development took place after data collection for the RCT study was completed.

Methods

Design

The design was a stratified random control trial with schools as the unit of assignment. To take into account pre-existing variations in demographics and achievement levels among the schools, these characteristics were used to create six clusters (strata); randomization of schools to condition (intervention or BAU) was applied within each cluster (see Participants section). The study was conducted over a five to six month period (20 to 22 weeks of instruction), beginning with the 2014 academic year. The intervention was tested against BAU by comparing post-scores across conditions taking into account pre-scores. Multilevel modeling was used to conduct statistical tests to account for the levels of nesting present in the design (students within classrooms; classrooms within teachers; teachers within schools) to test the differences between conditions. Details of these analyses are provided in the Results section.

For teachers, dependent variables were derived from a self-report teacher survey of knowledge, attitudes and practices, as well as classroom observations of teaching practices (see Assessment section, Teachers). For students dependent variables were derived from researcher – developed assessments of evidence-based argument, self-efficacy, and science epistemology. In addition, pre- and post-intervention, students completed reading comprehension assessments that had been developed by a third-party (see Assessment section, Students).

Participants

Schools were recruited from six districts in and around a large urban area. Six stratified groups were created based on achievement levels and socioeconomic status but all served diverse populations, including English Language Learners. Achievement was indexed by the percentage of students who met or exceeded the Prairie State Achievement Exam (PSAE) the previous year. This ranged from an average of 16% in Strata 1 to 68% in Strata 6. Socioeconomic status was indexed by the percentage of students qualifying for free or reduced lunch; the mean was 83% in Strata 1 and 40% in Strata 6. It was also the case that there were suburban as well as urban districts represented in each strata. Within strata, schools were randomly assigned to Intervention or BAU comparison and all participating teachers at that school were in the assigned condition. This produced 12 schools and 24 teachers of 9th grade biology in each condition. Table 4 shows the results of the stratified random assignment of schools to condition.

Teachers taught multiple sections of this course; each section defined a classroom for purposes of the design. We consented students in two classrooms per teacher, yielding 96 classrooms (48 READI and 48 BAU).[3] After consenting students, approximately 1400 students contributed analyzable data (60% Intervention). Preliminary analyses indicated that consent rates were consistent across strata and schools within districts. Thus, the consented sample did not introduce selection bias related to strata.

Assessment Instruments

Teacher survey. We developed and administered a self-report survey that tapped teacher knowledge, attitudes, and practices related to teaching science, science literacy, and their student populations. The teacher survey consisted of 72 items, reflecting 10 constructs. The items included one item asking about the teacher's familiarity with the Common Core State Standards, three scales developed for the purposes of this study, and six subscales adapted from Greenleaf et al. (2011). All items used a 5-point Likert-type response format with all response options labeled.

1. ***Common Core familiarity.*** Teachers were asked: "*How familiar are you with the Common Core State Standards?*" on a 5-point scale ranging from 1 = *not familiar* to 5 = *extremely familiar*.

Three scales (Attitude, Self-efficacy, and Argument and Multiple Source Practices) were developed for the purposes of this study and pilot-tested during 2011-2014 with a group of teachers who did not participate in the present study. Each scale (Attitude, Self-efficacy, and Argument and Multiple Source Practices) consisted of the same set of nine items; the majority of these items dealt with argumentation and use of multiple sources of information in science (e.g., *Use multiple sources of information presented in diverse formats and media in order to develop an argument; Evaluate the credibility and reliability of a source of information)*. The complete set of nine items is presented in Appendix C. For each of these three scales, a somewhat different prompt and a different response scale were used, as follows.

2. ***Attitude.*** Teachers rated importance for students of each of the nine items on a 5-point scale ranging from 1 = *not important* to 5 = *extremely important.*

3. ***Self-efficacy****.* Teachers rated their confidence in facilitating the acquisition of the 9 skills by their students. The 5-point scale ranged from 1 = *not confident* to 5 = *extremely confident*.

4. ***Argument and multiple source practices****.* Teachers indicated how frequently they worked on each of the nine skills with their students. The 5-point scale ranged from 1 = *never* to 5 = *all or almost all lessons*.

Six scales were adapted from Greenleaf et al. (2011) and tapped various teaching practices as well as the kinds of activities in which students had opportunities to engage. Five of these six scales asked teachers to rate how frequently they engaged in particular instructional practices as well as how frequently they had their students engage in particular kinds of activities or work with particular kinds of materials. They used the same 5-point rating scheme on five of six scales: 1 = *never* to 5 = *all or almost all lessons*. The sixth scale asked for their degree of agreement with items. (1 = *strongly disagree* to 5 = *strongly agree*).

1. ***Science reading opportunities: Learning structure.*** The six items on this scale referred to the kinds of reading and reading activities in which students engaged (e.g., *Read for homework assignment, Take turns reading aloud in whole class setting, Listen to teacher read aloud in whole class setting,* and *Read self-selected science materials.* Greenleaf et al. (2011) reported the reliability $\alpha = .73$ for the six items.

2. ***Content.*** This scale consisted of five items related to reading and discussing the content to be learned in whole class and small group discussion, taking notes, discuss homework assignments. Greenleaf et al. (2011) reported reliability $\alpha = .80$ for five items.

3. ***Metacognitive inquiry: Teacher modeling.*** Teachers indicated how frequently they made their thinking visible about making sense of reading materials on their own and in discussions with others, especially when confusions arose. There were 5 items of this type. Greenleaf et al. (2011) reported reliability $\alpha = .77$ for five items.

4. ***Metacognitive inquiry: Student practice.*** This 7-item scale asked teachers to indicate how frequently their students, annotated text materials, took notes on and/or discussed with others their confusions and ways to make sense of the reading materials. Greenleaf et al. (2011) reported reliability $\alpha = .87$ for seven items.

5. ***Negotiation success: Instruction***. This 7 item scale was adapted from a 9-item Greenleaf et al., (2011) scale. Teachers rated how often they assessed students reading and provide feedback on reading assignments, journals and related work. Greenleaf et al. (2011) reported reliability $\alpha = .74$ for nine items.

6. ***Teaching philosophy: Reading***. Teachers rated their agreement with 14 items (2 were added to the 12 used by Greenleaf et al. (2011)) that tapped beliefs about teaching reading, students' reading skills and work habits, and malleability of high school students' reading achievement (e.g., *It is virtually impossible to significantly improve students' reading in secondary school; Spending class time reading runs counter to the goal of building knowledge in my discipline*). The rating scale for these items ranged from 1 = *strongly disagree* to 5 = *strongly agree.* Greenleaf et al. (2011) reported reliability $\alpha = .47$ for 12 items.

Note that in the analyses, all items were recoded so that higher values on the scale reflect beliefs more aligned with the READI approach.

Classroom Observation Protocol. We conducted two classroom observations with a 6-construct observation protocol adapted from Greenleaf et al. (2011), augmented to reflect READI science learning goals. Observations occurred between the 4th and 7th weeks of the progression (Table 2) and 8 to 10 weeks later. The observation protocol tapped six constructs as follows:

1. Construct 1: science reading opportunities (is reading central to the intellectual work or not?) (four indicators)
2. Construct 2: teacher support for student efforts to comprehend science content from text (four indicators)
3. Construct 3: metacognitive inquiry into science reading and thinking processes (three indicators)
4. Construct 4: specific reading & science reading comprehension routines, tools, strategies and processes (two indicators)
5. Construct 5: science argumentation and/or building explanatory models from multiple sources (three indicators)
6. Construct 6: collaboration (three indicators).

Observers took field notes continuously during the observation and then assigned a rubric score point to each indicator using evidence from the field notes to provide justifications for the ratings. For all indicators higher score points reflect instructional practices that are more consistent with the READI approach. Score points generally indicated frequency with which evidence of the indicator was observed (range: 1 = *rare/nev*er to 4 = *almost always*) and quality/intensity of the enactment of a practice. For example, a score point of 1 was assigned when organization of the class activities did not provide social support for reading and understanding science content. Evidence for this might be absence of time during class when pairs of students talked to each other about texts they were reading.

Observations were conducted by six members of the READI staff, all of whom were familiar with the READI intervention, including three who had been directly involved in the development of the intervention. We used an external rater who had not been involved with the intervention development to check reliability of the ratings assigned by READI staff at each of the time-1 and time-2 observations. Training to achieve consensus on the criteria for the various score points was conducted prior to the time-1 and again prior to the time-2 observations. The "training" involved each of the 7 raters independently watching a video of a science class, taking field notes, and assigning score points. The 7 then met to discuss score points and rationales for each of the indicators. Discussion of each produced consensus regarding criteria for each of the score points on each of the indicators. Different videos were used for consensus training at the two time points. The time-1 video was of a middle-school teacher implementing an early iteration of a text-based investigation on the water cycle. The video at time–2 was of a 9th grade genetics lesson that used text but the teacher had not yet been part of READI activities.

To establish interrater reliability, the external rater observed one class with each of the six READI observers, thus resulting in six pairs of observations at time-1 and six pairs of observations at time-2. The external rater was not told whether the teacher was an

intervention or BAU teacher. Percent agreement was computed for exact score point agreement and agreement within 1 score point. Average exact percent agreement was 76.4% (range 51.7 – 96.6) at time-1 and at time-2, 65.5% (range 89.7% – 51.7%). Within 1 score, at time-1, average agreement was 93.1% (range 100% - 86.2%) for time-1 and for time-2 92.5% (range 100% - 89.7%).

Student: Evidence-based Argument. The READI science and assessment team designed the Evidence-based Assessment (EBA) to closely align with the text-based inquiry intervention. We used several task types, as shown in Table 5. The text set and each of the task types were designed so that appropriate responses required reading and synthesizing information across multiple texts. The essay and multiple choice formats were intended to assess students' comprehension of the underlying explanatory model for the topic. Multiple choice items provided evidence of comprehension not confounded with students' productive writing skills. Two additional tasks, peer evaluation and model selection/justification, were included to specifically provide evidence relative to critiquing and evaluating models (READI science learning goals 4, 5). Each of these two tasks also requires making a determination of the quality and adequacy of the model and thus tap skills beyond those required by the essay and the multiple choice items.

For the pre/post design, we developed text sets on two topics (skin cancer and coral bleaching), allowing us to counterbalance within classroom so that students completed the assessment on different topics at pre and post (Goldman, Britt, Lee, Wallace & Project READI, 2016). The text sets consisted of one text that provided background information about skin cancer or coral reefs plus four texts, two of which were graphs, that contained information needed to answer a prompt that asked for an explanation of a phenomenon associated with skin cancer/coral bleaching, depending on topic. Figures 2a, b are representations of the linked network of states and events derived from the information in the text set for skin cancer (a) and for coral bleaching (b). These reflect representations of explanatory models that provide complete and coherent responses to the prompt.

We anticipated that differences in prior knowledge of these topics and potentially of general science knowledge would affect reading comprehension consistent with findings from single text reading comprehension research (e.g., Kintsch, 1994). Prior to beginning the EBA assessment reading and tasks, we administered a 6-question survey that asked students to rate how much they knew about each of the six, with 1 = *I do not know anything* and 6 = *I know a lot.* For coral bleaching the six items were coral bleaching, life science, earth science, plant cell function, oceanography, the sun. For skin cancer, the items were skin cancer, life science, earth science, the earth's coordinates, cell reproduction, and the sun. The very brief nature of the prior knowledge assessment reflected time constraints for the pre and post assessments and that we wanted to maximize the time students had to read and complete the response tasks. These ratings were used to statistically control for differences in prior knowledge when examining the effects of the READI intervention relative to BAU comparison instruction.

The task instructions prior to reading informed students that one purpose of reading in science was to understand why and how science phenomena happen. The instructions continued as follows:

> ***For skin cancer****: Today you will be reading about what causes some people to experience abnormal cell growth like skin cancer.*
> ***For coral bleaching:*** *Today you will be reading about what causes coral bleaching. Coral, which lives in the ocean, can be many different colors, but sometimes it loses its color and turns white.*
> ***For both:*** You will have to piece together important information across multiple sources to construct a good explanation of how and why this happens. No single source will provide all of the important pieces of the explanation. Instead, you are the one making connections across sources and coming up with an explanation.
>
> *Your task is to read the following set of sources to help you understand and explain*
> ***For skin cancer:*** *what leads to differences in the risk of developing skin cancer.*
> ***For coral bleaching*****:** *what leads to differences in the rates of coral bleaching.*
> ***For both****: While reading, it is important to show your thinking by making notes in the margins or on the texts.*
>
> *You will be asked to answer questions and use specific information from the sources to support your ideas and conclusions.***"**

The instructions also specified that the information sources could be read in any order but that students should read the one labeled "Background" first because it gave general information on the topic.

The instructions for the writing task, the multiple choice task, the peer essay evaluation task, and the model preference task all referenced using the information sources students had been provided with to answer the questions. Appendix E contains the complete set of task instructions for each topic and task.

External assessments developed by third-party. At the beginning of the school year, we administered the RISE, a general reading skill assessment (Sabatini, Bruce & Steinberg, 2013) to look at impact of the intervention controlling for pre-intervention basic reading skills (e.g., word recognition, efficiency, vocabulary). Post intervention, the GISA, developed specifically to tap reading for understanding using multiple texts, was administered (Sabatini & O'Reilly, 2015). For the READI 9th grade biological sciences intervention, a GISA on the topic of mitochondrial DNA was developed by ETS in consultation with READI science content experts. The GISA contains an initial assessment of prior knowledge and then taps a variety of complex comprehension skills. For example, students read a text and then use the information in the text to construct a table of attributes indicating whether they are attributes of nuclear DNA, mitochondrial DNA, both, or neither. Thus, students must reason from the information provided in the text to construct the table. Other items provide inferences and students must decide if the inference is supported by the text or not. They are also asked to read and understand the

claims of two scientific theories, the evidence that supports them, and whether and what type of additional evidence would lend greater support to each theory. The final task involved reading a short article that presented new evidence. Students are asked to decide which theory the evidence supported and why. All responses on the GISA except the justification for the theory chosen are selected response items that are scored by ETS to produce a percent correct score.

Students: Science Epistemology scale. A number of researchers have reported that epistemic cognition about the topic of a task is often a significant predictor of comprehension in multiple source reading situations (e.g., Bråten, Strømsø, & Samuelstuen, 2008; Strømsø, Bråten, & Samuelstuen, 2008). Accordingly, we developed and administered an epistemology survey specifically related to reading multiple texts in science. This survey built on prior work on both general and topic specific epistemology surveys (e.g., Hofer & Pintrich, 1997). There were a total of 18 items on the epistemology scale that loaded on two factors: seven items on the simple/certain dimension of nature of knowledge and 11 items on the need for and importance of corroborating or integrating information when using multiple sources of science information (see Salas, et al., 2015, 2016). Students endorsed the items using a scale ranging from 1 = *strongly disagree* to 6 = *strongly agree*. Sample items for each scale are the following:

Corroboration

"To understand the causes of scientific phenomena you should consider many perspectives."

"Getting information from multiple sources is important when trying to explain the causes of scientific phenomena."

"You should consider multiple explanations before accepting any explanation for scientific phenomena."

Simplicity-Certainty

"Most scientific phenomena are due to a single cause."

"The best explanations in science are those that stick to just the one major cause that most directly leads to the phenomena."

Students: Self-efficacy. A long term goal of the Project READI intervention is that students' see themselves as competent readers and learners who have the confidence to persist at tasks and with texts that challenge them. Based on Bandura's (1997) definition of perceived self-efficacy as " beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (p. 3), academic self-efficacy refers to an individual's belief that he or she can successfully perform or achieve at a designated level an academic task. We adapted an existing self-efficacy scale (Nietfeld, Cao, & Osborne, 2006) to align with the science domain. The resulting scale contained six items measuring students' confidence to learn and perform well in science (e.g., *I am sure I could do advanced work in science*). The scale employs a 5-point Likert-type response scale with option labels for the middle and both end-points 1 = *nothing like me*; 3 = *somewhat like me*; 5 = *exactly like me*.

The Science Self-efficacy scale was pilot tested on 392 adolescents with similar demographic characteristics to the students participating in this study. The results

indicated that scores from the Science Self-efficacy scale produced adequate psychometric properties. That is, the six items loaded on a single factor and explained 63.48% of the variance; factor loadings ranged from .74 to .85. Further, a Cronbach's alpha of .91 indicated good internal consistency. Initial validity was evidenced by positive correlations between the Science Self-efficacy scale and interest in science (a single-item measure with a 5-point Likert-type response scale), $r(382) = .65$, $p < .001$. On average, students' self-efficacy toward science was above the mid-point of the response scale ($M = 3.60$, $SD = 0.92$).

Procedure

Teachers. For all teachers, the "pre" assessment was completed prior to the start of professional development for the intervention teachers (early 2014) and the "post" assessment occurred at the conclusion of the classroom intervention. BAU comparison teachers post surveys were timed to coincide with the intervention teachers from schools within their same randomization strata. Teachers assigned to the READI intervention condition participated in 11 Professional Development sessions (6.5 hours) spaced over the semester prior to and during implementation (two days during Feb 2014, two days during May 2014, five days during July, 2014 and two follow-up days in the early and mid Fall 2014 semester). Teachers who had been assigned to the BAU group were provided with six days of professional development after all post-intervention assessments were concluded. They were provided with all of the materials (modules and scaffolding tools) that the intervention teachers had received.

All teachers in both conditions were observed twice during the intervention period, with the first occurring "early" in the semester (within the first 4 - 6 weeks of school) and the second occurring sometime during the last three weeks of the intervention. Timing of observations in BAU classrooms coincided with the observations of the intervention classrooms in the same strata. Across teachers, the average time between observations was 108 days ($SD = 11$, range 93 to 132).

Students. The EBA assessment along with the epistemology and self-efficacy scales were administered in paper and pencil format over two successive days during the biology class period. The epistemology survey was distributed and completed first (10 minutes). Students then completed the brief topic knowledge rating for their pretest topic. Each student then received an envelope that contained the relevant texts for their topic arranged in the same order for all students but "clipped" rather than stapled so students could easily manipulate them. They were provided with the instructions for the overall task (Appendix E) but were told that for the rest of the first class period they would be reading and annotating the texts. These were collected at the end of the period in envelopes with students' names on them. On the second day of the assessment, the envelopes from the previous day were returned to students and they were provided with a response booklet that included the instructions again plus lined paper for their essays followed by the nine multiple choice questions, the model evaluation task, and the peer essay evaluation task. Students were explicitly told they could and should refer to the texts in doing the tasks in the response booklet. The last thing students completed was the self-efficacy scale. An additional class period was used for computer – based administration of the RISE reading comprehension test. Post-intervention was similar in terms of task order and organization of the materials, except that each student received

the topic and knowledge rating task t they had not been given on the pretest. The GISA was administered via computer within two weeks of completing the EBA assessment and the other response scales.

Student "pre" data were collected within the first eight weeks of school and "post" within two weeks of concluding the intervention. Data collection in BAU classrooms was yoked to that of their "corresponding" intervention classrooms in their district and strata to the degree possible due to scheduling issues. In all but one case, the BAU classrooms completed the assessments later into the year than the intervention so that any bias introduced by when the test was taken would favor the students in the BAU comparison classrooms. For ease of instructional management, all students in each class were administered the assessments, including the RISE and the GISA. Data from nonconsented students were removed and destroyed prior to analyses. Note that school start dates ranged from mid-August to just after Labor Day depending on the district.

Scoring

Coding and scoring of constructed responses on EBA assessment: Essays. The essays were scored based on a number of factors. The key variables of interest were the number of concepts and the number of connections provided in the essay. Additional variables included the number of words in the essay, the number of initiating factors mentioned, and the number of intervening concepts in the final compiled claim. The essays were coded on a sentence-by-sentence basis. For each sentence, all identified concepts were listed as well as those concepts connected by causal language (e.g., leads to).

Interrater reliability for the scoring of the essays was established through a multi-step process. Three individuals were trained as coders for the pre and post EBA essay data. Coders were trained on materials for each topic (coral bleaching or skin cancer) using annotated versions of the five documents, causal models indicating numberings associated with concepts in the models, and spreadsheets of ideal answers and vague answers. Two individuals were each trained on only one topic, and a third coder was trained on both. The single topic coders were responsible for scoring every essay on their topic, whereas the double topic coder was responsible for scoring 20% of the essays for each topic. Thus, two coders scored 20% of the essays, and one coder scored the remaining 80%.

Training on the scoring process began with a meeting to discuss the causal models, the scoring structure, and the numberings associated with each concept code. All three coders were given practice essays from a previous round of data collection using the same science topics and tasks. Cohen's Kappa was used to establish inter-rater reliability. To do this, the 13 concept codes in the coral bleaching model and the nine concept codes in the skin cancer model were displayed vertically in a spreadsheet for each participant's essay. The cells in the adjacent columns were filled with 1s and 0s depending on whether a given concept code was included in that coder's compiled claim. A Kappa score was calculated based on these sets of 1s and 0s. The Kappa scores for the three rounds of training were .76, .84, and .94 for the coral bleaching essays and .90, .93, and .97 for the skin cancer essays.

Following this, the two single topic coders began scoring subsets of essays with each subset consisting of about $1/6^{th}$ of the total set of essays. After each subset of essays

was scored, the double topic coder randomly selected 20% of the essays to score. Kappa scores were calculated for each round of essays, and disagreements were reconciled through discussion. This allowed for consistency in scoring throughout the essay scoring time frame. The Kappa scores for the coral bleaching essay sets were .75, .89, .85, .86, .86, and .93. The Kappa scores for the skin cancer essay sets were .64, .92, .88, .89, .85, and .93.

Scoring of constructed responses on EBA assessment: Model evaluation. The justification of the model evaluation item was scored as a 1 or 0 based on a brief rubric of acceptable answers. The language in the justification of the selection of the better model had to include some variation of the language from the following options: steps, step-by-step, order, cause and effect, the way it's organized, process, chain reaction, how they connect to each other. It could also include any language about specific concepts from the documents leading to one another. Following a set of practice scorings, two coders began scoring the responses in three separate subsets. One coder scored all of the responses, while the other coder only scored 20% of the subsets of responses. The kappa scores for the three subsets of coding were .90, .92, and .91

Coding and scoring of constructed responses on EBA assessment: Peer evaluation. The peer evaluation items were scored based on six variables of interest that were either present or absent in the two essays. These six variables included: relevance (staying on topic), coherence (connecting concepts to the final outcome), completeness (stating both initiating factors), the importance of sourcing, mentioning the graph, mentioning a concept tied to the graph. For each of the two essays, the variable was either addressed or not addressed in the peer's response (e.g., the peer essay said something about both initiating factors, the peer essay did not mention one of the initiating factors). Each variable was addressed in only one of the essays; none of the six variables were addressed in both essays. Because there were two peer essays to evaluate, the scores for each of the six variables were collapsed across the two essay evaluations such that a score of 1 was given for each variable if the student correctly spoke about the variable in at least one of the essay evaluations – correctly noting that the variable was present in the essay or correctly noting that the variable was absent. Two coders were trained on scoring using a rubric of acceptable and unacceptable language to represent each of the six variables. After training, one coder scored all of the essays, and the other coder scored 5% of the essays. The second coder periodically scored a small set of evaluations resulting in kappa scores of .86, .80, and .84.

Results

Data were analyzed in several phases. Generally speaking, we examined and report descriptive statistics for a measure based on the total number of participants for whom we had data on that measure. We did simple parametric tests of differences in means to establish equivalence of teacher samples prior to READI professional development and for the student participants at the beginning of the Fall, 2014 semester. Sample size for the multilevel modeling was determined by the number of participants who had data on each of the variables that were involved in the model. Thus, the number

of participants contributing to any given analysis varies.[4] Nevertheless, the sample sizes for all analyses reported met the criteria for the specific analyses that were conducted and that are reported. We first present results for teachers and then for students.

Teachers

Surveys. The pre-survey was completed by 43 teachers (24 READI and 19 BAU). The post-survey was completed by 46 teachers (23 READI and 23 BAU). Overall, we had complete data (pre and post, all items) from 41 teachers: 23 READI and 18 BAU). All teachers reported that they were familiar with the Common Core State Standards and we do not discuss these data further.

Preliminary analyses. Exploratory Factor Analysis (EFA) and reliability analyses were conducted for each of the nine scales prior to conducting comparisons across time and teacher groups. Two sets of analyses were conducted – one on the pre-data and one on the post – using data from all of the teachers who had provided data at each time point. Summary results for the EFAs and reliability analyses are provided in Appendix D, Table 1. Of the original 72 items on the survey, 15 items were removed because their factor loadings were below .40 on the pretest. The majority of removed items (nine) were from the Teaching Philosophy: Reading scale. Two items each were removed from three of the other scales (Science reading opportunities: Learning Structure, Content, and Negotiation success). The factor loadings, variance explained, and reliabilities provide evidence of the reliability and factor validity of these nine scales.

In addition, because six of the nine scales focus on teachers' practices we explored whether they loaded on a single, higher-order factor. The results of the EFA indicated that five of the six did load on a single factor, which we dubbed Higher-order Teacher Practices. The one that did not was Science Reading Opportunities: Learning Structure. After removing this scale, the final Higher-order Teacher Practices factor consisted of five mean scale scores. Factor loadings ranged from .63 to .86, explaining 51.3% of the variance for the pre score of Higher-order Teacher Practices factor, and from .86 to .88, explaining 74.2% of the variance for the post score. Reliability estimates were .83 for pre score and .93 for post score.

Comparisons of READI and BAU teachers. Table 6 provides means, standard deviations, and independent samples *t*-test statistics for pre scale scores by condition. There were no significant differences between teachers in schools that were randomly assigned to the READI as compared to those assigned to the BAU condition. In contrast to the absence of differences at pretest, the posttest comparisons, provided in Table 7, indicate that teachers in the READI Intervention scored significantly higher than those in the BAU condition on Higher-order Teaching Practices as well as on each of its components, with large effect sizes ($1.34 < d > 2.00$). As well, READI teachers indicated that they provided a variety of science reading opportunities more frequently than the BAU teachers reported doing so, also a large effect size, $d = 1.37$. READI teachers scored higher than BAU teachers on four individual differences variables: familiarity with CCSS, attitude, self-efficacy, and teaching philosophy. However, the differences were not statistically significant and Cohen's *d* effect sizes were small, ranging from .29 to .51. The lack of significant differences on these may well be related to the short time frame of the intervention. Attitudes, self-efficacy and teaching philosophies about reading might well be expected to change more gradually than practices.

The same pattern of significant differences between READI and BAU teachers was obtained from 2-level multilevel models, in which teachers (level-1) were clustered within schools (level-2). The multilevel analyses controlled statistically for pre-scores on the scales (grand-mean centered at level 1) as well as strata to test for effects of treatment condition. Effect sizes for treatment condition ranged from 1.34 to 2.21, indicating large effects (Elliot & Sammons, 2004).

Classroom observations. All 24 teachers in each condition were observed twice and contributed data on all 6 constructs.

Preliminary analyses. As indicated earlier there were 19 indicators that observers rated based on their observations as captured in the field notes. Indicators within each construct were submitted to EFA. The extraction method used was principal axis factoring with no rotation because a single-item solution was expected for each construct. Six exploratory factor analyses, one for each construct, showed that the indicators within each construct could be combined into a single construct score for time-1 observations. A second set for time-2 showed similar results (See Appendix D Table 2a, b). Indicators within each construct explained 51.4% to 87.0% of the variance at time-1 and 61.9% to 89.1% at time-2. Factor loadings were reasonable (e.g., time-1 range = .37 - .97; time-2 range .69 - .97), and estimates of internal consistency reliability (Cronbach's alphas) were adequate, ranging from .77 to .95 at time 1 and .86 to .93 at time 2. Accordingly, six mean scores were computed, one for each construct for time 1 and again for time 2.

Comparisons of READI and BAU teachers. At both observation time points, there were significant differences between READI and BAU teachers on all six constructs, with large effect sizes. Table 8 shows the descriptives, independent sample *t* tests, and effect sizes for the rubric scores at the level of the six constructs reflected in the observation protocol. The upper panel is for the time-1 observation and the lower for the time-2 observation. READI teachers achieved higher score points on each of the constructs and the differences between the two groups of teachers increased for the second observation. This is reflected in larger effect sizes at time-2 compared to time-1. Differences at time-1 were not unexpected because the READI teachers had had nine days of the PD prior to beginning the Fall, 2014 semester. Thus, the time-1 differences indicate that READI intervention students were indeed experiencing instruction and opportunities to learn that were substantively different from what the comparison BAU students were experiencing. These differences increased as the semester progressed.

We also examined whether there were significant differences in construct scores from time-1 to time-2 observations within each group of teachers. These analyses indicated that although there were increases for all constructs among the READI teachers, only two met conventional levels of statistical significance, Construct 2: Support ($M = 2.45$ with $SD = .84$ at time-1 and $M = 2.9$ with $SD = .80$ at time-2, $t\ (23) = 2.53$, $p = .019$) and Construct 6: Collaboration ($M = 2.19$ with $SD = .71$ at time-1 and M = 2.58 with $SD = .77$ at time-2, $t\ (23) = 2.67$, $p = .014$). There were no significant differences for the BAU teachers, although scores on each construct trended lower at time-2 than at time-1 observations.

The multilevel model results for the observation data indicated that teachers within schools shared a considerable amount of variance in time-2 observation scores, as

indicated by ICCs ranging from 10.2% to 57.9%. Full models were built to explore the differences between READI and BAU teachers, after controlling for school strata and time-1 observation scores. All of the construct scores were higher for READI than for the BAU teachers. For the Argumentation construct there was a medium effect size ($\beta = .42$, $p = .031$, $ES = 0.65$) but for the other five constructs, the effect sizes were large (β ranged from 0.62 to .98, $p < .05$, ES ranged from .83 to 1.49). The effects of READI intervention were the highest for observation construct 1 – science reading opportunities.

Teachers: Conclusions. In summary, the survey data indicated significant shifts in ratings among the READI intervention teachers after the professional development and implementation experiences, the impact of which cannot be separated in the current data. The classroom observation results indicate that teaching and learning practices in READI teachers' classrooms were more aligned with the READI approach than were those of the BAU teachers. Furthermore, the differences between the two groups of teachers increased over the course of the intervention as READI teachers' practices were observed to be more aligned with those practices advocated in the READI PD and built into the educative curriculum modules. We conclude therefore that the Intervention teachers were providing students with opportunities to engage in text-based science investigations and explanatory modeling.

Students

The results for the students first address the main question of interest in this efficacy study: Do participants in the READI intervention outperform those in the comparison BAU classrooms? We addressed this question with respect to performance on the internally-developed EBA assessment as well as on the ETS - developed GISA assessment. For both the internally and third-party developed assessments, we first present descriptive statistics for the two conditions overall at pre and post, along with independent samples t-tests. We then report multilevel modeling employed to take into account the nested design of the study.

EBA assessment. Only those consented students who were present for both the two-day pre and two-day post administration of the EBA assessment were included in these analyses. The resulting sample consisted of 964 students (567 READI and 397 BAU) from 95 classrooms (48 READI and 47 BAU) in 24 schools (12 READI intervention and 12 BAU) and 48 teachers, 24 in each condition.[5]

Preliminary analyses. As described earlier, the EBA assessment consisted of several tasks that varied in terms of the READI science learning goals they were intended to assess as well as the written production demands. Preliminary analyses of the model evaluation and peer evaluation indicated that differences between READI and BAU groups were small and that the variation within each group was large. In conjunction with the fact that these items reflect comprehension plus production plus evaluation and justification, we focus instead on the measures that are "closer" to traditional comprehension measures: the multiple choice and the essay performance. Of these, the multiple choice is less dependent on written production demands than is the essay.

We also examined pre and post intervention scores on epistemology, self-efficacy, and topic prior knowledge scales in evaluating the effects of the intervention. Preliminary

analyses of the epistemology scales using exploratory factor analyses showed that two distinct factors that corresponded to the a priori 11- item Corroboration scale and 7-item Simple/Certain scale. Factor loadings for the 11 items on the Corroboration scale ranged from .41 to .60 (Cronbach's alpha = .80) for pre-scores and from .44 to .72 (Cronbach's alpha = .84) for post-scores. Factor loadings for Simple/Certain ranged from .43 to .56 (Cronbach's alpha = .70) for pre-scores and from .43 to .58 (Cronbach's alpha = .72) for post-scores. Overall, the two subscales explained 27.73% of the variance for pre-data and 33.09% for post-data. Exploratory factor analyses on the Self-Efficacy scale indicated a single factor solution. At pretest, factor loadings ranged from .63 to .76 (Cronbach's alpha = .86) and explained 50.47% of the variance; at post loadings ranged from .68 to .78 (Cronbach's alpha = .87), accounted for 54.15% of the variance.

Comparisons of READI and BAU students on EBA assessments. The descriptive statistics in Table 9, upper panel, show that at the beginning of the semester, the READI and BAU groups showed no significant differences in performance on the multiple choice or essay task measures (percentages of nodes and links). There were also no significant differences between the groups on the corroboration scale of the epistemology survey, self-efficacy or ratings of how much they knew about the topics of the EBA assessment. On the complex/uncertain scale the BAU group scored significantly higher than the READI Intervention group at the $p = .03$ level. Thus with the exception of the single epistemology scale, the randomization procedure resulted in groups that were statistically equivalent on the major measures of interest.

Post intervention, however, the descriptives in the lower panel of Table 9 indicate significantly higher performance on the multiple choice items for the READI group (56% correct) compared to the BAU group's 51% correct. In addition, post intervention, none of the scales on the surveys of epistemology, self-efficacy, or topic knowledge differed significantly between the two groups. Preliminary analyses also indicated that the two topics used for counterbalancing purposes were differentially difficult (skin cancer was easier than coral bleaching). However, the effects of topic and the interaction of topic and time of test were similar across the intervention and control conditions. Thus in conducting the multilevel modeling to evaluate the effects of the READI intervention, we statistically controlled for differences among students due to the testing time at which they had each topic.

Multilevel Modeling of Multiple Choice. We first explored what and how many levels to use in the multilevel modeling of our data. We considered and compared three models:

(a) A 3-level model, in which students nested within classrooms nested within schools;
(b) A 3-level model, in which students nested within teachers nested within schools; and
(c) A 4-level model with students nested within classrooms nested within teachers nested within schools.

To make our decision on the model, we computed and compared the intra-class correlation coefficients (ICCs) at each level for each model. The ICC is the degree of similarity at each level of the analyses, expressed as the percentage of variance accounted for by clustering students at each level in a model. Table 10 shows the variance and the ICC at each level for the two different 3-level models and the 4-level model. When all

four levels were considered, the teacher level did not add any shared variance (0.05%), indicating that similarities of students within classrooms explained all shared variance that could have been potentially attributed to the similarities of students within teachers. Likewise, comparing the two 3-level models, there is more shared variance when students are nested within classrooms (8.80%) than when students are nested within teachers (5.27%). Based on these results, as well as considerations of choosing a more parsimonious model (Raudenbush & Bryk, 2002), we decided to proceed with a three-level students-classrooms-schools model. The design effect (DEFF) value of 3.78 confirmed that we should use a multilevel approach to our data analysis of the multiple choice performance (McCoach & Adelson, 2010).[6] Results of comparisons of the ICCs for these three models on other outcome variables (e.g., essay scores, GISA) replicated the ICC comparisons for the multiple choice data. Accordingly the same 3-level model was used in analyzing each of the outcome variables.

The initial multiple regression of the full model included all student attribute variables assessed at pretest (e.g., multiple choice, two epistemology scales, self-efficacy scale, prior knowledge of the topic). Also included were variables intended to statistically control for school strata (six levels) and the difference in difficulty of the two assessment topics, the latter reflected in the inclusion of pretest topic and the topic by pretest interaction. The results for the full model indicated non-significant effects for self-efficacy and prior knowledge of the topic. These were trimmed from the model and the multiple regression was rerun.

Table 11 shows the results for the trimmed model. The trimmed model explained 18.18% of the student-level variance ,76.92% of the classroom-level variance, and 96.62% of the school-level variance. Condition was significant ($\beta = 5.71$, $p = .010$, $ES = 0.26$), showing that READI Intervention students had on average 5.7% higher scores on MC-post than BAU comparison students. Additionally, the two Epistemology scales were significant: Corroboration-pre ($\beta = 4.69$, $p < .001$, $ES = 0.29$) and Complex/Uncertain-pre ($\beta = 2.96$, $p < .001$, $ES = 0.22$). That these were significant predictors indicates that students who at the start of the semester held more sophisticated epistemological beliefs scored higher on the post-intervention multiple choice measure. More sophisticated beliefs were reflected in stronger agreement about the need and value of cross validation in constructing explanations of science phenomena and disagreement with simple, single cause explanations of science phenomena. Not surprisingly, individual differences in multiple choice performance at the start of the semester predicted performance post intervention. Note that which topic students had at pretest was also significant. However, this does not compromise the interpretation of the significant effect of Condition since the same counterbalancing scheme was used in READI and BAU classrooms.

Essay performance. The descriptives in lower panel of Table 9 indicate that although students in the READI Intervention group mentioned more nodes and links in their essays, the difference from the percentages in the BAU comparison group were not statistically significant. We used the same process as described for the multiple choice percent correct data to test the full 3-level students– within-classrooms - within – schools - model. The results of the trimmed model for the concept nodes are provided in Table 12 upper panel and for links in the lower panel. Individual differences at pretest associated with the Corroboration epistemology scale and with the Self-efficacy scale

were significant predictors of the inclusion of concept nodes in the essays. That the Corroboration scale was a predictor is consistent with the design of the text sets to require cross-text comparison and integration to determine important and relevant concepts to include in the explanatory model for each topic. As well, prior knowledge of the posttest topic predicted performance on concepts: the higher the rating, the more concepts included. The final model for nodes accounted for 21.56% of the variance at the student level, 48.77% of the variance at the classroom level, and 99.92% at the school level. We found a similar pattern with the links that were included in the essays: a nonsignificant condition effect and significant predictors among the scales completed at the beginning of the semester. The trimmed model for links accounted for 7.04% of the variance at the student level, 39.6% at the classroom level, and 99.99% at the school level. [7]

We also examined a composite comprehension score using the multiple choice post along with the number of nodes and link connections students mentioned in their essays. The composite score was based on the results of a factor analysis that indicated that these three measures, but not scores on the other two (model and peer essay evaluation) clustered on a single factor. The multilevel model analyses on the composite score showed the same pattern as the multiple choice post: a significant condition effect, $p = .029$ but a slightly smaller effect size (ES = 0.18). We attribute this reduction in the effect size to the production demands of the essay task and the relative lack of instructional emphasis and supports for writing explanations of science phenomena.

Third-party assessment of reading comprehension. As indicated earlier, students completed two computer-administered assessments that had been developed by a third party. The RISE administered at the beginning of the year, indicated that basic comprehension was not significantly different between the READI and BAU groups: For READI, $M = 271.37$ (SD=13.43), $N = 671$ and for BAU, $M = 270.09$ (SD=13.94), $N = 540$. RISE scores were used as a covariate in examining the effect of treatment condition on the GISA outcome comprehension score.

Following the Intervention, descriptive statistics for percent correct out of total items on the GISA indicated higher performance for the READI Intervention group ($M = 55.93$ ($SD = 16.84$), $N = 642$) compared to the BAU comparison group ($M = 52.79$, ($SD = 17.31$), $N = 452$). The 3-level model with students nested in classrooms within schools was again the most appropriate to employ based on the previously discussed criteria.[8] To test whether performance in the READI Intervention was significantly different from the BAU comparison full models were built, controlling for school strata and covarying beginning of year scores on the RISE, the two scales on the epistemology survey, and self-efficacy (each grand-mean centered). The full model allowed random variation of intercepts at all three levels, and random variation of the level-2 slopes. Results (see Table 13) showed that treatment condition emerged significant ($\beta = 4.41$, $p = .038$, $ES = 0.32$) with READI students scoring significantly higher on GISA than BAU students. The effect size of 0.32 was small from a statistical point of view. However, estimates of the magnitude of change associated with one year of reading growth at the high school level are .19 (Hill, Bloom, Black, & Lipsey, 2008). This indicates that the READI students were about 1.5 years ahead of the comparison BAU students at the end of the study.

Discussion

The READI approach to text-based investigation produced performance benefits over typical classroom instruction in 9th grade biological sciences on both highly aligned assessments and on third-party developed assessment of comprehension of science material across multiple sources. It is important to note that the topics on the assessments were related to biological sciences but were not part of the curriculum of either the READI or the BAU classrooms. Thus, all students were being asked to explain phenomena on which they had not received instruction. Although effect sizes were larger on the highly aligned, researcher - developed assessment, the effects were significant for both researcher - and third party – developed assessments.

Within the research-developed assessments, effects were strongest on the multiple choice task. Of the four task types within the EBA assessment, the multiple choice is perhaps the "purest" test of comprehension within and across texts in the multiple text sets provided for each topic. The essay task involves a heavy writing production component that may have interfered with students being able to demonstrate all that they had understood. The READI intervention did not emphasize rhetorical forms for expressing explanatory models, although students did spend instructional time working together in small groups to produce models (homeostasis and MRSA modules). The model and peer essay evaluation tasks required that students not only comprehend but then invoke evaluative criteria for models and for written explanations of models. These learning goals were introduced during the semester but given the limited instructional time that was devoted to them, it is not surprising that performance on these items did not differ from that of the BAU students. Thus, our interpretation of the generally weaker and nonsignificant findings for the essay data as well as for the two evaluation tasks is that the outcome that instruction really emphasized was achieved, namely close reading of texts for purposes of understanding key content ideas and how they might be synthesized and connected to explain a phenomenon in the physical world. However, students need additional instruction and support to express their ideas in independently produced written essays and to develop criteria and language frames for writing critiques of representations (pictorial models or verbal representations) produced by others.

It was also the case that trends in epistemic cognition for science as well as confidence in reading to learn science did not show significantly different degrees of change for READI and BAU students. However, the trends suggested that among the READI students there was a trend toward an increase in agreement with the need to corroborate science information while the trend in the BAU students' data suggested less agreement with the need to corroborate. Confidence showed similar trends in both groups, although among the BAU students the trend suggested a greater decrease in confidence than the trend for the READI students.

Our theory of change posited that teachers determine what students have opportunities to learn in their formal schooling. Hence, we constructed professional development to insure that teachers understood and had opportunities to engage in the practices and explore materials and instructional strategies they might use in creating these kinds of opportunities for their students. Teachers in the present RCT efficacy study were teaching with the READI approach for the first time. The READI approach calls for significant shifts in positioning of texts, tasks, and students' roles as agentive learners in

the classroom. Other research indicates that to make such shifts in practice, it typically takes multiple iterations of teachers trying out new practices, reflecting on "how it went" and revising for the next classroom iteration (Ball & Cohen, 1999; Joyce & Showers, 1982). In-classroom coaching supplementing "out of the classroom" professional development experiences as well as opportunities to reflect and revise with colleagues can facilitate adaptive shifts in practice (Cochran-Smith & Lytle, 1999; Darling-Hammond & McLaughlin, 1995) Despite the short duration of the READI PD, the first time implementation and the absence of in-classroom coaching, at the end of the intervention implementation, READI teachers reported practices significantly more aligned with those called for in the READI approach when compared to their own ratings at the start of the professional development and in comparison to the ratings of the BAU teachers. These differences cannot be attributed to pre-existing differences because prior to the start of the READI professional development, self-reported teaching practices were indistinguishable for teachers randomly assigned to the READI intervention condition versus the BAU comparison condition. Classroom observations both validated the self-reports of the READI and BAU teachers and showed that over the course of the semester-long intervention observable practices and instructional routines in the classrooms of the READI teachers were more aligned with those central to the READI approach than they were at the beginning of the semester. Thus the self-report and classroom observations indicate that the READI teachers shifted their practice to be more aligned with the READI approach and its emphasis on socially supported reading, reasoning and argument based on information presented in multiple information resources. Thus, the present study demonstrates that such shifts in practice are possible, can be accomplished to some degree within what was basically a one-year time frame for this RCT. We caution however that in all of the READI classrooms, teachers and students were just getting started with this type of instruction and learning. Additional opportunities for PD and classroom experiences of this type are likely needed to more firmly establish these practices of teaching and learning.

It is also interesting that the multi-level modeling of the student performance indicated that more variance in outcomes was associated with the classroom level of clustering than with the teacher level of clustering. This finding suggests that the types of changes in instructional practices called for by the READI approach require changes in the classroom culture – in the expectations and responsibilities of both teachers and students. That is, teachers and students constitute a sense-making system that is dynamic and interactive. Outcomes are related to the character of the social and intellectual interactions in that system. These in turn rely on a community in which both teachers and students are respected and valued for what they bring and contribute to the learning environment.

The positioning of this RCT of the READI approach in biological sciences was necessitated by the need to recruit a sufficient sample size of schools and teachers to achieve sufficient power to detect an effect. As noted, across grades 6 – 12 and history, literary reading/literature, and the various sciences, 9th grade biological sciences for 9th graders was the only grade level and subject area where this was possible. However, this resulted in a semester-long curriculum that engaged students in sense-making from text, where *text* is defined as the multiple forms of representation in which biological information is expressed. This sense-making involved many of the NGSS science

practices, including asking questions, developing models, interpreting data, constructing explanations, and arguing from evidence. Typically text and representation is positioned solely in practice 8, Obtaining, evaluating, and communicating information. However the present study demonstrates the efficacy of text-based investigations for involving students in almost all of the practices of science. Of course, this may be a function of the particular subdiscipline of biological sciences. Nevertheless, there are many topics and phenomena in the sciences where it is simply not feasible and in some cases not possible for students to engage directly with the phenomenon. Thus, we argue for the repositioning of texts in the practices of science and the implementation of the NGSS and other practice-based approaches to science teaching, learning, and assessment.

References

Achieve (2013). N*ext generation science standards.* Washington, DC: National Academies Press.

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

Ball, D., & Cohen, D. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3–32). San Francisco: Jossey-Bass.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* NY, NY: Freeman.

Bazerman, C. (1985). Physicists reading physics: Schema-laden purposes and purpose-laden schema. *Written Communication, 2,* 3-23.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, *93*, 26-55.

Bråten, I., Strømsø, H.I., & Samuelstuen, M.S. (2008). Are sophisticated students always better? The role of topic-specific personal epistemology in the understanding of multiple expository texts. *Contemporary Educational Psychology, 33,* 814-840.

Bromme, R. & Goldman, S. R., (2014). The public's bounded understanding of science. *Educational Psychologist, 49,* 59-69.

Brown, W., Ko, M., Greenleaf, C., Sexton, U., George, M., Goldman, S. R., with CA Teacher Inquiry Network Science Teachers. (2016). *Life sciences: The spread of MRSA. High school, grade 9, spring 2013, RCT fall 2014*. Project READI Curriculum Module Technical Report CM #27. Retrieved from URL: www.projectreadi.org

Cavagnetto, A. (2010). Argument to foster scientific literacy: A review of argument interventions in K–12 science contexts. *Review of Educational Research, 80,* 336-371.

Chiappetta, E. L., Fillman, D. A. (2007). Analysis of five high school biology textbooks used in the United States for inclusion of the nature of science. *International Journal of Science Education, 29*, 1847-1868.

Chin, C., & Osborne, J. (2010). Supporting Argumentation Through Students' Questions: Case Studies in Science Classrooms. *Journal of the Learning Sciences*, *19*(2), 230–284.

Cobb, P., diSessa, A., Lehrer, R., Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13.

Cochran-Smith, M. , & Lytle, S. L. (1999). Relationships of knowledge and practice: Teacher learning in communities. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education (Vol. 24, pp. 249-305*). Washington, DC: American Educational Research Association.

Cognition and Technology Group at Vanderbilt. (1997). *The Jasper Project: Lessons in curriculum, instruction, assessment and professional development.* Mahwah, NJ: Erlbaum.

Council of Chief State School Officers (CCSSO) (2010). *The Common Core Standards for English Language Arts and Literacy in History/Social Studies and Science and Technical Subjects.* Downloaded from http://www.corestandards.org.

Darling-Hammond, L., & McLaughlin, M. W. (1995). Policies that support professional development in an era of reform. *Phi Delta Kappan, 76(8), 597–604.*
Davis, E.A. & Krajcik, J.S, (2005) Designing educative curriculum materials to promote teacher learning. *Educational Researcher, 34(3*), 3 -14.
Elliot, K., & Sammons, P. (2004). Exploring the use of effect sizes to evaluate the impact of different influences on child outcomes: Possibilities and limitations. In I. Schagen & K. Elliot (Eds.), *But what does it mean? The use of effect sizes in educational research.* (pp. 6-24). Slough: National Foundation for Educational Research. Retrieved from https://www.nfer.ac.uk/publications/SEF01/SEF01.pdf.
Fang, Z., & Schleppegrell, M. J. (2010). Disciplinary literacies across content areas: Supporting secondary reading through functional language analysis. *Journal of Adolescent & Adult Literacy*, *53*, 587–597.
Ford, M.J. (2012). A dialogic account of sense-making in scientific argumentation and reasoning. *Cognition and Instruction, 30*(3), 207-245.
Gee, J. P., (1992). *The social mind: Language, ideology, and social practice.* NY, NY: Bergin and Garvey.
Goldman, S. R. (2012). Adolescent literacy: Learning and understanding content. *Future of Children, 22*, 89–116.
Goldman, S. R. (2005). Designing for scalable educational improvement. In C. Dede, J. P. Honan, & L. C. Peters (Eds.), *Scaling up success: Lessons learned from technology-based educational improvement* (pp. 67-96). San Francisco, CA: Josey Bass.
Goldman, S.R. (2004). Cognitive aspects of constructing meaning through and across multiple texts. In N. Shuart-Ferris & D.M. Bloome (Eds.), *Uses of intertextuality in classroom and educational research.* (pp. 313-347).Greenwich, CT: Information Age Publishing.
Goldman, S. R. & Bisanz, G. (2002). Toward a functional analysis of scientific genres. In J. Otero, J. A. León, & A. C. Graesser (Eds.), *The Psychology of Science Text Comprehension,* (pp. 19-50). Mahwah, NJ: Routledge.
Goldman, S. R., Britt, M. A., Brown, W., Cribb, G., George, M., Greenleaf, C., Lee, C. D., Shanahan, C. & Project READI. (2016). Disciplinary literacies and learning to read for understanding: A conceptual framework for disciplinary literacy. *Educational Psychologist*, 51(2), 219-246.
Goldman, S. R., Britt, M. A., Lee, C. D., Wallace, P. & Project READI. (2016). *Assessments of evidence-based argument in three disciplines: History, science and literature.* Project READI Technical Report #10. Retrieved from URL: projectreadi.org
Goldman, S. R., Ko, M., Greenleaf, C. & Brown, W. (in press). Domain-specificity in the practices of explanation, modeling, and argument in the sciences. To appear in F. Fischer, C. Chinn, K. Engelmann, & J. Osborne, (Eds.) *Scientific Reasoning and Argumentation: Domain-Specific and Domain-General Aspects*. NY, NY: Taylor Francis.
Greenleaf, C., Brown, W., Goldman, S. R., & Ko, M. (2014). *READI for science: Promoting scientific literacy practices through text-based investigations for middle and high school science teachers and students.* Washington, D.C.: National Research

Council. Available at http://sites.nationalacademies.org/DBASSE/BOSE/DBASSE_085962 [January 2014].
Greenleaf, C., Brown, W., Ko, M., Hale, G., Sexton, U., James, K. & George, M. (2016). *Updated Design Rationale, Learning Goals, and Hypothesized Progressions for Text-Based Investigations in Middle and High School Science Classrooms.* Project READI Technical Report #25. Retrieved from URL: www.projectreadi.org
Greenleaf, C., Brown, W., Litman, C. Charney-Sirott, I., Cribb, G. & Jensen, R. (2016). *Iterative design of professional development to impact teacher knowledge, beliefs, and practices.* Project READI Technical Report #24. Retrieved from URL: projectreadi.org
Greenleaf, C.; Litman, C.; Hanson, T.; Rosen, R.; Boscardin, C. K.; Herman, J.; Schneider, S.; with Madden, S. & Jones, B. (2011). Integrating literacy and science in biology: Teaching and learning impacts of Reading Apprenticeship professional development. *American Educational Research Journal, 48*, pp. 647 – 717.
Greenleaf, C. & Schoenbach, R. (2004) Building capacity for the responsive teaching of reading in the academic disciplines: Strategic inquiry designs for middle and high school teachers' professional development. In D. S. Strickland & M. L. Kamil, (Eds.), *Improving Reading Achievement through Professional Development*, Christopher-Gordon Publishers, Inc., pp. 97 – 127.
Greenleaf, C., Schoenbach, R., Cziko, C. & Mueller, F. (*2001*) Apprenticing adolescent readers to academic literacy. *Harvard Educational Review*: April 2001, Vol. 71, No. 1, pp. 79-130.
Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177.
Hofer, B.K., & Pintrich, P.R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning *Review of Educational Research*. *67,* 88-140.
Joyce, B., & Showers, B. (2002). *Student achievement through staff development* (3rd ed.). Alexandria, VA: ASCD.
Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research*. Retrieved from: http://rer.sagepub.com/content/early/2016/01/29/0034654315626800.abstract
Kintsch, W. (1994). The psychology of discourse processing. In M. A. Gernsbacher. (Ed.) Handbook of Psycholinguistics: 721-739, Academic Press: San Diego, CA.
Ko, M., Sarna, J., Stites, J., Goldman, S. R., Brown, W., James, K., & Greenleaf, C. (2016). *Life sciences: Homeostasis, high school, 9th grade.* Project READI Curriculum Module Technical Report CM #28. Retrieved from URL: projectreadi.org
Krajcik, J., Reiser., B., Sutherland, L., & Fortus, D. (2011). *IQWST: Investigating and questioning our world through science and technology (middle school science curriculum materials).* Greenwich, CT: Sangari Active Science.
Kress, G. (1989). *Linguistic processes in sociocultural practice* (2nd ed.). Oxford: Oxford University Press.
Kress, G., & Van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. London, UK: Edward Arnold.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. NY, NY: Cambridge University Press.
Lee, C. D., & Spratley, A. (2010). *Reading in the disciplines: The challenges of adolescent literacy.* New York, NY: Carnegie Corporation of New York.
Lemke, J. L. (1998). Multiplying meaning: Visual and verbal semiotics in scientific text. In J.R. Martin & R. Veel (Eds.), *Reading Science* (pp.87-113). London: Routledge.
Litman, C., Marple, S., Greenleaf, C., Charney-Sirott, I., Bolz, M., Richardson, L, Hall, A., George, M., & Goldman, S.R. (2017). Text-based argumentation with multiple sources: A descriptive study of opportunity to learn in secondary English language arts, history and science. *Journal of the Learning Sciences,* 26, 79-130.
Loucks-Horsley, S., Hewson, P. W., Love, N., & Stiles, K. E. (1998). *Designing professional development for teachers of science and mathematics*. Thousand Oaks, CA: Corwin Press.
McCoach, D. B., & Adelson, J. L. (2010). Dealing with dependence (Part1): Understanding the effects of clustered data. *Gifted Child Quarterly, 54*(2), 152-155.
McNeill, K. L., & Krajcik, J. S. (2011). *Supporting Grade 5-8 Students in Constructing Explanations in Science: The Claim, Evidence, and Reasoning Framework for Talk and Writing.* NY, NY*:* Pearson.
Moje, E. B. (2008). Foregrounding the disciplines in secondary literacy teaching and learning: A call for change. *Journal of Adolescent & Adult Literacy, 52*, 96-107.
Moje, E. B., & O'Brien, D. G. (Eds.). (2001). *Constructions of literacy: Studies of teaching and learning in and out of secondary classrooms*. Mahwah, NJ: Erlbaum.
Moon, J. A. (2013). *Reflection in learning and professional development: Theory and practice*. NY, NY: Routledge.
National Assessment of Educational Progress (2009a). *NAEP 2008 Trends in Academic Progress (NCES 2009–479).* Prepared by Rampey, B.D., Dion, G.S., and Donahue, P.L. for the National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.
National Assessment of Educational Progress (2009b). *Achievement Gaps: How Black and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress, (NCES 2009-455).* Prepared by Vanneman, A., Hamilton, L., Baldwin Anderson, J., and Rahman, T. for the National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.
National Center for Educational Statistics (2012). *The Nation's Report Card: Science 2011* (NCES 2012–465). Washington, DC: Institute of Education Sciences, U.S. Department of Education.
National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century.* Committee on Defining Deeper Learning and 21st Century Skills, J. W. Pellegrino and M. L. Hilton, Editors. Washington, DC: The National Academies Press.
New London Group (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review, 66*, 60-92.
Next Generation Science Standards Lead States (2013). *Next Generation Science Standards: For States, by States.* Washington, DC: National Academies Press.
Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring

exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition Learning*, 1, 159-179.
Norris, S. P. & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, *87,* 224–240.
Organization of Economic and Cultural Development (2013). *PISA 2012: Results i focus.* Paris: OECD.
Osborne, J. (2002). Science without literacy: A ship without a sail? *Cambridge Journal of Education*, *32*(2), 203-218.
Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science Magazine (328)*, pp. 463 – 467.
Osborne, J. F., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? *Science Education*, *95*, 627-638.
Passmore, C. M., & Svoboda, J. (2012). Exploring opportunities for argumentation in modeling classrooms. *International Journal of Science Education*, *34*, 1535-1554.
Pearson, P.D., Moje, E.B., & Greenleaf, C. (2010).“Science and literacy: Each in the service of the other.” *Science Magazine (328)*, 459-463.
Penney, K., Norris, S. P., Phillips, L. M., & Clark, G. (2003). The anatomy of junior high school science textbooks: An analysis of textual characteristics and a comparison to media reports of science. *Canadian Journal of Science, Mathematics and Technology Education, 3, 415–436.*
Pluta, W. J., Chinn, C. A., & Duncan, R. G. (2011). Learners’ epistemic criteria for good scientific models. *Journal of Research in Science Teaching*, *48*, 486–511.
RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Prepared for the Office of Educational Research and Improvement. Santa Monica, CA: RAND.
Raphael, T., Au, K., & Goldman, S. R. (2009). Whole school instructional improvement through the Standards Based Change Process: A developmental model. In J. Hoffman and Y. Goodman (Eds.), *Changing literacies for changing times* (pp. 198-229). New York, NY: Routledge/ Taylor Frances Group.
Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
Reinking, D., & Bradley, B. A. (2008). *On formative and design experiments: Approaches to language and literacy research*. New York: Teachers College Press.
Rouet, J-F. & Britt, M.A. (2011). Relevance processes in multiple document comprehension. In M.T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Relevance instructions and goal-focusing in text learning* (pp. 19 - 52). Greenwich, CT: Information Age Publishing.
Sabatini, J., Bruce, K., & Steinberg, J. (2013). *Research Report, ETS RR-13-08: SARA Reading Components Tests, RISE Form: Test Design and Technical Adequacy.* Princeton, New Jersey: Educational Testing Service.
Sabatini, J. & O’Reilly T, (2015). *Is the Moon a Satellite? “No, it is a Big Piece of Rock. It’s a Moon!” Examining Scientific Reasoning in Elementary Students’ Performance on Scenario-Based Assessments.* Presentation at Society for Text & Discourse, Minneapolis, MN.
Salas, C., Griffin, T. D., Wiley, J., Britt, M. A., Blaum, D., & Wallace, P. (2015). *Validation of New Epistemological Scales Related to Inquiry Learning.* Presentation

at Society for Text & Discourse, Minneapolis, MN.
Salas, C., Griffin, T., Wiley, J., Britt, M. A., Blaum D., & Wallace, P. (2016). *Validation of new epistemological scales related to inquiry learning*. Project READI Technical Report #6. Retrieved from: www.projectreadi.org
Schoenbach, R., Greenleaf, C., & Murphy, L. (2012). *Reading for understanding: How Reading Apprenticeship How Reading Apprenticeship Improves Disciplinary Learning in Secondary and College Classrooms*, 2nd Edition. SF: Jossey-Bass, Inc.
Schoenbach, R., Greenleaf, C., & Murphy, L. (2016). Leading for Literacy: A Reading Apprenticeship Approach. Jossey-Bass. Retrieved from http://www.wiley.com/WileyCDA/WileyTitle/productCd-1118437268.html
Sela, R., Brown, W., Jaureguy, A., Childers, E., & Ko, M. (2016). *Reading science models, high school, grade 9*. Project READI Curriculum Module Technical Report CM #29. Retrieved from URL: www.projectreadi.org.
Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content area literacy. *Harvard Educational Review, 78*(1), 40-59.
Snow, C. E., & Biancarosa, G. (2003). *Adolescent literacy and the achievement gap: What do we know and where do we go from here?* Adolescent Literacy Funders Meeting Report. NY: Carnegie Corporation.
Strømsø, H.I., Bråten, I. & Samuelstuen, M.S. (2008). Dimensions of topic-specific epistemological beliefs as predictors of multiple text understanding. *Learning and Instruction, 18,* 513-527.
Toulmin, S. E. (1958) *The uses of argument.* Cambridge, UK: Cambridge University Press.
Toulmin, S., Rieke, R., & Janik, A. (1984). *An introduction to reasoning*. NY, NY: Macmillan Publishing Company.
van den Broek, P. (2010). Using texts in science education: Cognitive processes and knowledge representation. *Science*, *328*, 453–456.
Vaughn, S., Swanson, E.A., Roberts, G., Wanzek, J., Stillman-Spisak, S.J., Solis, M., & Simmons, D. (2013). Improving reading comprehension and social studies knowledge in middle school. *Reading Research Quarterly*, 48(1), 77–93. doi:10.1002/rrq.039.
Von Aufschnaiter, C., Erduran, S., Osborne, J., & Simon, S. (2008). Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching*, *45*, 101-131.
Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model‐based inquiry as a new paradigm of preference for school science investigations. *Science education*, *92*(5), 941-967.
Yore, Larry D. (2004). "Why do future scientists need to study the language arts?" In E. W. Saul (Ed.), *Crossing borders in literacy and science instruction: Perspectives on theory and practice* (pp. 71–94). Newark, DE: International Reading Association
Yore, L. D., Bisanz, G. L., & Hand, B. M. (2003). Examining the literacy component of science literacy: 25 years of language arts and science research. *International Journal of Science Education, 25,* 689–725.
Zech, L. K., Gause-Vega, C. L., Bray, M. H., Secules, T., & Goldman, S. R. (2000). Content based collaborative inquiry: A professional development model for sustaining educational reform. Educational Psychologist, 35(3), 207–217.

Endnotes

[1] Of course, reasoning practices operate on entities and relationships among them.
[2] In the geographical area in which the study was conducted biological sciences at the 9th grade was the only topic area in science, literature, or history that was taught with enough consistency in topic coverage and grade level to achieve sample sizes of schools and teachers needed to achieve sufficient power in a randomized controlled efficacy study.
[3] One teacher taught only one section. Due to small class size in the sections of another teacher, we consented students in three of her classrooms. If a teacher had more than two sections, we randomly selected which two for consenting.
[4] Although all of the EBA assessments were proctored by READI staff, and students were instructed to work through the entire response booklet, staff reported that especially at post-test they observed students skipping around in the booklets.
[5] One class with few consented students had no consented students who were present for all 4 days of the pre/post assessments.
[6] The DEFF is "design effect." When the DEFF > 1, the standard error in the simple regression model is underestimated, meaning a multilevel model should be used. Typically, a criterion of DEFF ≥ 2 is used to indicate that a multilevel approach is suitable and needed. We computed the DEFF based on a combined ICC for school and classroom (30.4%), and an average cluster size observed in classrooms. DEFF = 1+ (average cluster size – 1)*ICC = 1 + (10.15 – 1) * .3043 = 3.78.
[7] The high levels of explained variance at level-3 in the cases of nodes and links are likely due to differences on these variables across strata.
[8] ICCs indicated significant variance in percent correct scores was accounted for by student clustering at the classroom (16.8%) and school levels (additional 18.0%).

Appendix A
READI Science Modules Used in the Intervention Classrooms

The Project READI curriculum materials focused on high school biology and consisted of two text sets, instructional tools, an investigation of science models and two text-based investigations of science phenomena linked to the instructional progression. The modules are intended to support student growth across a sequence of learning activities within and across modules to address the READI learning goals targeting science literacy and inquiry practices as well as epistemological understanding. Reading models, Homeostasis, MRSA

Reading Science Models

To develop the Reading Science Models, we reviewed the emerging literature from the field focused on argumentation as well as the work focused on building empirical learning progressions for science explanation and science modeling. As a result of that reading, we were introduced to an elicitation task for students' understanding of science models that was developed by Pluta, Chinn, and Duncan(Pluta, Chinn, & Duncan, 2011). We requested and received permission to use these elicitation materials in our science design work. READI science design team members analyzed the model criteria elicitation materials used by Pluta, Chinn, and Duncan (2011) to find commonalities with our design principles and core constructs and to identify gaps that module development would need to address.

As a result of that analysis, and to address the identified needs of teachers and students from our early implementation, we augmented the model criteria elicitation task to focus more explicitly on inquiry with science texts, rather than assessment of student ideas alone. Thus, we introduced pedagogical tools and practices for engaging students in close reading of a variety of science texts, including the varied models in the elicitation materials, through teacher and student Think Aloud and discussion routines, in sync with our other science design work and modules and drawing on our prior work (Cynthia Greenleaf, Gina Hale, Irisa Charney-Sirott, Ruth Schoenbach, 2007; Schoenbach et al., 2012). To this end, we searched for an authoritative text on why and how scientists use models and received permission from Reiser to excerpt a text from the IQWST materials (J. Krajcik, B. Reiser, L. Sutherland, & D. Fortus, 2011). This early unit is designed to put in place early scaffolds for reading science texts (Think Aloud and metacognitive conversations) as well as discourse routines to support ongoing sense-making, and ultimately, argumentation.

The Reading Models module supports students' close reading of science visuals and models germane to high school biology while building knowledge about the conventions of scientific models and the criteria for evaluating them. The module is comprised of a set ten texts within a notebook with integrated routines, scaffolds, tasks and an annotated teacher guide. The initial tasks engage students in reading and discussion of "Explanatory Models in Science," an authoritative text about science models (The Board of Regents of the University of Wisconsin System, 2002). The subsequent tasks engage students in reading and discussion of nine visuals and visual models. The discussions include metacognitive conversations about the reading and

sense making processes to build knowledge of the conventions of science models. To support knowledge building of the model criteria, the module engages students in an evaluation tasks requiring application of science model criteria. These tasks are adapted from an elicitation task for student understanding of science models developed by Pluta, Chinn, and Duncan (2011).

Homeostasis Investigation

The Homeostasis module supports science reading and introduces the use of texts as tools for constructing, revising and defending explanatory models of phenomena that relate to disciplinary core ideas in Biology. The Homeostasis Investigation focuses on two examples of homeostasis in a human body: the balance of sodium and sugar concentrations in blood serum. In line with our design principles, the module includes multiple texts of multiple modalities (e.g. clinical studies of hypernatremic patients, news stories, diagrams from Biology textbooks, and texts from the New York Times and The New Yorker). The module includes explicit modeling and explanation tasks, peer review with scientific argumentation for the modeling and explanations tasks. Each of these tasks requires that students use information from texts in the text set to develop, refine, and apply explanatory models to related phenomena around the disciplinary core idea of Homeostasis. The Homeostasis Investigation is linked to the third learning phase, *deepening scientific literacy and discourse practices for reasoned sensemaking*, in which multiple text synthesis and modeling are introduced. The intention of this phase is for students to use the knowledge, skills, and dispositions introduced and practiced in the first two phases for purposes of constructing a causal explanation of science phenomena.

We collaborated with two high school biology teachers to select human homeostasis as a topic for the module, based on the college readiness frameworks (Annenberg Institute for School Reform et al., 2014), the Common Core State Standards (Council of Chief State School Officers, 2010), the Chicago Public School Biology content frameworks and the Next Generation Science Standards (NGSS Lead States, 2013). During our design-team meetings, we brainstormed the breadth and depth of scientific principles that we wish to target. In the subsequent meetings, we brought text candidates and discussed potential sequencing possibilities, based on the causal models that were created for each phenomenon and identified affordances each text provided for building a repertoire of close reading practices. Next, the two design teachers used these texts with their students, which informed the final set of meetings. We debriefed with teachers and probed for how these texts were used in the classroom, its affordances for close reading and knowledge building, and the kinds of texts and tasks that would support these sense-making discussions.

Through this iterative process, we decided to focus on 2 cases that exemplify how feedback within and between organ systems maintains homeostasis in the human body. The first half of the module focuses the maintenance of sodium ion levels in the blood – both cases of when the balance is in maintained and when it is disrupted (hypo- and hypernatremia). The second half of the module focuses on

how the body maintains appropriate blood sugar levels, and cases when this balance is disrupted (e.g. Diabetes). The Homeostasis module text set includes both texts that are specific to mechanisms that govern salt and sugar balance, as well as more generalized texts that describe on the principles of human homeostasis.

After selecting these two phenomena as the central focus of the Homeostasis module, research members of the design team studied the phenomena in detail (consulting with multiple reputable online and textbook sources) to generate causal model that would accurately describe and explain the feedback mechanisms that regulate salt and sugar balance in the human body. These explanatory models were then brought to the teacher partners and discussed and revised for accuracy and simplicity. Through this discussion, we also identified the critical features of these explanatory models that would be set as targets for students learning, as well as the features developmentally or instructionally inappropriate for 9th grade high school Biology students. These complex, evidence-based models served as the guideline for text-selection process, helping us determine the affordances of a given text for scientific knowledge building. We simultaneously evaluated whether or not texts afforded opportunities to engage students in discussions of the close reading practices and variety of texts representative of those encountered in science.

The MRSA Investigation

The MRSA module supports students' science reading and engagement in modeling, explanation, and argumentation to build knowledge of core science concepts. The topic of methicillin resistant *Staphylococcus Aureus* (MRSA) affords the opportunity to learn natural selection, adaptation, and human impact on evolution. As such, it involves cross cutting concepts (NGSS Lead States, 2013) such as cause and effect; systems and interactions and, to a degree, scale, proportion, and quantity. MRSA also offers direct relevance to students since adolescents are at increased risk for contracting MRSA, and entails sufficient complexity to foster questioning, argumentation, and modeling. The MRSA Text-Based Investigation consists of four sections: MRSA infection, transmission and spread of MRSA, evolution from SA to MRSA, and managing the public health challenge of MRSA which build on each other conceptually and in literacy practices. Each section engages students in reading multiple texts and in science argumentation. The first three sections engage students in developing an explanatory model for MRSA and in argumentation about their models. The final section focuses on designing interventions based on the models and in argumentation about their interventions. Throughout the MRSA TBI, students do the challenging work of reading, evidence gathering, piecing together explanatory models, and arguing about their models.

The MRSA Module text set consisted of 13 texts representing a range of sources; five from university websites, three from news agencies, two from science research journal reports, one each from the CDC website, a high school biology textbook excerpt, and a popular science magazine. They also offered a range of information in diverse representations: MRSA news stories, statistics on MRSA deaths, MRSA-human ecology, timelines showing antibiotic use and antibiotic resistance, models of evolution, and potential interventions. Four texts featured visuals: three graphs and one visual explanatory model.

The MRSA Module Interactive Notebook includes integrated routines, scaffolds, and tasks. Inquiry questions support student engagement with the phenomena. Notetakers support students in identifying and reasoning about evidence in the texts; and modeling and argumentation tasks to engage students in these science practices. The routines for the reading and modeling tasks provided opportunities to assess students' reading for modeling and to incorporate scaffolds responsively – orchestrating student groupings and timing; chunking texts and reading; offering strategic modeling of reading and modeling processes; and facilitating metacognitive conversations to deepen engagement, solve problems, and build knowledge of science principles and practices. An annotated teacher guide provides support for implementation.

The MRSA investigation is linked to the fourth learning phase, utilizing scientific literacy and discourse practices for disciplinary knowledge building, in which reading is framed as investigation, and the work of argumentation, drawing on science principles, to develop and refine models and explanations is foregrounded. It may be used subsequent to the Reading Science Models module and the Homoeostasis investigation to provide ongoing opportunities to learn.

Appendix B
Items on the Teacher Survey of Attitude, Self - Efficacy, and Argument/Multiple Text Practices

1. Identify claims and evidence in expository text passages
2. Develop disciplinary vocabulary, concepts, and principles
3. Use multiple sources of information presented in diverse formats and media in order to develop an argument
4. Determine the central idea of a text
5. Draw evidence from disciplinary (literary, historical, scientific) texts to support analysis, reflection, and research
6. Evaluate the credibility and reliability of a source of information
7. Identify points of agreement and disagreement across authors addressing the same issue or topic
8. Evaluate the claims, evidence, and reasoning presented by the author of an expository text passage
9. Understand the criteria for what counts as evidence in your discipline

Appendix C
Classroom Observation Protocol Constructs with Indicators

1. Construct 1: science reading opportunities (is reading central to the intellectual work or not?) (4 indicators)
   - Role of Reading
   - Breadth of Reading
   - Teacher Support of Reading
   - Accountability for Reading

2. Construct 2: teacher support for student efforts to comprehend science content from text (4 indicators)
   - Task Structure - Social Support for Reading Comprehension
   - Nature of Teacher Support
   - Student Practice
   - Accountability/Formative Assessment of Content from Reading

3. Construct 3: metacognitive inquiry into science reading and thinking processes (3 indicators)
   - Task Structure
   - Teacher Support
   - Student Practice

4. Construct 4: specific reading & science reading comprehension routines, tools, strategies and processes (2 indicators)
   - Teacher Support (Explicit Instruction and Modeling)
   - Student Practice (Routines, Tools, and Strategies)

5. Construct 5: science argumentation and/or building explanatory models from multiple sources (3 indicators)
   - Task Structure - science argumentation and/or model building through reading
   - Teacher Support - science argumentation and/or model building through reading
   - Student Practice - science argumentation and/or model building through reading

6. Construct 6: collaboration (3 indicators).
   - Student Practice - Collaboration
   - Task Structure - Collaboration
   - Teacher Support - Collaboration

Appendix D
Preliminary Analyses: Teachers

Table D1.
Teacher Survey: Summary Results for Exploratory Factor Analyses and Reliability Analyses at Pre and Post

		Pre			Post		
Scale	# of items	Factor Loadings: Range	Variance Explained	α	Factor Loadings: Range	Variance Explained	α
Attitude	9	.51 - .81	48.01%	.88	.55 - .90	54.21%	.91
Self-efficacy	9	.75 - .92	68.90%	.95	.71 - .96	69.10%	.95
Teaching philosophy: Reading	5	.51 - .77	43.30%	.78	.42 - .80	44.90%	.77
Science reading opportunities: Learning structure	4	.40 - .73	37.30%	.69	.54 - .77	46.90%	.77
*Argumentation and multiple source practices	9	.40 - .73	37.30%	.69	.54 - .77	46.90%	.77
*Content	3	.47 - .74	42.20%	.67	.33 - .97	61.30%	.70
*Metacognitive inquiry: Teacher modeling	5	.44 - .78	34.00%	.70	.66 - .81	54.50%	.85
*Metacognitive inquiry: Student practice	7	.47 - .74	34.00%	.70	.65 - .78	51.30%	.87
*Negotiation success: Instruction	5	.34 - .90	42.80%	.75	.48 - .91	46.90%	.79
Higher-order Teacher Practices	*5	.63 - .86	51.3%	.83	.86 - .88	74.2%	.93

α is Cronbach's α.
*Scales included in the *Higher-Order Teacher Practices*.

Table D2a.
Exploratory Factor Analysis for Six Constructs (Observation Time 1), N = 48

PRE scores	# of items	Variance Explained	Loadings Range	Cronbach's Alpha
C1: Opportunities	4	51.37%	.37-.89	.79
C2: Support	4	71.12%	.75-.90	.91
C3: Inquiry	3	66.77%	.79-.86	.84
C4: Strategies	2	62.57%	.79-.79	.77
C5: Argumentation	3	86.98%	.90-.97	.95
C6: Collaboration	3	65.42%	.77-.84	.84

Table D2b.
Exploratory Factor Analysis for Six Constructs (Observation Time 2), N = 48

POST scores	# of items	Variance Explained	Loadings Range	Cronbach's Alpha
C1: Opportunities	4	61.92%	.69-.91	.86
C2: Support	4	71.37%	.80-.91	.90
C3: Inquiry	3	75.96%	.83-.95	.90
C4: Strategies	2	77.47%	.88-.88	.87
C5: Argumentation	3	89.11%	.92-.97	.93
C6: Collaboration	3	71.26%	.75-.89	.88

Appendix E
Instructions for EBA Assessment

Task Introduction and Instructions for Reading	
Skin Cancer	**Coral Bleaching**
One purpose of reading in science is to understand the causes of scientific phenomena; in other words, we read to understand how and why things happen. To do this, we often need to gather information from multiple sources. **Today you will be reading about what causes some people to experience abnormal cell growth like skin cancer**. You will have to piece together important information across multiple sources to construct a good explanation of how and why this happens. No single source will provide all of the important pieces of the explanation. Instead, you are the one making connections across sources and coming up with an explanation. **Your task is to read the following set of sources to help you understand and explain what leads to differences in the risk of developing skin cancer. While reading, it is important to show your thinking by making notes in the margins or on the texts.** **You will be asked to answer questions and use specific information from the sources to support your ideas and conclusions.** You can read the sources in any order you wish, but you should read the sheet called "Background: Skin Damage" first, because it gives general information on the topic.	One purpose of reading in science is to understand the causes of scientific phenomena; in other words, we read to understand how and why things happen. To do this, we often need to gather information from multiple sources. **Today you will be reading about what causes "coral bleaching". Coral, which lives in the ocean, can be many different colors, but sometimes it loses its color and turns white.** You will have to piece together important information across multiple sources to construct a good explanation of how and why this happens. No single source will provide all of the important parts of the explanation. Instead, you are the one making connections across sources and coming up with an explanation. **Your task is to read the following set of sources to help you understand and explain what leads to differences in the rates of coral bleaching. While reading, it is important to show your thinking by making notes in the margins or on the texts.** **You will be asked to answer questions and use specific information from the sources to support your ideas and conclusions.** You can read the sources in any order you wish, but you should read the sheet called "Background: What is 'Coral Bleaching?'" first, because it gives general information on the topic.

Writing task	
Using this set of documents, write an essay **explaining what leads to differences in the risk of developing skin cancer**. Make sure to connect the ideas within your explanation to the differences in the risk of developing skin cancer. Be sure to use specific information from the documents to support your ideas and conclusions.	Using this set of documents, write an essay **explaining what leads to differences in the rates of coral bleaching**. Make sure to connect the ideas within your explanation to the differences in the rates of coral bleaching. Be sure to use specific information from the documents to support your ideas and conclusions.
Multiple Choice Items	
Based on the documents you read, please select the option that best fills in the blanks to answer the question: "**explain what leads to differences in the risk of developing skin cancer**."	Based on the documents you read, please select the option that best fills in the blanks to answer the question: "**explain what leads to differences in the rates of coral bleaching."**

Table 1. Core Constructs instantiated for Text-based Investigation in Science

Core Construct: General Definition	**Science: Text-based Investigation**
Epistemology: Beliefs about the nature of knowledge and the nature of knowing. What counts as knowledge? How do we know what we know?	Description, classification, and explanation of the natural and engineered worlds expressed as models and theories that are • approximations and have limitations • based on sound empirical data • socially constructed • meet criteria of parsimony, and logical cohesion • subject to revisions with successive empirical efforts that reflect changes in technology, theories and paradigms, and cultural norms.
Inquiry Practices, Reasoning Strategies: Ways in which claims and evidence are established, related, and validated	Scientific knowledge is built by: • developing coherent, logical classification systems, explanations, models or arguments from evidence • advancing and challenging classification systems and explanations • converging/corroboration of evidence • comparing/integrating across sources and representations • evaluating sources and evidence in terms of scope, inferential probability, reliability, and extent to which it accounts for evidence.
Overarching concepts, themes, principles, frameworks: Foundational concepts, ideas, reasoning principles, and assumptions. These serve as a basis for warranting, justifying, legitimizing connections between evidence and claims.	Scientists connect evidence to claims using • cross-cutting concepts (patterns; cause and effect; scale, proportion and quantity; systems and system models; energy and matter in systems; structure and function; stability and change of systems). • disciplinary core ideas in the physical sciences, earth and space sciences; life sciences; and engineering, technology, and applications of science.
Forms of information representation/types of texts: Types of texts and media (e.g., traditional print, oral, video, digital) in which information is represented and expressed.	• Scientific texts may have different explanatory purposes (e.g., cause effect, correlation, comparison, process sequence, chronology, enumeration, description). • Science texts convey meaning with multiple representations (e.g., verbal, diagrams, equations, graphs, tables, simulations, flowcharts, schematics, videos).

	• Different types of sources (genres) are written for different audiences and purposes, with implications for their content and structure (e.g., bench notes, refereed journal articles, textbooks, websites, blogs).
Discourse and language structures: The oral and written language forms in which information is expressed.	Science texts contain • distinctive grammatical structures (e.g., nominalizations, passive voice). • technical and specialized expressions. • signals to the degree of certainty, generalizability, and precision of statements. Argumentation is a scientific discourse practice in which evidence is used to support knowledge claims, and scientific principles and methods are used as warrants. Conventions for claim and evidence presentation in oral and written forms include • one-sided, two-sided arguments, multi-sided • two-sided, multi-sided refutational arguments • implicit arguments (embedded in descriptive and narrative structure) • oral arguments (debates, discussions, conversations)

Table 2. Progression of READI Science Goals across the Four Learning Phases

	Learning Phase			
READI Science Learning Goal	Building Classroom Routines To Support Science Literacy and Meaning Making	Building a Repertoire of Science Literacy and Discourse Processes	Deepening Scientific Literacy And Discourse Practices For Reasoned Sensemaking	Utilizing Scientific Literacy And Discourse Practices For Disciplinary Knowledge Building
1. Close reading. Engage in close reading of science information to construct domain knowledge, including multiple representations characteristic of the discipline and language learning strategies. Close reading encompasses	Setting a purpose for reading in science and science learning. Introducing • Annotation as persistent close reading practice. • Discussion of meta-comprehension in context of sense making. • Language for describing reading and reasoning processes.	Building confidence and range with science genre, text types and text structures (including scientific models). Previewing to set reading purpose and process based on topic, genre, text type, level of interest and level of challenge. Identifying and Handling Roadblocks while reading.	Set reading purpose based on text-based indicators of reliability and scope of content. Nascent Modeling Reading processes: attending to and clarifying/inquiring into the Science, phenomena, elements and relationships thereof, and model generation. Multi-text synthesis	Attending to scientific principles (theories such as mass-energy conservation, Hardy-Weinberg model) and Unifying Concepts of science (paradigms such as Evolution, Scale, Equilibrium, Matter and Energy, Form and Function, model and explanations, Evidence and representations) while reading.

metacomprehension and self-regulation of the process.				
2. Synthesize within and across multiple text sources	Reading multiple texts on same topic or related topics	Making connections to schema and in-text connections Building knowledge of key concepts across multiple texts	Attending to how multiple texts are connected (i.e. complimentary, additive, or even contradictory) and the affordances of various text types (i.e. personal anecdotes, primary data) Building explanations for inquiry questions across multiple texts	Viewing texts as investigations, setting purpose and inquiry for reading single and multiple texts. Attending to the new information afforded with additional texts and how information provided in those texts addresses the inquiry question
3. Construct explanations of science phenomena (explanatory models) using science principles, frameworks, enduring understandings, cross-cutting	Developing Norms for Classroom Discourse that holds students accountable to one another's ideas. Students begin to increasingly explicate their ideas and make them visible to the classroom and their peers.	Making norms for reading, writing, talking, speaking for text based science inquiry / sensemaking discussion routine Increased attention to building off of one another's ideas, attending to the logical coherence of one another's claims. Constructing gists of	Deepen language, representation and discourse patterns/conventions that attend to disciplinary norms for knowledge building. Attention to evidence, claims, and the links that one puts forth and that others propose within classroom discussion. Developing and making	Using disciplinary criteria for knowledge building as students engage in multiple cycles of reading, talking, and writing Constructing multi-text models from larger text sets Using models to predict implications of proposed solutions and answers to authentic science questions

concepts, and scientific evidence.		phenomena from single texts treating the text as an authority (i.e. noticing causal or correlation relationships between elements)	public disciplinary norms for model construction, justification, critique and revision Constructing models based on text evidence	
4. Justify explanations using science principles, frameworks and enduring understandings, cross-cutting concepts, and scientific evidence. (Includes evaluating the quality of the evidence.)	Citing text in sense making/meta-comprehension discussions. Reasoning and support based on authority (text, teacher, or one's own experience)	Identifying relevant evidence in single text that responds to inquiry questions Increasing attention to the distinction between evidence and inference in both texts and classroom talk	Identifying relevant evidence that informs the model while reading single and multiple texts Specifying how evidence informs the model Developing criteria for scientific models and explanations (writ large and for particular systems) Justifying models based on criteria for scientific models and reliability of text sources	Justifying explanations by appealing to scientific principles or unifying concepts of science. Refining explanatory models and explanations through careful attention to claims, evidence and reasoning
5. Critique explanations using science principles, frameworks and enduring understandings, cross-cutting concepts, and	Offering and tolerating alternative explanations, viewpoints, opinions in class discussions.	Disagreeing and offering evidence/rationale for it, Asking probing questions of each other in class discussions. Questioning while reading (to clarify, challenge or build knowledge). Increased	Offering alternative explanations in response to the explanations of others Using criteria for scientific models and explanations (writ large and for particular systems) as basis for critique (I think that part of the model	Critique the reliability of models and explanations based on the quality of evidentiary support (convergence, corroboration) Critique the scope of the model based on appeals to scientific principals and

scientific evidence. Critique explanations using science principles, frameworks and enduring understandings, cross-cutting concepts, and scientific evidence.		attention to how the ideas presented in text "fit with" one's prior knowledge and other texts.	may be wrong because …) and consensus building. Critique models and explanations based on the purpose of the model (the question it is trying to answer). Compare multiple, alternative models for single phenomena.	unifying concepts of science.
6. Science Epistemology and Inquiry. Demonstrate understanding of epistemology of science through inquiry dispositions and conceptual change awareness/orientation (intentionally building and refining key	Promoting the understanding that scientific findings have both practical and theoretical implications for science and society Taking inquiry stance as a basis for interacting with text	Viewing science findings as limited and tentative, based on available evidence Tolerating ambiguity and seeking the best understanding, given the evidence, while reading	Recognize that science knowledge is socially constructed by peer critique and public dissemination (advancing and challenging explanations/models) to create scientific explanations that meet certain criteria (based on sound empirical data, parsimonious and logically cohesive) as a basis for co-construction of knowledge while reading.	Recognize that • Science knowledge building is shaped by (and shapes) scientific principles (theories) and Unifying Concepts (paradigms) • Theories and paradigms are used as a basis for constructing, justifying and critiquing models while reading.

concepts through multiple encounters with text); seeing science as a means to solve problems and address authentic questions about scientific problems				

Table 3. Overview of READI Professional Development for Intervention Teachers

Timing	Topics of the Sessions
Days 1-4 4 Single day sessions over the Spring Semester 2014	• READI and the Reading Apprenticeship Framework • Inquiry into Science Reading • Metacognitive Conversation and Text Complexity • Formative Assessment and Reciprocal Modeling • Using READI pedagogies in classroom practice
Days 5-9 5 consecutive days during Summer 2014	• READI Science Learning Goals • READI Science Text Based Investigations • READI science goals and Instructional Progression • Planning use of READI intervention pedagogy and materials in RCT study semester
Days 10-11 2 single days during ongoing implementation (Fall, 2014)	• Reflection on classroom experience with READI intervention pedagogy and materials • Formative assessment of student learning • Planning ongoing use of READI intervention pedagogy and materials
Day 12 Single session that occurred after conclusion of post-testing.	• Reflection on READI Intervention pedagogy and materials

Table 4. Stratified random assignment of schools and teachers for each strata

Strata	Intervention		Control	
	# of Schools	# of Teachers	# of Schools	# of Teachers
1	1	1	1	3
2	1	1	2	2
3	3	7	2	6
4	3	5	2	3
5	3	7	2	5
6	1	3	3	5

Table 5. Item Types on the Evidence-Based Argument Assessment for Students

Task Type	Description of Task Components	Science Learning Goals Targeted
Essay		
Reading Task	• Read and annotate texts.	• Close reading of task relevant information
Write essay with texts present	• Explanation criteria: relevance, accuracy, coherence, completeness • Support criteria: selecting citation of texts • Rhetorical organization of ideas	• Synthesize task relevant information within and across text set. • Construct explanations of elements and connections among them
Non-essay		
Multiple Choice (4 alternatives)	• 9 items that tapped connections among elements in the causal model • Required inferring relations among causes	• Synthesize task relevant information within and across text set. • Construct explanations of elements and connections among them
Graphical model comparison	• Select better of 2 models (connected element chain versus distinct causal connections	• Critique other's models/explanations • Synthesize text(s) • Construct explanations
"Peer" Essay evaluation task	• 2 explanations written by students varied on relevance, coherence, completeness, deviation from normal, citing texts • What is well done? • What advice for improving explanation?	• Critique other's models/explanations • Synthesize text(s) • Construct explanations

Table 6
Comparison of PreTest Mean Scores on Survey Scales for READI Intervention and BAU Teachers

	READI $n = 24$		BAU $n = 19$				
Scale	*M*	*SD*	*M*	*SD*	Δ*M*	*t*(41)	Cohen's *d*
Familiarity with CCSS	2.58	1.02	2.75	0.91	-0.17	-0.57	0.18
Attitude	3.76	0.46	3.81	0.82	-0.05	0.22[a]	0.08
Self-efficacy	3.39	0.76	2.96	0.80	0.43	-1.79	0.55
Teaching philosophy: Reading	3.58	0.66	3.56	0.70	0.02	0.08	0.03
Science reading opportunities: Learning structure	2.79	0.76	2.72	0.74	0.07	-0.30	0.09
Higher-order Teacher Practice	2.87	0.47	2.94	0.54	-0.07	-0.45	0.14
*Argumentation and multiples source practices	3.13	0.54	3.23	0.62	-0.10	0.59	0.17
*Content	3.03	0.67	3.26	0.79	-0.24	1.06	0.33
*Metacognitive inquiry: Teacher modeling	2.99	0.58	2.88	0.63	0.11	-0.58	0.18
*Metacognitive inquiry: Student practice	2.45	0.70	2.59	0.60	-0.13	0.66	0.20
*Negotiation success: Instruction	2.73	0.61	2.72	0.77	0.02	-0.08	0.03

Notes. CCSS = Common Core State Standards.
Teaching philosophy: Reading – This was reverse coded so that higher scores reflect beliefs more consistent with the READI perspective.
[a] Equal variance is not assumed (Levene's test: $F = 6.73$, $p = .013$), independent samples t-test: $t = .22$, $df = 28.71$. For all other scales, equal variance was assumed.
Cohen's $d = M2 - M1/\sqrt{[(S1^2 + S2^2)/ 2]}$; *d* values between 0.2 and 0.5 constitute a small effect, 0.5 to 0.8 a medium effect, and 0.8 or greater a large effect.

Table 7
Comparison of PostTest Mean Scores on Survey Scales for Intervention and BAU Teachers

	READI $n = 23$		BAU $n = 23$				
Scale	*M*	*SD*	*M*	*SD*	Δ*M*	*t*(44)	Cohen's *d*
Familiarity with CCSS	3.17	0.78	2.96	0.75	0.22	0.97	0.29
Attitude	4.21	0.64	3.90	0.66	0.31	1.64	0.48
Self-efficacy	3.44	0.83	3.02	0.82	0.42	1.70	0.51
Teaching philosophy: Reading	3.92	0.74	3.72	0.56	0.20	1.03	0.30
Science reading opportunities: Learning structure	3.70	0.45	2.85	0.74	0.84***	4.72[a]	1.37
Higher-Order Teacher Practice	3.95	0.31	3.05	0.56	0.90***	6.90[b]	2.00
*Argumentation and multiples source practices	3.90	0.40	3.20	0.62	0.70***	4.60[c]	1.34
*Content	4.12	0.47	3.26	0.75	0.86***	4.66	1.37
*Metacognitive inquiry: Teacher modeling	3.94	0.40	3.02	0.71	0.92***	5.46[d]	1.60
*Metacognitive inquiry: Student practice	3.87	0.43	2.78	0.70	1.09***	6.37[e]	1.88
*Negotiation success: Instruction	3.91	0.41	2.99	0.57	0.92***	6.30	1.85

Notes. Teaching philosophy: Reading – This was reverse coded so that higher scores reflect beliefs more consistent with the READI perspective.
[a] Equal variance is not assumed (Levene's test: $F = 6.96$, $p = .011$), independent samples *t*-test: $t = 4.72$, $df = 38.30$.
[b] Equal variance is not assumed (Levene's test: $F = 6.29$, $p = .016$), independent samples *t*-test: $t = 6.90$, $df = 36.14$.
[c] Equal variance is not assumed (Levene's test: $F = 4.78$, $p = .034$), independent samples *t*-test: $t = 4.60$, $df = 39.67$.
[d] Equal variance is not assumed (Levene's test: $F = 11.95$, $p = .001$), independent samples *t*-test: $t = 5.46$, $df = 34.68$.
[e] Equal variance is not assumed (Levene's test: $F = 8.21$, $p = .006$), independent samples *t*-test: $t = 6.37$, $df = 36.70$.
Cohen's $d = M2 - M1/\sqrt{[(S1^2 + S2^2)/ 2]}$; *d* values between 0.2 and 0.5 constitute a small effect, 0.5 to 0.8 a medium effect, and 0.8 or greater a large effect.
*** $p < .001$.

Table 8

Comparisons of the Mean Rubric Score Points on Classroom Practices Constructs for READI and BAU Teachers

Time-1 Observations

	READI		BAU		*t*-test			ES
Construct	*M*	*SD*	*M*	*SD*	*t*	*df*	*P*	*d*
C1: Opportunities	2.74	0.68	1.97	0.66	-4.00	46	.000	1.16
C2: Support	2.45	0.84	1.77	0.72	-3.00	46	.004	0.87
C3: Inquiry	2.04	0.73	1.33	0.61	-3.66	46	.001	1.06
C4: Strategies	1.77	0.63	1.38	0.47	-2.48	46	.017	0.71
C5: Argumentation	1.60	0.75	1.04	0.20	-3.51	26.4 [a]	.002	1.01
C6: Collaboration	2.19	0.71	1.63	0.75	-2.70	46	.010	0.78

Time-2 Observations

	READI		BAU		*t*-test			ES
Construct	*M*	*SD*	*M*	*SD*	*t*	*df*	*P*	*d*
C1: Opportunities	2.92	0.80	1.77	0.64	-5.49	46	.000	1.58
C2: Support	2.90	0.80	1.78	0.59	-5.49	46	.000	1.58
C3: Inquiry	2.36	0.95	1.25	0.43	-5.21	32.0 [a]	.000	1.50
C4: Strategies	2.04	0.79	1.17	0.38	-4.87	33.1 [a]	.000	1.41
C5: Argumentation	1.71	0.94	1.00	0.00	-3.69	23	.001	1.07
C6: Collaboration	2.58	0.77	1.51	0.67	-5.15	46	.000	1.49

Notes. ES = effect size.

[a.] Levene's Test of Equal Homogeneity was significant; thus, variance is not equal across two cohorts.

Cohen's d = (M2 - M1) /(SQRT((SD1^2 + SD2^2)/2)); .2-.5 = small; .5-.8 = medium; >.8 = large.

Table 9. Descriptives Statistics for EBA Assessment Measures and Scales at Pre and Post Intervention for READI and BAU Students[a]

PRE/Beg of Semester		READI			BAU			Indep Samples		
EBA Measures and Scales	Total *N*	*N*	*M*	*SD*	*N*	*M*	*SD*	*T*	*df*	*p*
MC-pre%	964	567	53.03	25.15	397	54.77	25.18	1.06	962	.290
Nodes (% mentioned of possible)	959	566	30.68	21.21	393	32.65	22.95	1.37	957	.172
Links (% mentioned of possible)	959	566	13.82	16.78	393	14.99	18.51	1.02	957	.307
Epistemology Scales										
Corroboration	964	567	4.85	0.65	397	4.91	0.69	1.44	962	.152
Complex/Uncertain	964	567	3.79	0.80	397	3.90	0.83	2.17	962	.030
Self - Efficacy	949	556	3.65	0.81	382	3.68	0.77	0.61	947	.542
Topic Prior Knowledge Pre	962	565	2.92	0.76	397	2.97	0.73	1.00	960	.317

POST/End of Intervention		READI Intervention			BAU Control			Indep Samples		
EBA Measures and Scales	Total *N*	*N*	*M*	*SD*	*N*	*M*	*SD*	*T*	*df*	*P*
MC % correct	964	567	55.67	25.50	397	50.94	26.16	2.81	962	.005
Nodes (% mentioned of possible)	954	561	34.83	21.59	393	32.55	22.53	1.58	952	.115
Links (% mentioned of possible)	954	561	16.45	17.02	393	15.16	16.52	1.17	952	.244
Epistemology Scales										
Corroboration	946	558	4.89	0.70	388	4.80	0.75	1.88	795.065[b]	.060
Complex/Uncertain	946	558	4.00	0.85	388	4.00	0.83	0.02	944	.983
Self - Efficacy	937	555	3.61	0.84	382	3.54	0.84	1.28	935	.202
Topic Prior Knowledge	954	564	3.00	0.74	390	3.00	0.67	0.05	882.901[b]	.957

[a]Note that differences in sample sizes reflect missing data. There were some students despite being present for all 4 days failed to provide any written essay in their assessment booklets. Thus the sample size for the analyses of the essays was lower than that for the multiple choice; similarly for the various scales.
[b]Levene's Test of Equal Homogeneity was significant; thus, variance is not equal across two cohorts.

Table 10. Variances and ICCs at each level for three multilevel models for the multiple choice performance

	3-level: students, classrooms, schools	3-level: students, teachers, schools	4-level: students, classrooms, teachers, schools
Variances			
Students	470.69773	494.85288	461.86623
Classrooms	59.53255	n/a	95.23247
Teachers	n/a	35.31683	0.28521
Schools	146.34189	140.56513	133.66919
ICCs			
ICC-students	69.57%	73.78%	77.52%
ICC-classrooms	8.80%	n/a	13.79%
ICC-teachers	n/a	5.27%	0.05%
ICC-schools	21.63%	20.96%	22.43%

Table 11. Trimmed model resulting from multilevel modeling of multiple choice posttest performance[a]

Variable	β Coefficient	*SE*	*t*	*p*	ES[b]
Level-3: School (df = 17)					
Condition: READI vs BAU	5.71	1.97	2.90	.010	0.26
Strata 1	43.53	3.79	11.47	.000	2.01
Strata 2	44.85	3.96	11.32	.000	2.07
Strata 3	47.48	2.89	16.44	.000	2.19
Strata 4	54.87	3.07	17.88	.000	2.53
Strata 5	53.94	2.72	19.83	.000	2.49
Strata 6	56.88	2.66	21.37	.000	2.62
Level-1: Students (individual) (df = 840)					
Corroboration-pre	4.69	1.08	4.33	.000	0.29
Complex/Uncertain-pre	2.96	0.89	-3.33	.001	0.22
MC-pre	0.45	0.04	11.08	.000	1.03
Topic	-4.19	3.05	-1.37	.170	-0.19
Topic X MC-pre Interaction	-0.11	0.05	-2.18	.029	-0.34

[a]These multilevel models reflect students nested within classrooms, nested within schools. Because there were no predictor variables at level 2 = classroom, this level is not displayed in the table.

[b]Effect Size for dichotomous variables (ES) = $\beta 1 / \sigma$. Effect size for continuous variables (ES) = $\beta 1 * 2SD_{iv} / \sigma$. These effect sizes are interpreted as Cohen's *d*, with d = 0.2 a small effect, 0.5 a medium effect and 0.8 or greater a large effect.

Table 12. Results of Multilevel Modeling Comparing Intervention and BAU Control groups on Concept Nodes (upper panel) and links (lower panel) included in essays[a]

Concept Nodes	Coefficient	*SE*	*t*	*p*	*ES*[b]
Level-3: School (df = 17)					
Condition	2.11	1.50	1.41	.178	0.11
Strata 1	27.14	2.88	9.42	.000	1.38
Strata 2	27.97	3.26	8.59	.000	1.42
Strata 3	32.12	2.14	15.03	.000	1.63
Strata 4	41.08	2.33	17.59	.000	2.09
Strata 5	41.27	1.97	20.91	.000	2.10
Strata 6	40.57	1.97	20.63	.000	2.06
Level-1: Students (individual) (df = 801)					
Corroboration-pre	3.00	1.02	2.94	.003	0.20
Self-Efficacy-pre	2.05	0.82	2.51	.012	0.16
Prior Knowledge-post	-1.91	0.89	-2.14	.032	-0.14
Topic	-9.49	2.13	-4.47	.000	-0.48
Elements-pre	0.48	0.05	9.37	.000	0.05
Topic X Elements-pre Interaction	-0.27	0.06	-4.58	.000	-0.02

Links	Coefficient	*SE*	*t*	*p*	*ES*[b]
Level-3: School (df = 17)					
Condition	1.21	1.20	1.01	.328	0.08
Strata 1	8.31	2.26	3.68	.002	0.54
Strata 2	11.75	2.59	4.54	.000	0.76
Strata 3	11.71	1.59	7.37	.000	0.76
Strata 4	19.84	1.74	11.38	.000	1.28
Strata 5	18.33	1.47	12.49	.000	1.18
Strata 6	19.97	1.45	13.77	.000	1.29
Level-1: Students (individual) (df = 812)					
Complex/Uncertain-pre	-1.63	0.65	-2.49	.013	-0.17
Self-Efficacy-pre	1.86	0.63	2.94	.003	0.19
Topic	-2.69	1.28	-2.11	.035	-0.17
Link-pre	0.29	0.05	6.36	.000	0.04
Topic X Link-pre Interaction	-0.12	0.06	-2.17	.030	-0.01

[a]These multilevel models reflect students nested within classrooms, nested within schools. Because there were no predictor variables at level 2 = classroom, this level is not displayed in the table.
[b]Effect Size for dichotomous variables (ES) = $\beta 1 / \sigma$. Effect size for continuous variables (ES) = $\beta 1 * 2SD_{iv} / \sigma$.

Table 13. Multilevel Model for GISA[a]

Variable	Coefficient	*SE*	*t*	*p*	*ES*
Level-3: School					
Condition	4.41	1.96	2.25	.038	0.32
Strata 1	51.95	3.26	15.92	.000	
Strata 2	52.40	3.28	15.97	.000	
Strata 3	54.08	2.40	22.52	.000	
Strata 4	53.15	2.51	21.19	.000	
Strata 5	56.24	2.19	25.70	.000	
Strata 6	59.16	2.44	24.20	.000	
Level-1: Students (individual)					
RISE	0.46	0.04	11.37	.000	
Corroboration_pre	1.77	0.75	2.35	.019	
Simple/Certain_pre	-1.54	0.59	-2.60	.010	
Self-efficacy_pre	0.16	0.57	0.29	.772	

[a]These multilevel models reflect students nested within classrooms, nested within schools. Because there were no predictor variables at level 2 = classroom, this level is not displayed in the table. Level-3 df = 17; level-1 df = 810.

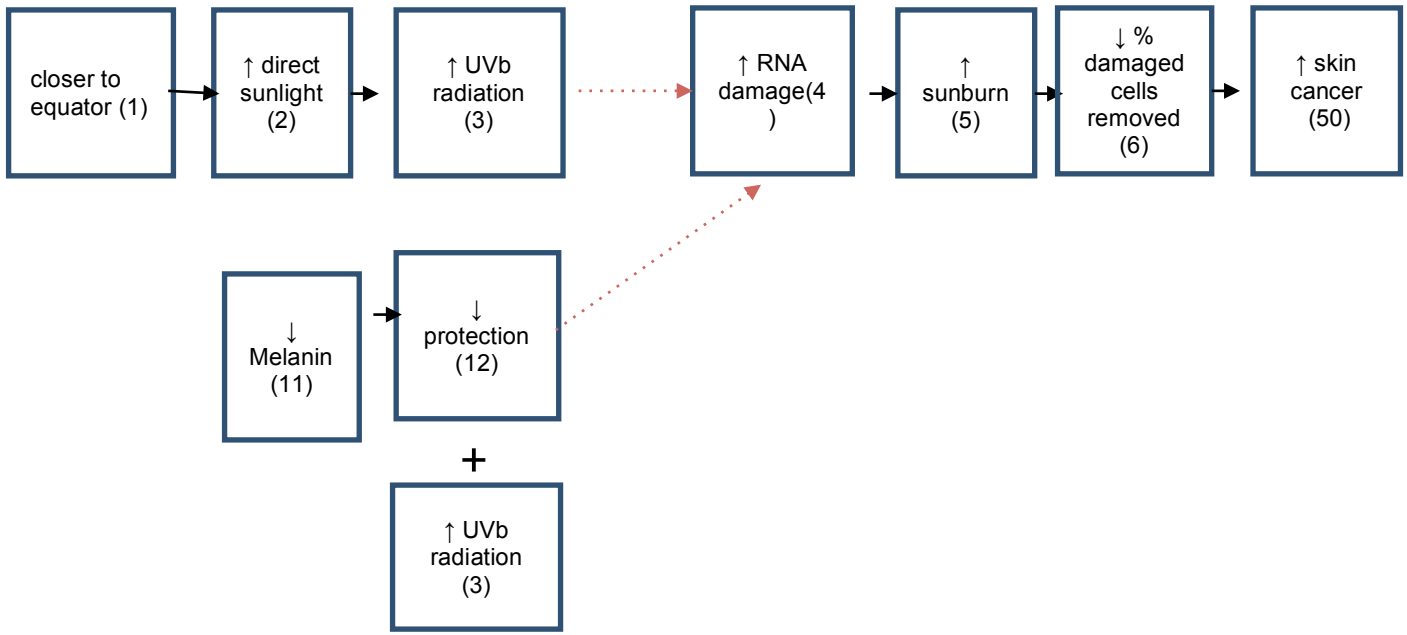
Figure 1. Progression of Science Learning Goals, READI Modules, and Biology Topics across the Semester

Week in semester	1 - 4	5 - 8	9 - 12	13 - 17
READI Science Learning Goal Progression across Learning Phases (see Table 2)	**Building classroom routines to support science literacy and meaning making.** Students begin to see text as a part of *scientific practice* and that scientific knowledge is built through *close reading of text,* and also through class-wide *knowledge-building* discourse. Students begin to see themselves as readers of science, increasingly interact with texts and view the classroom as a place where their knowledge is valued.	**Building a repertoire of science literacy and discourse processes.** Text is increasingly used to deepen understanding of scientific phenomena. Attention on *kinds of evidence* that are embedded in *various text types* (written, visual representations), *interpretations* that can be made from different kinds of evidence, and how this helps *construct explanations* of science phenomena. Students use a READI science module to build knowledge of conventions of scientific models and criteria for *evaluating* them. Increasing *awareness, confidence, ownership* of science reading and reasoning practices.	**Deepening scientific literacy and discourse practices for reasoned sensemaking.** Students dig into a READI science module to continue building *close reading* and *multiple text synthesis* practices in order to develop a causal explanatory account for scientific phenomena. Students take active role in building explanations of scientific phenomena in the world and increasingly *view models as representations* that facilitate their own sense making activities: to *clarify, refine, and modify or revise* their own science thinking.	**Utilizing scientific literacy and discourse practices for disciplinary knowledge building.** Students utilize a READI science module to deepen *close reading* and *multiple text synthesis* in order to *construct, justify, and critique* causal explanatory accounts for scientific phenomena. Students work more independently in building explanations of scientific phenomena in the world as well as taking an active role in *justification and critique* of scientific explanations.
READI Modules	Use of READI candidate texts		Reading Models	Homeostasis Module (~ 3 – 4 weeks)	MRSA Module (~4 – 5 weeks)
Science	Introduction	Cell Biology		•Feedback mechanisms	•Natural selection (variation in traits,

Topics	to Biology • Community and Ecosystem Ecology (Inter-dependence and energy flow in ecosystems) • Energy production in plants (Photosyn-thesis) • Scientific evidence of Evolution • Cell biology: cell division, communica-tion	• Basic cell biochemistry • Enzymes/substrate interactions • Cell differentiation and specialization • History of cell biology • Technology and advancement of science knowledge		• Cell communication • Homeostasis (both cellular and organism levels – human focus) • Role of specialized organs and systems (e.g., kidneys, pancreas, endocrine system) in maintaining balance in the human body. • Diabetes and hypo/hypernatremia as cases of homeostasis disruption • Behavior and its impact on homeostasis	genertic inheritance, selection) ad adaptation • Antibotic resistance (focused on staphylococcus aureaus) • Microbes: bacteria and viruses; human microbiota (staphylococcus aureus in particular); Binary fission of bacteria • Human contributions to evolution and evolutionary engineering.

Figure 2. Representations of complete and coherent models that could be constructed from text sets for a. Skin Cancer and b. Coral Bleaching.

a. Skin Cancer



b. Coral Bleaching