AN ABSTRACT OF THE DISSERTATION OF

Timothy A. Collins for the degree of Doctor of Philosophy in Science Education presented on April 21, 2011.
Title: Science Inquiry as Knowledge Transformation: Investigating Metacognitive and Self-regulation Strategies to Assist Students in Writing about Scientific Inquiry Tasks

Abstract approved: ____
Lawrence B. Flick

Science inquiry is central to the science education reform efforts that began in the early 1990’s. It is both a topic of instruction and a process to be experienced. Student engagement in the process of scientific inquiry was the focus of this study. The process of scientific inquiry can be conceived as a two-part task. In the initial part of the task, students identify a question or problem to study and then carry out an investigation to address the issue. In the second part of the task, students analyze their data to propose explanations and then report their findings. Knowing that students struggle with science inquiry tasks, this study sought to investigate ways to help students become more successful with the communication demands of science inquiry tasks.

The study took place in a high school chemistry class. Students in this study completed a total of three inquiry tasks over the course of one school year. Students were split into four experimental groups in order to determine the effect of goal setting, metacognitive prompts, and sentence stems on student inquiry tasks. The quality of the student written work was assessed using a scoring rubric familiar to the students. In addition, students were asked at four different times in the school year to respond to a self-efficacy survey that measured student self-efficacy for chemistry content and science inquiry processes.

Student self-efficacy for the process of scientific inquiry was positive and did not change over the course of the study while student scores on the science inquiry tasks rose significantly. The metacognitive prompts and instruction in goal setting did not have any effect on student inquiry scores. Results related to the effect of the sentence stems were mixed. An analysis of student work indicated that students who received high marks on their initial inquiry task in this study were the ones that adopted the use of the sentence stems. Students who received low marks on their initial inquiry task did not tend to use the sentence stems. An analysis of word counts that compared the number of words used in the Framing section to the number of words used in the Analysis section indicated that students may have been using insufficient writing strategies. This study concludes with implications for classroom practice and recommendations for future research around student writing in the science classroom.

Science Inquiry as Knowledge Transformation: Investigating Metacognitive and Self-regulation Strategies to Assist Students in Writing about Scientific Inquiry Tasks

by
Timothy A Collins

A DISSERTATION

Submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented April 21, 2011
Commencement June 2011

Doctor of Philosophy dissertation of Timothy A. Collins presented on April 21, 2011.

APPROVED:

Major Professor representing Science Education

Chair of the Department of Science and Mathematics Education

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Timothy A. Collins, Author

ACKNOWLEDGEMENTS

The author expresses sincere appreciation to

My students, fellow staff, and district leadership who helped make this study possible.

Ron Doel, Larry Enochs, Nam-Hwa Kang, and Vincent Remcho for their feedback, support, and willingness to serve as advisors.

Larry Flick for his many roles (major professor, mentor, teacher, advisor, boss, etc.) in helping to see this project to completion.

David & Susan Collins and family who put up with me for two years; taking over your guestroom and eating your food while I commuted to and from OSU.

Madison & Allie Collins who gave up weekends and summer days with their daddy so that he could finish this study.

Tia Collins, my wife. You have invested as many hours in this work as I have. Your support and encouragement have made all the difference. It’s been an incredible journey so far, and I am honored to have been able to share it with you.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF APPENDICES

LIST OF APPENDICIES (Continued)

LIST OF APPENDICIES (Continued)

LIST OF APPENDIX TABLES

LIST OF APPENDIX TABLES (Continued)

LIST OF APPENDIX TABLES (Continued)

LIST OF APPENDIX TABLES (Continued)

LIST OF APPENDIX TABLES (Continued)

LIST OF APPENDIX TABLES (Continued)

LIST OF APPENDIX TABLES (Continued)

LIST OF APPENDIX TABLES (Continued)

LIST OF APPENDIX TABLES (Continued)

LIST OF APPENDIX TABLES (Continued)

LIST OF APPENDIX TABLES (Continued)

LIST OF APPENDIX TABLES (Continued)

DEDICATION

This work is dedicated to my parents, Dennis & Judy Collins, who have modeled what it means to truly be life-long learners.

Science Inquiry as Knowledge Transformation:
Metacognition and Self-regulation in Writing Scientific Inquiries

Chapter 1: Introduction

Problem Statement

Scientific Inquiry is at the heart of science reform efforts that began in the mid 1990's. The National Research Council (NRC, 1996, pg 13) stated that two of their four goals for science education were:

1) Students should experience the richness and excitement of knowing about and understanding the natural world.
2) Students should use appropriate scientific processes and principles in making personal decisions.

These goals implied that students should understand both the content of science as well as the process of science. In a subsequent publication, the NRC (2000, pg 13) further explained their position on inquiry to state inquiry is both a topic to be taught and a process to be experienced. In this sense, science inquiry is an ongoing process that is continually embedded into a curriculum. The current goals of science reform are not that students necessarily learn more content, but rather, that through the content students ought to be continually engaged in the processes of science.

These processes, referred to as scientific inquiry (AAAS, 1990, 1993; NRC, 1996, 2000), involve skills such as making observations, developing questions, designing experiments, collecting data, developing conclusions based on data, and communicating results. In order for students to become more competent in science inquiry, they need continual, ongoing practice. As students develop these skills through the application of science content, these process oriented activities then serve to bolster and support student understanding of content. The state of Oregon, for example, requires that students take standardized tests to assess science content knowledge but relies on teacher reporting to assess student abilities in regards to scientific inquiry processes.

This led to the question, how proficient are our students in applying the processes of scientific inquiry? One indicator of student ability in science inquiry was measured by the NAEP given in 2005. Among the eighth graders who were tested, only 29% were rated as at least proficient in scientific inquiry. This dropped to 18% for students in the twelfth grade (NAEP, 2005). Proficiency in science inquiry at the eighth grade level was defined as "the ability to design an experiment and have an emerging understanding of scientific phenomena, and can design plans to solve problems" (NAEP, 2005). At the twelfth grade, proficiency in science inquiry was defined as, "a working ability to design and conduct scientific investigations" (NAEP, 2005). From eighth grade to twelfth grade, the standards increased for what students should be able to do. In the 8th grade, the focus was on the ability to design an experiment. As students progressed, the goal was that students demonstrate ability to both design and carry out multiple investigations. While the proficiency rates decreased from the 8th to the 12th grade, both of these proficiency rates were low. In addition, these proficiency rates were virtually unchanged from the NAEP assessment that was given in 2000 (NAEP, 2005) which indicated that these findings were fairly stable.

The results from this large-scale national assessment mirrored assessments conducted at the local level. At a large suburban high school (grades 9 – 12) located in the western United States, only 28% (424 / 1540) of the student body had completed an in-class science inquiry work-sample that qualified as “passing” according to the scoring guide standards established by the state for the 2007 – 2008 school year (personal communication, April 9, 2008). In this particular district, all high school students were required to have attempted at least one science inquiry project in order to graduate from high school. This school was predominantly Caucasian with a growing Hispanic population. On state tests this school tended to be at or slightly above the state average for math, reading, and writing. Overall, the state rated this high school as “satisfactory”.

In another large suburban high school (grades 9 – 12) located in the same state, only 31% of the student body had achieved passing scores on their science inquiry work samples for the 2007 – 2008 school year (personal communication, January 10, 2010). This school was more evenly split between Caucasian and Hispanic students. In terms of test results, this high school tended to score at or slightly below state averages on state tests for math, reading, and writing. Again, this school was rated as “satisfactory” by the state. For both schools, the passing rates were in line with the results collected at the national level. While these results were by no means exhaustive, they indicated that students do indeed struggle with scientific inquiry.

This led to the question: what can be done to help students become more successful in understanding and applying the concepts of scientific inquiry in a high school classroom? What was it about these science inquiry tasks that students found most challenging? Inquiry not only entails the actions that a student goes through in the investigation of a problem, but also includes an individual's ability to clearly communicate their findings. One possibility is that students lack the basic understanding of science content to be successful on these science inquiry tasks. Recent NAEP data (NAEP 2005) indicated that 57% of the eighth graders tested had at least a basic understanding of science. Only 27%, though, of the 8th graders tested were considered to have had a proficient understanding of science (NAEP 2005). Difficulty with science content cannot be ignored.

While difficulty with science content may be one hurdle that students face in relation to science inquiry tasks, there is another possibility. Student difficulties with science inquiry may also be grounded in the communication demands of the inquiry task. A number of recent studies have begun to investigate the central role that writing plays in the process of scientific inquiry (Baker, 20004; Baker et al. 2008; Klein, 2006; Warrick et al 2003; Yore & Treagust, 2006). Yore and Treagust (2006) speak specifically to the need to develop a more current conception of scientific literacy, of which scientific inquiry is a part.

Scientific inquiry then, can be conceptualized as two parallel, yet connected tasks. First is the task of investigation. Based on their understanding of science content, a student must choose a question, an appropriate methodology which will yield satisfactory data, and then carry out the investigation. Once this is complete, the student must engage in a second task where they must determine a form which will allow them to adequately communicate their study methodology and its subsequent conclusions (Champagne, Kouba, & Hurley 2000). The focus of this study was on continuing to build our understanding of the cognitive and metacognitive skills necessary for high school students to carry out an inquiry task and then effectively communicate their findings.

Conceptual Frameworks

Science Inquiry

Early science education reform documents provided a common language to talk about inquiry processes and also helped to define how student inquiry tasks might be assessed (AAAS, 1993; NRC 1996, 2000). Studies such as those by White & Frekeriksen (1998) and Magnusson & Palincsar (1995, 2005) sought to understand how students learn inquiry in science. Even though the framework for this study was significantly based on the work by Magnusson & Palincsar (1995, 2005), the Magnusosn & Palincsar (1995 & 2005) framework builds on previous work in the field of science education.

In 1993, the American Association for the Advancement of Science (AAAS) published *Benchmarks for Science Literacy*. AAAS recommended that students, "participate in scientific investigations that progressively approximate good science" (1993, pg 13). By the end of 12th grade, students should be able to understand "The nature and importance of prediction in science…[and] the use of statistics, probability, and modeling in making scientific predictions" (AAAS, 1993, pg 13). The *Benchmarks* outlined the understandings that students should develop from grade school through high school related to scientific inquiry. Student understanding of science inquiry in the *Benchmarks* was built on a spiraling curriculum concept where ideas were revisited continually but in greater depth than before.

The model of inquiry outlined by the National Research Council (NRC, 1996) shared much of the same philosophical framework as the *Benchmarks* (1993). The NRC (1996), similar to the *Benchmarks* (1993), recommended that students in grades 9 - 12 "must actively participate in scientific investigations, and they must actually use the cognitive and manipulative skills associated with the formulation of scientific explanations" (pg 173). These skills were then further defined. To be successful in scientific inquiry, the NRC (1996) recommended that students be able to do the following: ask scientific questions, design investigations to answer these questions, collect and analyze data, explain results, and defend conclusions with evidence (pgs 175 – 176). In this framework, there was a much greater emphasis on students communicating their work and conclusions through writing or speaking (pg 176).

A follow up publication by the NRC (2000) helped to further define this framework. Instruction on, and in, scientific inquiry should incorporate five complimentary, yet distinct phases. Science inquiry begins as students are engaged with a scientific problem or process. Following this initial engagement, students should then be given the chance to explore the science question or phenomena of interest. After exploration, students should propose explanations related to the information gathered during the exploration phase. To test these explanations, students should try to extend these understandings to new situations to see if their ideas are still valid. In the final phase, students evaluate their learning with peers and teachers. The NRC (2000) noted that these phases should not be seen as "lockstep, prescriptive devices" (pg 35), but rather "as general guidelines for designing instruction" (pg 35). In addition, these phases were envisioned to guide teacher practice related to the development of an entire unit as well as inform teacher practice as they created day to day learning experiences.

In order to better understand how to help students understand the inquiry process, White and Frederiksen (1998) developed their own model called "The Inquiry Cycle". White & Frederiksen (1998) noted that their framework was developed in order to help students "construct conceptual models of scientific phenomena and…monitor and reflect on their progress" (pg 5). Similar to the NRC (1996, 2000), White & Frederiksen (1998) employed a cyclical model (see Figure 1) that was explicitly used with students involved in their study. A number of the phases were also similar such as identification of a question, experimentation, and then applying the results to a new setting.

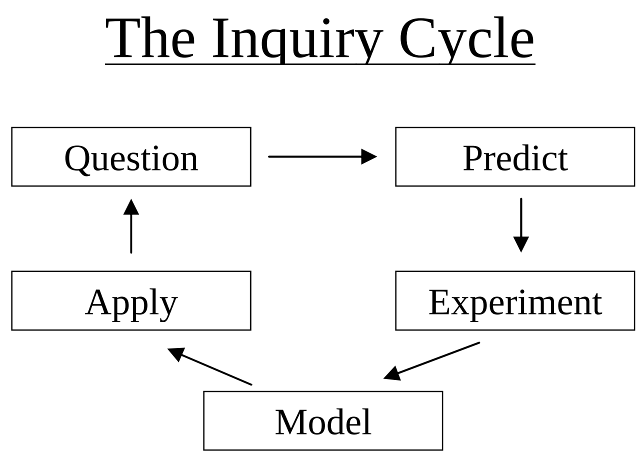


Figure 1: One Model of Scientific Inquiry – The Inquiry Cycle proposed by White & Frederiksen (1998). (Note. From "Inquiry, modeling, and metacognition: making science accessible to all students", by B. White and J. Frederiksen, 1998, Cognition and Instruction, 16(1), p. 5. Copyright 1998 by Routledge. Reprinted with permission)

While the previous three frameworks focused heavily on the processes that would take place during the planning and data collection phase of an experiment, Magnusson & Palincsar (1995 & 2005) proposed a model where science inquiry was split into two main phases: the investigation phase and the reporting phase. They outlined a process called guided inquiry (see Figure 2). In line with the recommendations of the National Research Council (1996, 2000), this inquiry cycle is both a model that informs curriculum and teaching practice, and a process which could be explicitly taught. Similar to White & Frederiksen (1998), the model is cyclical in nature implying that inquiry is not a "one and done" event, but rather is a continual part of a curriculum. Again, many elements of the prior models are present in this one such as the development of a question, designing an investigation, and proposing explanations. In the Magnusson & Palincsar (2005) model, though, there is also a very strong emphasis on the communication of results.

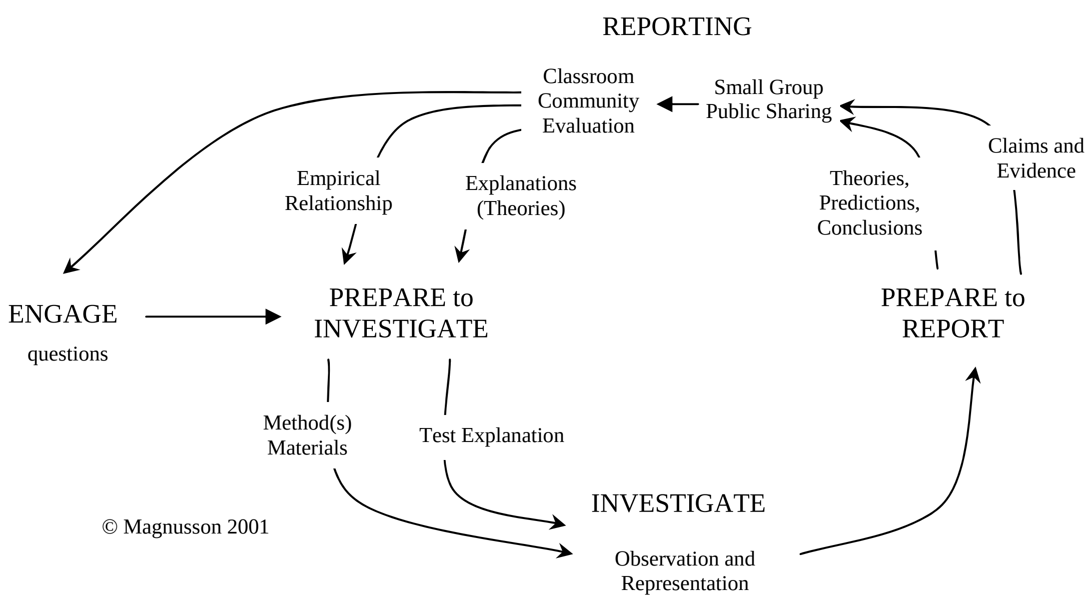


Figure 2: Another Model of the Scientific Inquiry Process – The Guided Inquiry Cycle proposed by Magnusson & Palinscar (2005). (Note. From “Teaching to promote the development of scientific knowledge and reasoning about light at the elementary school level” by S. Magnusson & A. Palincsar, in *How Students Learn: History, Mathematics, and Science in the Classroom*, M. Donovan & J. Bransford (Eds.), Figure 10-1, p. 427, Copyright 2005 by the National Academy of Sciences. Reprinted with permission.)

The guided inquiry model is based on three principles: cycles of engagement, investigation, and reporting. While the model is cyclical, Magnusson & Panliscar (2005) proposed that there are two parallel emphases that take place in a science inquiry cycle. The outer loop of the cycle focuses on the development of science inquiry skills such as method design or the use of data to support a claim. The inner loop, in contrast, focuses on the development of science concepts and content. While separate and distinct, the investigation and reporting phases both begin with a planning space. Prior to the activities of an investigation, students must first plan what it is they intend to do and how they intend to do it. In a similar way, prior to reporting their results, students begin by thinking through what it is that they want to communicate and then determine the most effective way of communicating these ideas to their audience.

A fundamental assumption of this model is that students must be taught how to think in scientifically literate ways. Magnusson and Palincsar (2005) note, "Engaging children in science, then, means engaging them in a whole new approach to questioning. Indeed it means asking them to question in ways most of us do not in daily life" (pg 426). In order to successfully engage students in science inquiry as this model intends, students must develop the metacogntive skills of "thinking about their own thinking" that are necessary to facilitate this kind of exploration. These metacognitive skills are developed as students engage in developing new conceptual understandings, build on prior knowledge, and reflect on their own learning as they move through the phases of this guided inquiry process.

In this model, inquiry is initiated at the engagement stage. At this stage, students identify what they already know and propose questions they may have related to a specific scientific idea. In this phase, students also identify how it is that they have come to "know" the knowledge that they possess. As students reflect on their current knowledge, their task is to create a question or problem that will guide their subsequent inquiry activities. This phase focuses primarily on metacognition and the activation of prior knowledge (i.e. How to do I know what I know?).

Once students have identified a question / problem to explore, the next task is to outline a set of procedures that will provide data relevant to the question / problem at hand. How the procedures are developed (given directly from a teacher, teacher / student collaboration, students ideas alone) is secondary to the understanding of why various procedures are important. What does it mean to have a control? What data should be collected? How will you know when you have collected enough data? What is the importance of various steps in the process? This also includes thinking about how to construct data tables and what roles various individuals will play if this exploration is being done in a group setting. In this step of the cycle, the emphasis is on developing the metacognitive understanding of *why* something is being done as opposed to simply understanding *what* each step entails.

The investigation phase has significant overlap with the planning stage and although they are separated in this model, in many ways, they form two parts of the same process. As students begin to investigate their question, the goal is that students monitor their activity to assess ideas such as: Is this working? Does this fit what I expected to see? Will these procedures produce the intended data? Did anything happen that I did not anticipate? Do the procedures need to be revised / modified? While students formed their initial procedure in the Prepare to Investigate phase, the planning process is actually ongoing as they carry out their experiment. Again there is a strong emphasis in this stage on metacognitive thinking about the process in addition to the pursuit of data. In this stage as well, Magnusson & Palincsar (2005) highlight that the initial movement towards student conceptual understanding should start to begin in this phase as well. Students should begin to not only look at what data they have collected, but they should also start the initial thinking about what this data actually means. This provides the opportunity to continue to collect data if students feel that this is necessary.

Once the data has been collected, the students move to the stage where they prepare to report their findings. This model follows a claims / evidence format where students make claims and then demonstrate how their data supports these claims. The focus of these claims is on conceptual understanding such as identifying patterns, constructing models, or relating findings to other ideas in science. In this model, the students present their findings to the rest of their classmates in a class discussion. As they prepare to present, they must consider: Who is my audience? What can be done to make my claim most compelling? How should the data be presented? Are there other ways to explain the data? Why are these alternatives more / less desirable? The focus has shifted away from data collection to conceptual development. Again, metacognition plays a key role in this phase.

The final phase of this guided inquiry cycle is that of actual reporting. Students report out to their peers what they chose to investigate, the data they collected, and their subsequent conclusions. The role of the class is to provide a chance to critique and analyze the work presented. Students do this by asking questions, clarifying misunderstandings, critiquing other group’s data and subsequent conclusions, proposing alternative explanations, etc. The goal of this phase is to generate questions / problems which could potentially lead to another cycle of inquiry so that the process is then repeated.

The guided inquiry framework is well suited to this study. It envisions a science inquiry process as composed of two distinct inquiry phases. Both phases involve the construction of a plan and then the subsequent actions that carry out that plan. Both phases ask students to be metacognitive about their work. The main difference, though, is in the products that are produced from each phase. In the Investigation phase, students produce data. In the Reporting phase, the students produce a product that allows them to share their data and conclusions with others. Magnusson & Palincsar (2005) noted that in a study of fourth graders studying the concept of light, the reports were expected to "include a statement of the group's knowledge claim(s) as well as data backing up the claim(s)" (pg 444). In addition, students were asked to consider "how to present their findings to best enable others to understand them, and be convinced of the group's claim" (pg 444). Magnusson & Panlicsar (2005) place communication on equal footing with investigation.

The implication of this framework is that learning from science inquiry tasks requires being proficient in both the Investigation and the Reporting phases. While distinct, the phases are complementary. Both phases have an end goal in mind and a process that is developed to meet the requirements of that target. Both phases ask that students monitor their progress. The main distinction, though, is in what the students are asked to do during the phase. In the Investigation phase, the focus is on experimentation and manipulation of tools in order to collect data. In contrast, the Reporting phase places a much greater emphasis on a student's ability to communicate. Looking at inquiry as a two-stage process leads to a question this study sought to address. Do students struggle with inquiry because of limited skills in their abilities to conduct scientific investigations, or in their abilities to report their results once the data has been collected, or rather, some combination of the two?

A strength of the guided inquiry framework is the extent to which it has been used to inform other studies in science education. The framework has been used to teach concepts such as sound (Magnusson, 1996), animals (Magnusson, 1996), light (Magnusson & Palincsar, 2005; Palincsar, Magnusson, Collins, & Cutter, 2001), and floatation (Palincsar, Magnusson, Collins, & Cutter, 2001). In addition, the framework has been used to help inform instruction with both mainstream students (Magnusson, 1996; Magnusson & Palincsar, 2005) and students who require special accommodations (Palinscar, Collins, Marano, Magnusson, 2000; Palincsar, Magnusson, Collins, & Cutter, 2001). All of the work on the guided inquiry framework up to this point, though, has been at the elementary level. This study seeks to extend the use of this framework into a high school setting.

Knowledge Telling versus Knowledge Transforming

While the principles of science inquiry are broadly applicable across grade levels, the nature of what constitutes sufficient output from the Reporting phase will change based on grade level. At the high school level, greater skills and abilities in writing are required for the Reporting phase. Since the guided inquiry framework does not speak as specifically about the writing process, another framework perspective was needed to inform the writing process in this study.

The focus on writing was chosen for two reasons. First, there was evidence in the literature that suggests that students do indeed possess the abilities to carry out valid investigations with appropriate scaffolding. At the elementary level, Magnusson & Palincsar (2005) documented the natural curiosity of fourth grade students and also tracked students' abilities to conduct investigations with the phenomena of light. Also at the elementary level, Flick & Tomlinson (2006) worked with students as they investigated the nature of circuits and electricity. At the high school level, White & Frederiksen (1998) worked with 9^{th} graders who were studying the physics of motion. In that study, students worked with computer models and simulations, and demonstrated that they were capable of designing and conducting scientific investigations. Also at the high school level, Stewart, Cartier, & Passmore (2005) studied students as they carried out inquiry investigations in a high school biology classroom. The students were studying genetics using a computer program that generated hypothetical offspring data from fruit fly crosses. These studies support the idea that students across a wide age range do indeed possess the abilities necessary to conduct investigations in scientific inquiry.

Second, writing has become a recent focus in science education. Both Yore & Treagust (2006) and Klein (2006) commented on the difficulty that students face as they write in science. Yore & Treagust (2006) state that, "Effort must be given to both inquiry teaching and embedding language issues". They continue, "Additional research is needed to document how students learn to talk, write, and read science". Scaffolded instruction in science writing has been linked to student gains in science content (Hohenshell & Hand, 2006) and increases in the quality of student writing products (De La Paz & Graham, 2002; McNeill, 2009). Warrick et al (2003) looked specifically at the effect of sentence frames on students' ability to understand science procedures (identifying controls and variables, recognizing the need for multiple trials, etc.). They found when students used sentence frames, students were more able to express their understanding of science processes. Warrick et al (2003) commented, "There is, as yet, no research evidence to suggest in which areas of science the use of writing frames may yield better outcomes." The intent of this study was to build on this body of knowledge and investigate ways to help students become more effective in their abilities to communicate scientifically.

To understand how and why the process of writing might present difficulties for students, this research project used the theoretical models proposed by Bereiter and Scardamalia (1987). They stated that writing can often be described as the product of one of two modes. A knowledge-telling mode is one that requires fewer cognitive resources and skills, but often produces work of that is of limited quality. In contrast to this is the knowledge-transforming mode. A knowledge-transforming mindset requires greater cognitive engagement, but it often results in higher quality writing products.

The knowledge-telling model grew out of Bereiter & Scardamalia's (1987) work with novice and expert writers. They noted that novice writers often expressed that their greatest struggle in writing was that of finding adequate content in response to a prompt. Even though novice writers quickly came to a place where they felt that they had nothing more to write, Bereiter & Scardamalia (1987) found that when these novice writers were given simple prompts such as "I know that this is tough, can you write more about this?", the word output doubled.

In addition to limited word output, writing done from a knowledge-telling perspective tended to lack a clear purpose. Knowledge-telling writing often seemed to be a written "stream of consciousness" or a simple list. Students would write down words that related to a specific prompt in the order that the words came to mind. It appeared that little consideration was given to the text that came before or the text that came after. For example, Bereiter & Scardamalia (1987) gave an example of a student's work in response to a prompt about whether or not boys and girls should play sports together. The student wrote, "I think they should because sports are for girls and boys and there is no difference between girls and boys. The girls might not be good in sports, so that's why the boys don't like the girls to play" (pg 121). While one sentence was in support of girls and boys playing sports together, the next sentence actually contradicted what the first one said. Bereiter & Scardamalia (1987) proposed that this knowledge-telling model produced this "stream of consciousness" writing because the novice writer was attempting to write text, but by using conventions from conversational speech.

This poses a significant problem for novice writers. Conversation contains natural breaks, and allows for others to assist in providing ideas. In contrast, writing from this conversational perspective produces a much greater cognitive load because the writer is carrying the entire conversation by them self. Ideas are written as they come to mind because content generation is the foremost concern for these novice writers. The content does not necessarily need to connect in any way. Novice writers simply want the content to be present with the hope that the reader will make sense of the information. Lacking knowledge of any alternate ways of constructing text, novice writers become very proficient at this knowledge-telling strategy.

This knowledge-telling strategy is not limited to young writers either. Many high school and college student utilize the same knowledge-telling strategy on essay type exams. If they come to a question that cannot be answered, the typical response is to write down everything that the student can recall that is even remotely associated with the prompt. Brown et al (1983) commented that strategies such as knowledge-telling are particularly resistant to change because they are effective in common day to day exchanges where the product demands are less stringent. Also, in studies that looked at how student strategies change over time, it was evident that students were reluctant to give up less adequate strategies for more effective ones (Brown et al 1983). This was true even as students were learning the new strategies. This reluctance to give up the less effective, but more attractive strategies actually posed a greater challenge to learning than did the work of learning new, more effective strategies. The implication is that it is not enough to simply teach students more effective writing strategies, students must also be convinced that the knowledge-telling strategy is not adequate to the task.

This does not imply that the knowledge-telling strategies are deficient or undesirable in all circumstances. In fact, knowledge-telling strategies may be an excellent mode of response for a particular prompt. If knowledge recall is all that a prompt requires, then knowledge-telling is more than adequate. The process of reporting in scientific inquiry, though, is meant to be more than simply a knowledge-telling event. Reporting in scientific inquiry is intended to be a process that expands student thinking and helps students build on their prior knowledge. As will be seen, the knowledge-telling model requires fewer cognitive resources than the knowledge-transforming process. While student reasons may vary, a contention of this proposal is that one reason students do poorly when asked to communicate their findings in science inquiry is because they adopt a knowledge-telling mindset. This strategy produces work that does not align with the knowledge-transforming intent of scientific inquiry and so the result is a product of minimal quality.

The knowledge-telling model (Bereiter & Scardamalia, 1987) is a linear process (see Figure 3). Based on a prompt, the writer constructs a mental representation of the assignment. This mental representation dictates both the content and the writing style (narrative, expository, etc.) that will meet the requirements of the prompt. The writer then does a mental search for ideas and matches these ideas to the writing genre constraints related to the prompt. Content recalled from memory that meets these requirements for appropriateness is written down and this process is then repeated until the writer feels that either enough has been written or there is nothing left to say.

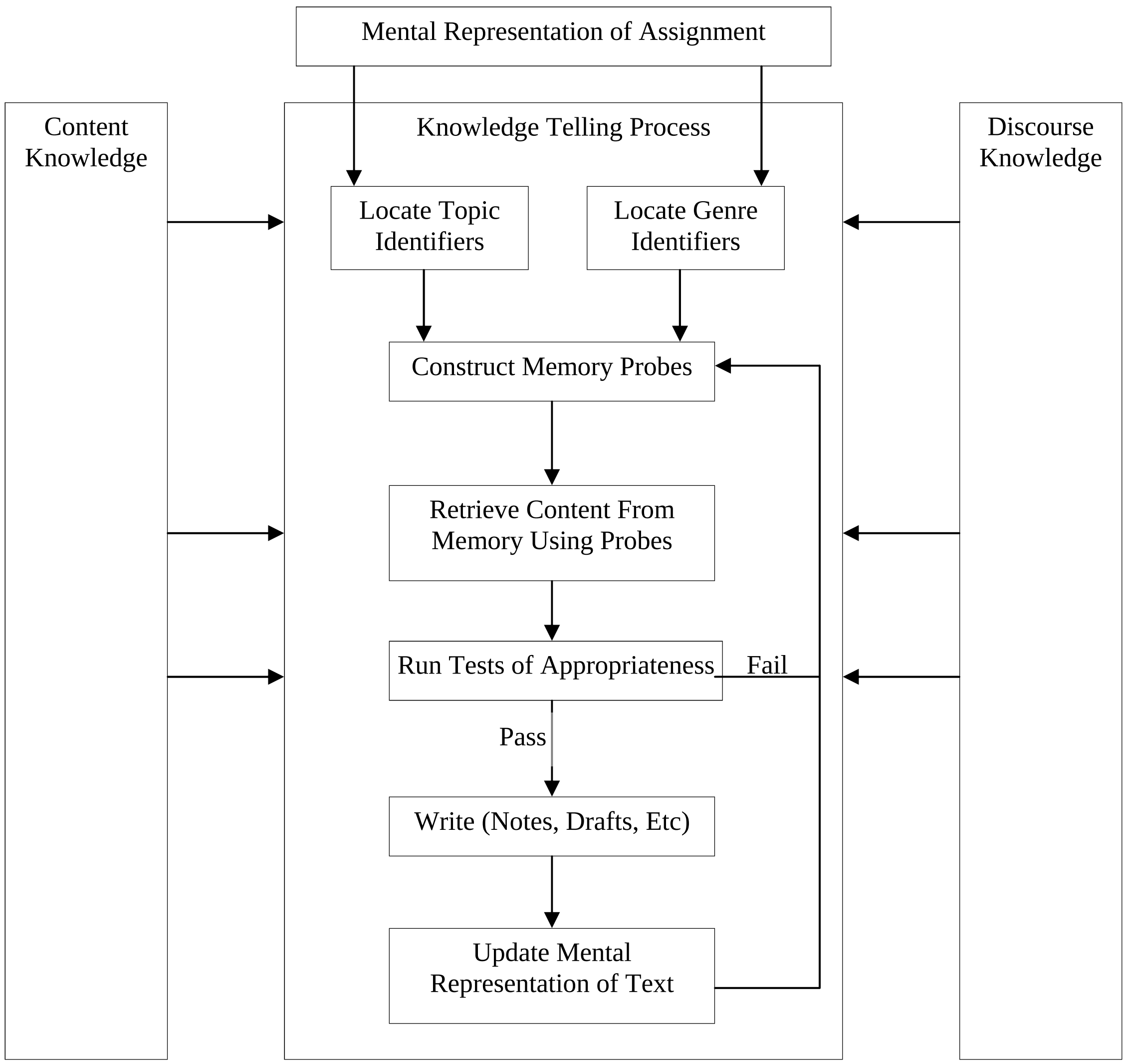


Figure 3: The Knowledge-Telling model of writing proposed by Bereiter & Scardamalia (1987). (Note. From “Two models of composing process”, by C. Bereiter & M. Scardamalia, The Psychology of Written Composition, p. 8. Copyright 1987 by Lawrence Erlbaum Associates, Inc. Reprinted with permission.)

Brown et al (1983) stated that there were four traits that described writing from a knowledge-telling perspective. First, the writing lacked any goal related planning.

There was no clear sense of a beginning or end to the writing, nor was there a clear sense of purpose to what was written. Second, there was no sense of cohesion within the text from one sentence to the next. The sentences did not build ideas, rather they stated unconnected thoughts. Third, the writing occurred in a forward-only manner. There was no indication that a forward-backward revision process took place as the writing occurred. Finally, if any revision were done between a draft and final copy, the revisions were merely cosmetic.

Here are two examples of what knowledge-telling writing looks like. They come from Bereiter & Scardamalia (1987, pg 163) in a study where they asked elementary students to write for 15 minutes on the topic “Should students be able to choose what things they study in school?” These two examples are the complete works that the students wrote, and they are copied directly from the text (including the students’ punctuation and spelling)

> In School We Should Be Able To Do Any Kind Of We Want To Do We Are Free We Could Do Anything We Want God Us Free We Could Do Anything We Want To Do I’d Like Spelling And Math In School We Should Do It Any Time We Want
>
> Spelling is my subject because I like it because it is fun to do it, and right after one lesson I can go one to another lesson. Spelling sounds exciting, to me because I can get high marks, and when I have Spelling tests, I get high marks like: 25 out of 25, and 24 out of 25, and those look like good marks to me.

In both of these examples, it is clear to see the evidence of a knowledge-telling heuristic. Both examples lack a clear sense of direction or planning. In the second example, the sentences show more of a focus on the topic of spelling, yet there is no clear ordering to this student’s thoughts. The order of the sentences could be rearranged, and yet the meaning of the second paragraph would remain unchanged. Again, this is not to say that a knowledge-telling strategy is never warranted, but in terms of reporting in science inquiry, the demands of the task require that a superior methodology be used.

In contrast to the knowledge-telling model, Bereiter & Scardamalia (1987) have proposed the knowledge-transforming model (see Figure 4). In this model, writing is seen as a process of discovery and as a means of creating new understandings about the topic at hand. Whereas knowledge-telling involves simply restating what one has stored in memory, the knowledge-transforming process attempts to lead the writer to synthesize their current understandings in the construction of new ideas. As Bereiter & Scardamalia (1987) note, the knowledge-transforming model seems to better approximate the cognitive and metacognitive actions of expert writers.

Similar to the knowledge-telling process, the knowledge-transforming process begins with building a mental representation of the text as dictated by a prompt. Unlike the knowledge-telling process, though, the next stage in knowledge-transforming process is where the writer develops a general plan as to how the prompt might be addressed. They ask them self: What style of writing would be most effective? What kinds of arguments would be most persuasive? How should the ideas be ordered? In this phase, the writer also begins to anticipate potential problems that might arise.

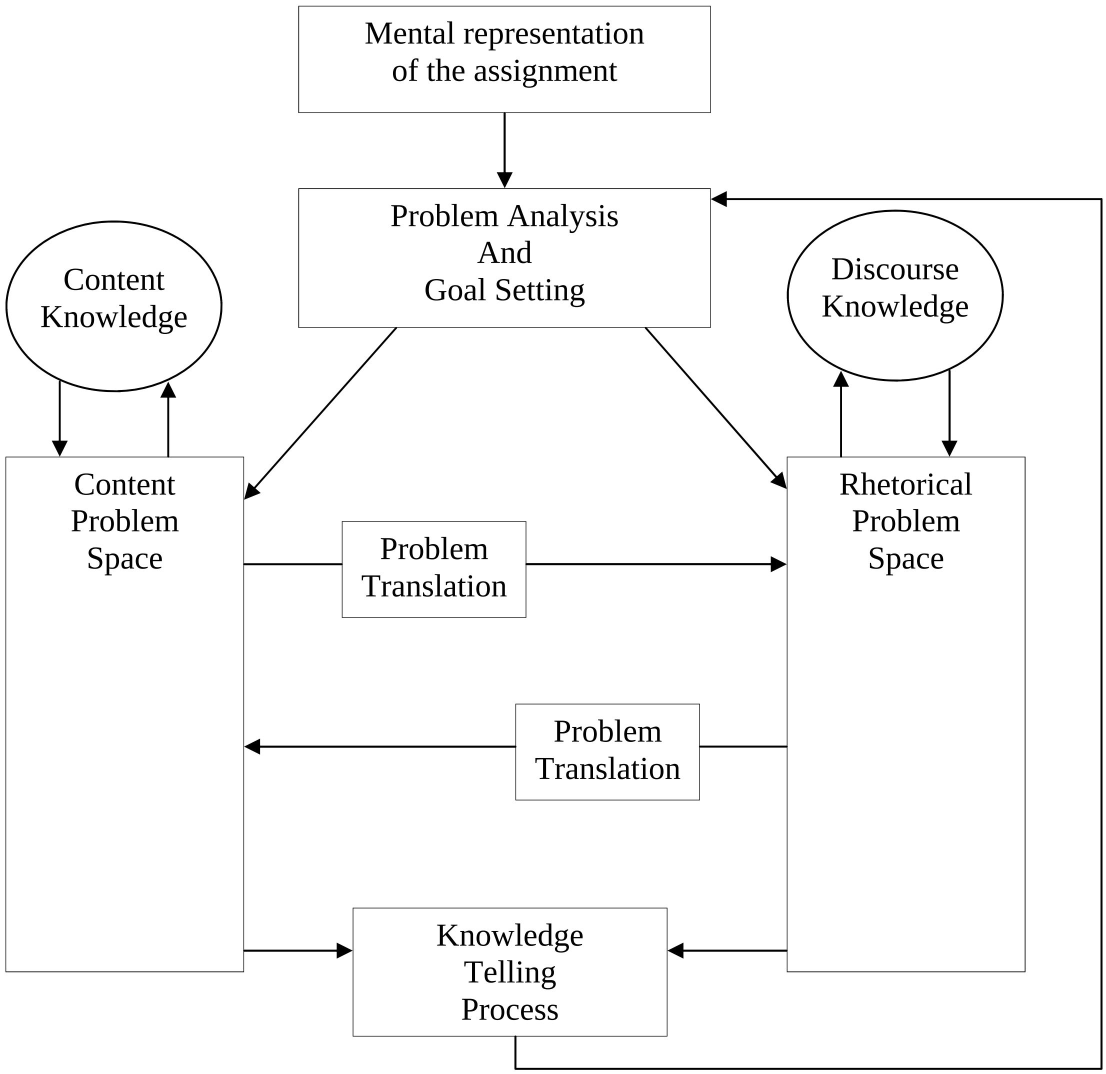


Figure 4: The Knowledge-Transforming model of writing proposed by Bereiter & Scardamalia (1987). (Note. From "Two models of composing process", by C. Bereiter & M. Scardamalia, *The Psychology of Written Composition*, p. 12. Copyright 1987 by Lawrence Erlbaum Associates, Inc. Reprinted with permission.)

These problems tend to be one of two types. One problem addresses issues related to content and assessing the extent to which the content is accurate. The other problem addresses issues specific to the limitations of language and text. Rules related to diction and syntax combined with the often ambiguous nature of language require that extra care be taken by the writer to ensure that meanings and intentions are clearly communicated.

Once goals have been established and an initial problem analysis has taken place, the actual process of writing begins. The writing process in the knowledge-transforming model is significantly different from that of the knowledge-telling process. Whereas the writing process in knowledge-telling is fairly linear, in the knowledge-transforming model, text production is the result of an ongoing interaction between two distinct spaces of content knowledge and discourse knowledge. Bereiter & Scardamalia (1987, pg 11) describe these spaces as,

> In content space, problems of belief and knowledge are worked out. In rhetorical space, problems of achieving goals of the composition are dealt with. Connections between the two problem spaces indicate output from one space serving as input for the other.

As an idea is formed, this idea is sent from the content space to the rhetorical space where issues of diction and syntax are worked out. As problems arise related to diction and syntax, these ideas are sent back to the content space to see if there are other parallel ideas or knowledge that could be utilized in order to remove any perceived ambiguity. As these problems are resolved, the writer produces text. Written text is then analyzed against pre-determined goals. This written text also provides a source of feedback to analyze if the pre-determined goals are still adequate or if these goals need to be revised. This process continues until the writer feels that the text has met the desired outcomes. Bereiter & Scardamalia (1987, pg 11) go on to state that, "It is this kind of interaction between problem spaces that we argue…is the basis for reflective thought in writing." It is also this kind of reflective, goal directed thought, I would argue, that science teachers desire to see in their students through the process of scientific inquiry.

The following is an example of what writing looks like from this knowledge-transforming perspective. The prompt and time constraints were the same as the previous two examples about subject and student choice. Again, this example is from Bereiter & Scardamalia (1987, pg 166). Student spelling and syntax were preserved.

> I think you should not be able to choose you own things to study because of a lot of reasons. The first reason is that some kids might pick easy subjects to learn. Also, I wouln't know what to learn about it, or write about. I think it alright for projects to do at home to pick a subject but not at school. If you they were to pick there own book for reading theyed propely pick the easiest one there. For math they would propely just do grade on and two work, and not learn any new things. School won't be like school if you picked your own subject.

This student paragraph shows a sequence of arguments that all stem from the main idea that students should not be the ones to pick what subjects they study. In addition, the ideas as they are presented seem to build on each other in the writers mind. At the close of the paragraph, there is a distinct conclusion that summarizes what the student was trying to say. This paragraph is not perfect by any means, but that is not the point. This paragraph is much more focused and cohesive than the two previous writing examples (see page 21), and this is the type of writing needed for students to be successful on science inquiry tasks.

The aims of guided inquiry and the aims of knowledge-transforming writing are complementary. Magnusson & Palincsar (2005) stated that one of their goals with guided inquiry was to help students ask questions in ways that are different from how most people ask questions in day to day life. Brown et al (1983) stated that one of the reasons that students tended to resist abandoning a knowledge-telling framework is that the knowledge-telling framework is what many people use in daily interactions. Guided inquiry and knowledge-transformational writing are both asking students to work and think in ways that are different from their common daily experiences.

In addition, guided inquiry and knowledge-transforming writing share the same metacognitive strategies. In guided inquiry, both the Investigation and the Reporting phases begin with a planning space. Students choose and design processes that are matched to their question of interest. The chosen methodology must not only address the question of interest, but must also match the intended outcome. Both the Investigation and Reporting phases are goal directed in that for both phases, there is an end product in mind. The same process is mirrored in knowledge-transformational writing. Also, in guided inquiry, as the plan is being implemented in both the Investigation and Reporting phases, students must still monitor their activity to see if what they are producing matches their intended outcome. In knowledge-transformational writing this is also known as monitoring meaning, where students continually rework, and revise their work so that the writing matches their intended meaning. The cognitive interaction between Content space and Rhetorical space that the knowledge-transformation model presumes (Bereiter & Scardamalia, 1987) is analogous to the parallel emphases of content and process that Magnusson & Palincsar (2005) emphasize as a part of the guided inquiry model. Putting this all together, then, the two inquiry phases of Investigation and Reporting (the phases include both planning and execution) that form the infrastructure of the guided inquiry cycle are in fact knowledge-transformational experiences that require students to think in ways that are often different from their every day experiences.

In a review of the literature, only one article was found where this knowledge-transforming framework informed a research project in science education. Coleman (1998) applied this model to work with 4th and 5th grade students. The goal of this particular research, though, was to determine if elementary students were able to process new information in a knowledge-transforming way. Coleman's results indicated that these students did indeed possess the cognitive capacity to utilize knowledge-transforming strategies. The limitation, though, is that the Coleman study (1998) only looked at the development of science ideas while students took part in group discussions. Also, work with the knowledge-transforming model has taken place predominately at the elementary level (Bereiter & Scardamalia, 1987; Scardamalia & Bereiter, & Steinbach, 1984). This study will seek to extend this work to high school students.

Self-efficacy

Because of the noted resistance to change that may occur with students as they learn new skills and strategies (Brown et al, 1983), a construct was needed that could measure the extent to which cognitive change was taking place. To that end, the construct of self-efficacy was chosen. Self-efficacy is an individual's assessment of their ability to carry out a specific task (Bandura, 1986, 1997). It is fundamentally a belief statement that may or may not correlate with an individual's actual ability to successfully complete a specified task. Of interest to this study, an individual's self-efficacy is highly predictive of their actions. Bandura (1997, pg 2) states, "people's level of motivation, affective states, and actions are based more on what they believe than on what is objectively true." An individual's self-efficacy is a stronger predictor of behavior than actual ability.

This relationship between self-efficacy and action has been studied extensively in education. In studies of student success, student self-efficacy is often the greatest predictor of academic achievement (Bandura, 1995, 2006a; Pajares, 2006; Zimmerman 1995; Zimmerman & Cleary, 2006). In addition, students with higher measures of self-efficacy often persist longer with tasks and set higher goals for themselves (Pajares, 2006; Zimmerman 1995; Zimmerman & Cleary, 2006). Since student self-efficacy plays such a pivotal role in student success, understanding how to help students increase their self-efficacy should also help these students be more successful.

Bandura (1986, 1997) has proposed that self-efficacy is most often affected through one of four means: mastery experiences, vicarious experiences, verbal persuasion, and affective states. The simple experience of these means, though, is no guarantee that one's self-efficacy will be affected. Cognition plays a vital role in translating and interpreting these experiences in order for the individual to recognize how that experience relates to their specific situation (Tschannen-Moran et al. 1998; Woolfolk Hoy & Davis, 2006; Zimmerman, 2006). In the same way that metacognition is an essential component of the guided inquiry model, this same sort of metacognitive processing is central to bringing about change in one's self-efficacy.

While there has been extensive research on the role of self-efficacy in adolescent education, (Bandura, 1995; Pajares & Urdan, 2006), a clear model of how adolescent self-efficacy can be changed has not yet been outlined. Bandura has argued that the factors effecting change in an individual's sense of self-efficacy are broadly applicable across age groups (Bandura, 1997). Instead of creating a model, then, for effecting change in adolescent self-efficacy, this study modified a framework proposed by Tschannen-Moran, Woolfolk Hoy, and Hoy (1998).

Building on the work of Bandura (1986, 1997) Tschannen-Moran, Woolfolk Hoy, and Hoy (1998) suggested the following model for affecting teacher self-efficacy (see Figure 5). While this model is specific to teaching, note the role that cognition plays in this recursive cycle. As a teacher processes a new source of efficacy information, the cognitive process is divided into two categories. Part of the cognitive process involves a teacher's analysis of the teaching task itself. The teacher assesses if they have the tools and resources necessary to successfully complete the task. The second part of the cognivite process involves a teacher's assessment of their own personal competence to actually carry out the task. The cognitive processes of both task analysis and personal competence, then, feed into and inform a teacher's sense of self-efficacy. This sense of efficacy affects the goals the teacher sets and even the amount of effort this teacher will be willing to expend in meeting the goal. Pajares states (2007)

> Self-efficacy beliefs also help determine how much *effort* people will expend on an activity, how long they will *persevere* when confronting obstacles, and how *resilient* they will be in the face of adverse situations. The higher the sense of efficacy, the greater the effort, persistence, and resilience. People with a strong sense of personal competence approach difficult tasks as challenges to be mastered rather than as threats to be avoided.

In this way, self-efficacy affects performance and then performance feedback in turn affects self-efficacy and so this cycle repeats itself.

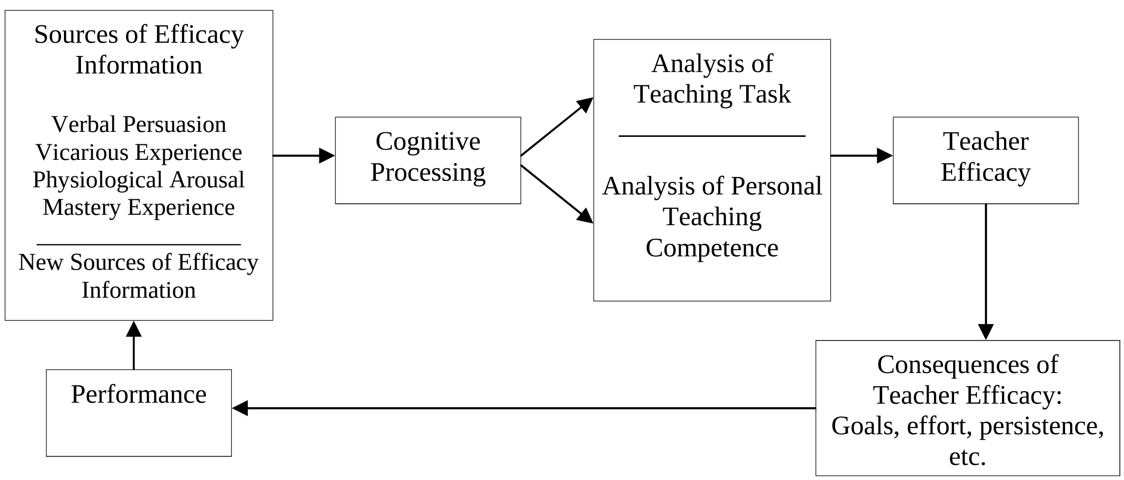


Figure 5: A model of the cyclical nature of Teacher Efficacy from Tschannen-Moran, Woolfolk Hoy, & Hoy (1998). (Note. From "Teacher efficacy: its meaning and measure", by M. Tschannen-Moran, A. Woolfolk Hoy, & W. Hoy, 1998, *Review of Educational Research,* 68(2), p. 228. Copyright 1998 by Sage Publications Inc. Reprinted with permission)

In a similar fashion, I propose that the same model is applicable to student self-efficacy (see Figure 6). In order to bring about a change in student self-efficacy, there must be a cognitive assessment of the efficacy information available. The cognitive processing would also be a two-fold process. A student would assess the extent to which they have the necessary tools and resources necessary to complete a specific task. In addition, they would also assess the extent to which they feel that they have the personal skills required to complete a given task. Similar to teachers, this cognitive processing would lead to changes in efficacy beliefs resulting in changes in performance. In relation to the guided inquiry model, student self-efficacy related to science inquiry is actually, then, the combination of a student's self-efficacy to carry out two distinct tasks: a student's self-efficacy to carry out an investigation and a student's self-efficacy to report one's findings.

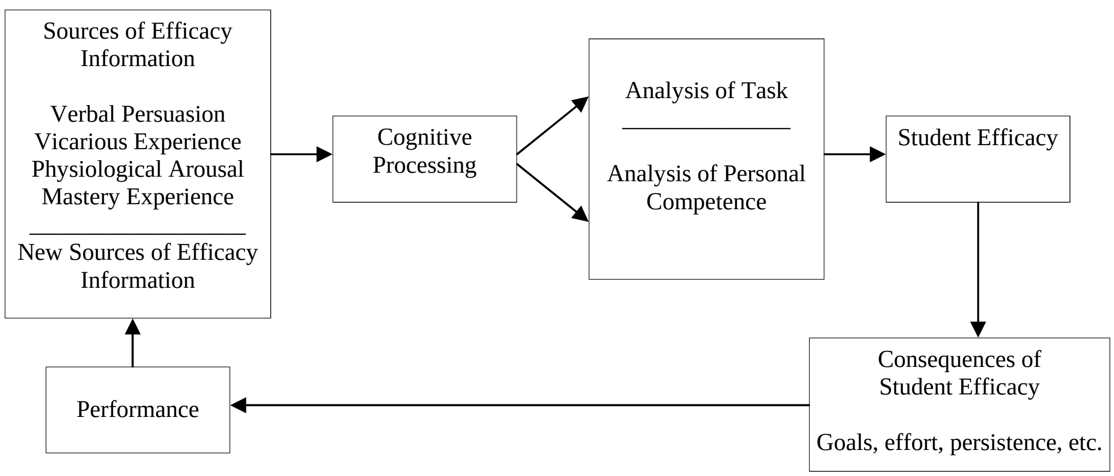


Figure 6: Proposed cyclical nature of Student Efficacy adapted from the work of Tschannen-Moran, Woolfolk Hoy, & Hoy (1998). (Note. This figure adapted from "Teacher efficacy: its meaning and measure", by M. Tschannen-Moran, A. Woolfolk Hoy, & W. Hoy, 1998, *Review of Educational Research,* 68(2), p. 228. Copyright 1998 by Sage Publications Inc. Reprinted with permission)

Also of interest, gender seems to play a role in a student's self-efficacy. The research in this area, though, has produced mixed results. Pajares (2003) found that 9^{th} grade boys tended to have a greater sense of self-efficacy in relation to writing than their female peers, but girls tended to have better writing scores. A study by Zimmerman & Martinez-Pons (1990) found similar results in that boys from grades 5 to 11 tended to have greater verbal self-efficacy than their female peers. In contrast, Jacobs et al (2002) found that self-efficacy of girls from kindergarten through grade 12 were consistently higher than their male counterparts. Self-efficacy related to math also showed the same inconsistency. Zimmerman & Martinez-Pons (1990) found male and female students did not differ significantly in their mathematics self-efficacy. Jacobs et al (2002) in contrast, found that males tended to have greater self-efficacy in math from kindergarten through grade 10. Students need to be proficient in both math and language skills in order to be successful on a science inquiry task. As research has indicated that gender may play a role specifically in science inquiry, gender will be investigated to see if it plays a role in student achievement.

Self-Regulation

One final framework informed the nature and direction of this study. In many ways, self-regulation was the driving construct behind this research proposal and was the construct that helped to tie each of the previous three frameworks together. Broadly defined, self-regulation is the combination of one's self generated thoughts, feelings, and actions that are planned and adapted in order to attain a specified goal (Zimmerman, 2000; Zimmerman & Cleary, 2006). While metacognition plays an important role in self-regulation, Zimmerman's model (2000, Zimmerman & Cleary, 2006) added a social cognitive perspective that included a strong self-efficacy component. It is not enough to have the metacognitive capacity to self-regulate one's actions, one must also possess the efficacy beliefs that they are indeed able to carry out these actions as well.

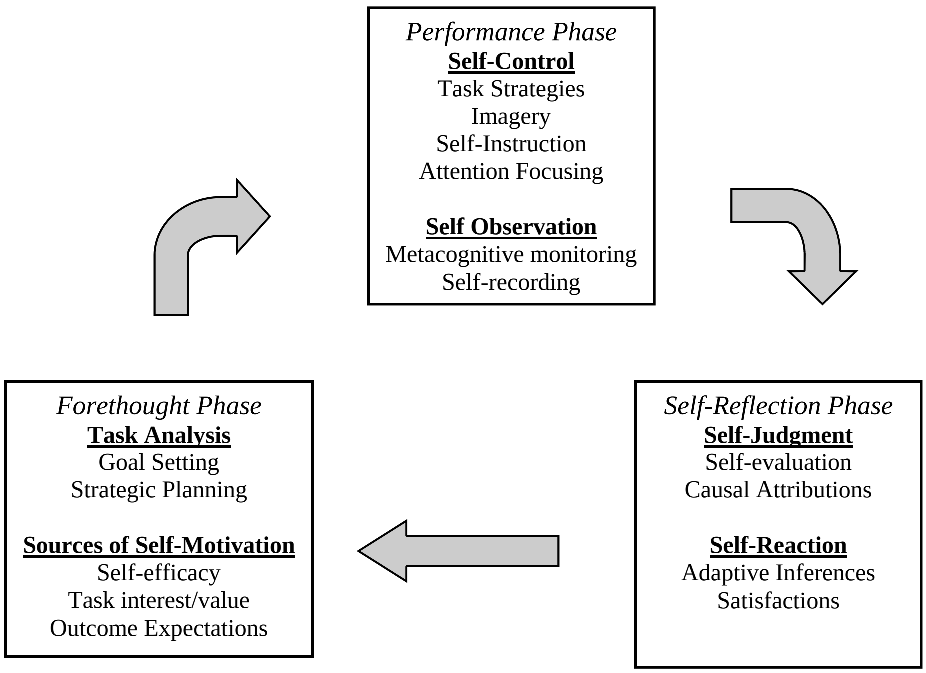


Figure 7: Phases and subprocesses of Self-Regulation proposed by Zimmerman & Campillo (2003). (Note. From "Motivating self-regulated problem solvers", by B. Zimmerman & M. Campillo (2003), in *The Psychology of Problem Solving*, J. Davidson & R. Sternberg(Eds.), Figure 8.1, p.239. Copyright 2003 by Cambridge University Press. Reprinted with permission.)

Zimmerman proposed that self-regulation can be conceived of as a three phase process (see Figure 7). The initial phase is called the Forethought phase. In this phase, goals are set and specific strategies are identified that will aid in the attainment of those goals. Proximal goals tend to promote success at a greater rate than distal goals. Proximal goal orientation also plays a significant role in self-efficacy development (Bandura & Schunk, 1981; Page-Voth & Graham, 1999; Zimmerman & Kitsantas, 1999). In the same way, goals that are more specific in nature tend to be more advantageous than general goals. In relation to strategy choice, self-efficacy has also been demonstrated to play a significant role (Zimmerman & Cleary, 2006). In a number of studies that were reviewed, individuals chose strategies based on their belief that they could successfully carry out these strategies (Schunk & Ertmer, 2000; Zimmerman, 1995, 2000; Zimmerman & Cleary, 2006). In many cases, less successful strategies were chosen because the individuals did not feel that they were capable of executing strategies that were more closely correlated with success. This many help to explain why some students are so resistant to adopt new, more effective strategies as was observed by Brown et al (1983).

In relation to the models relevant to this study, this Forethought phase shows up in the following ways. In the guided-inquiry model (Magnusson & Palincsar, 2005), the Forethought phase actually occurs twice. During the Prepare to Investigate phase of the model, students determine the goals of their investigation and select appropriate strategies. Individuals sense of self-efficacy plays a key role in determining what goals will be pursued in addition to what types of strategies will be utilized. A similar process takes place as students prepare to report. Again, this process is goal directed and influenced by self-efficacy.

In the knowledge-transforming model of writing, the problem analysis and goal setting phase also mirrors the Forethought phase of the self-regulation model. Again, while Bereiter & Scardamalia (1987) do not explicitly talk about the role of self-efficacy in strategy choice with the knowledge-transforming model, social cognitive theory implies that self-efficacy must play a role in which strategies and goals are pursued. In the self-efficacy model proposed by Tschannen-Moran, Woolfolk Hoy, and Hoy (1998), the cognitive activity of the forethought phase is very similar to cognitive processing of both the task and personal competence leading to an efficacy judgment.

The Performance phase follows the Forethought phase. In the Performance phase, metacognitive monitoring plays a primary role. The individual monitors the extent to which their actions will lead to desired outcomes, isolates sources of error, and also uses pre-determined strategies to plan subsequent actions. Zimmerman & Cleary (2006) noted the ongoing role of self-efficacy in this process. They stated (pg 62), “there is a reciprocal relation between efficacy judgments and self-monitoring behaviors.” An individual’s self-efficacy will determine their courses of action. As one monitors their performance, this feedback serves to reinforce, either positively or negatively, ones efficacy beliefs. These new beliefs then in turn effect decisions about future courses of action, and so the cycle continues.

This metacognitive monitoring is explicitly stated in the guided-inquiry model (Magnusson & Palincsar, 2005). There is a greater use of metacognitive monitoring that takes place in the investigate phase as compared to the reporting phase, but metacognition plays a role in each phase, nonetheless. In the knowledge-transforming model (Bereiter & Scardamalia, 1987), the interplay between the cognitive space and the rhetorical space is, in fact, a metacognitive interchange. Both spaces monitor the text that is produced.

The final phase in this self-regulation cycle is where the individual reflects on their prior performance. In the Self-Reflection phase, the individual evaluates the extent to which they were satisfied with the outcomes of their performance, the value of chosen strategies, and the causal attributions of the performance. Again, self-efficacy plays a key role in self-evaluation in that students who feel confident about their abilities to self-regulate a task tend to set higher performance standards for themselves and are less satisfied with mediocre performance (Zimmerman & Cleary, 2006). When students were asked to evaluate why they were or were not successful, highly efficacious students tended to cite factors that were within their ability to control. In contrast, students with low self-efficacy tended to list factors that were beyond their control (Zimmerman & Cleary, 2006). One other interesting finding related to efficacy and self-evaluation was that when students were given feedback to aid in self-evaluation, these students often exhibited an increase in their efficacy (Schunk & Swartz, 1993; Zimmerman, 1995; Zimmerman & Cleary, 2006).

Combined Framework

Merging the ideas of self-regulation with the guided-inquiry cycle yielded the model that this study employed (see Figure 8). The ideas from the Forethought Phase of the self-regulation framework informed the nature of the goals that students set as they prepared to investigate and also as they prepared to report their findings.

Strategic planning was already a part of the guided inquiry model, and self-regulation theory reinforced the importance of this activity.

The Performance Phase of the self-regulation cycle was identical to what takes place in the guided inquiry cycle at both the Investigation and Reporting phases. To be successful, students would need to assess the tasks they were completing in order to determine if their work was being done according to the outlined plan. Students would also need to monitor their work to determine if what they were doing was leading to data that made sense.

The unique contribution that the self-regulation framework added to the guided inquiry cycle was the reflection that would take place as part of the post-task analysis in the Self-Reflection phase. This phase was composed of both an evaluation of personal competence as well as the extent to which the students felt satisfied with their work. This is precisely what Tschannen-Moran, Woolfolk Hoy, and Hoy (1998) called for in their cyclical model of self-efficacy. In order for something to become a "new source of efficacy information", the performance from the prior task must be analyzed. As was already stated, simply completing an action will not in and of itself lead to an increase in self-efficacy. It is the cognitive processing of that task which leads to greater self-efficacy.

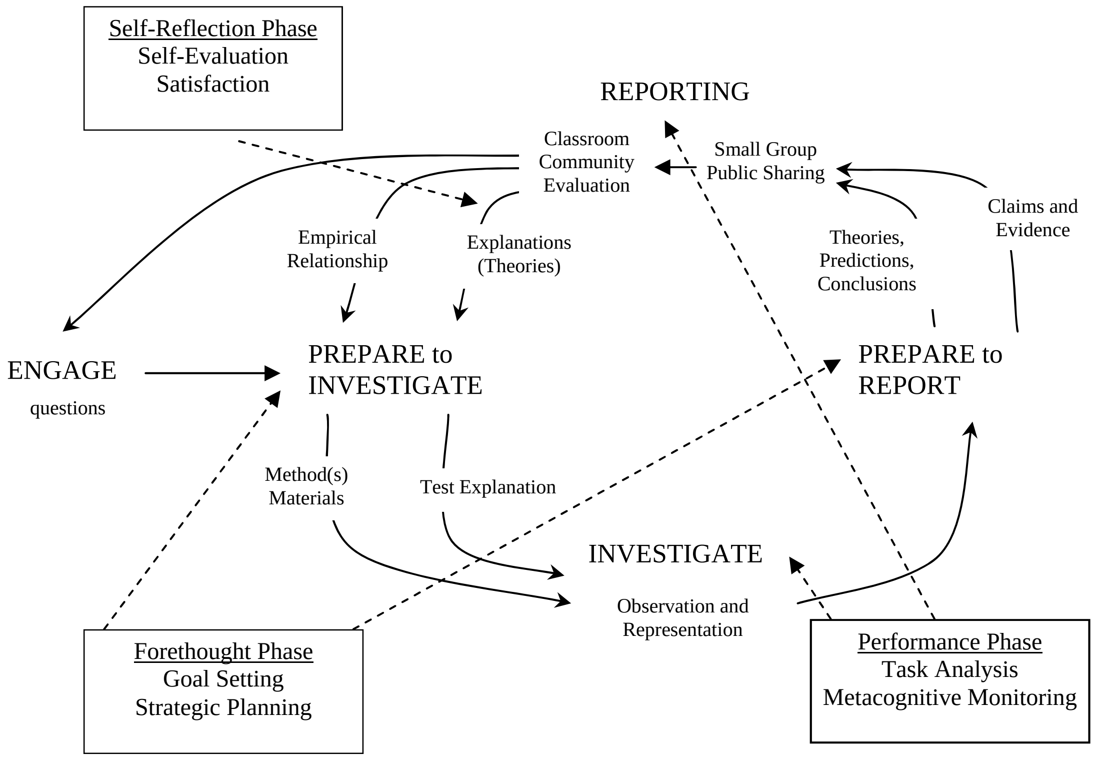


Figure 8: Guided-Inquiry and the Self-Regulation cycle adapted from work by Magnusson & Palinscar (2005) and Zimmerman & Campillo (2003).

As students become more adept at using the metacognitive strategies implicit in the science inquiry framework proposed for this study, they should see an increase in their abilities related to carrying out investigations. In addition, students should also become more proficient at creating subsequent reports. Providing time for post-task analysis will create a space for students to incorporate feedback from the prior task into their own self-efficacy judgments which should influence their performance in subsequent inquiry cycles.

Purpose of the Research

The aim of this study was to investigate ways to help high school students become more successful in the completion of science inquiry tasks. Specifically, it was hypothesized that self-regulation via metacognitive prompts and the use of sentence stems would result in greater student success on science inquiry tasks. In addition it was anticipated that increased student success would lead to a positive increase in student self-efficacy related to their abilities in science. As has already been shown, positive self-efficacy is significantly related to achievement (Bandura & Schunk, 1981; Schunk, 1996; Schunk & Swartz, 1993; Zimmerman & Bandura, 1994; Zimmerman, Bandura, & Martinez-Pons, 1992). It was anticipated that this study would have the following outcomes:

1. This study sought to advance our understanding of the role of prompted metacognitive reflections and the use of sentence stems as a writing strategy to assist high school students to be successful in the completion of science inquiry tasks.

2. This study was designed to result in a tool that could be used to measure student self-efficacy related specifically to chemistry content and also the skills necessary to carry out scientific inquiry.

Research Questions

1. What is the relationship between self-efficacy for science inquiry / chemistry content and performance on inquiry tasks?

2. Do metacognitive reflective prompts given prior to, throughout, and following the science inquiry processes improve student performance on inquiry tasks?

3. What impact does the use of sentence stems have on a student's ability to successfully communicate their ideas related to science inquiry tasks?

Significance of the Study

This study is the first that the author is aware of that blends the frameworks of guided inquiry (Magnusson & Palincsar, 2005), knowledge transformation (Bereiter & Scardamalia, 1987) and self regulation (Zimmerman & Cleary, 2006) to help students navigate the communication demands of an inquiry task. Both research (De La Paz & Graham, 2002; Klein, 2006; Yore & Treagust, 2006) and anecdotal experience suggest that students struggle with the task of reporting their scientific inquiry results and findings in a written format. At times, student work looks more like composition from a knowledge-telling as opposed to a knowledge-transforming mindset. While this may be due in part to limitations of the science inquiry prompts, Bereiter & Scardamalia, (1987) note that expert writers tend to take prompts that could easily be satisfied with a knowledge-telling format and instead tend to pursue a knowledge-transforming mode. As has already been stated, the goal of science inquiry in the classroom is that students use these inquiry experiences to develop new understandings and knowledge.

This study also responds to the recommendation that research in science education be classroom based (Bransford, Brown, & Cocking 2000; Magnusson & Palincsar, 1995). Bransford, Brown, & Cocking (2000) state that there is a need of "developing strong metacognitive strategies and learning to teach those strategies in a classroom environment" (2000, pg 21). This study will be classroom based, and as such, will be subject to the constraints and benefits that a typical science teacher would experience.

In addition to the classroom based nature of this study, this study also responds to the NRC recommendation that more work needs to be done specifically in relation to science education (Bransford, Brown, & Cocking, 2000). The NRC recommends that research be done to see if strategies that work for science education also work to improve instruction in other subject areas. This study actually asks the opposite question. What are some of the strategies used in other content areas (writing strategies in specific) and how can these strategies be incorporated into science education?

Finally, this study builds upon the large body of work that has been done related to self-efficacy (Bandura, 1995, 1997; Pajares, 2002; Pajares & Urdan, 2006). The relationship between student self-efficacy and academic success is well established (Bandura, 1995, 2006a; Pajares, 2006; Zimmerman 1995, 2006). In addition, that students with higher measures of self-efficacy often persist longer with tasks and set higher goals for themselves is also well established in the literature (Pajares, 2006; Schunk, 1996; Zimmerman 1995, 2006). While much work has been done in relation to the broad construct of self-efficacy, the literature review for this study yielded only three papers in which student self-efficacy in science was assessed in some measure (Shaw, 2004; Smist, 1993; Zusho & Pintrich, 2003). All three of these studies were conducted with college students and assessed their efficacy to learn physics, chemistry, or biology content in very general terms. One of these studies did assess student efficacy as it related to laboratory skills, but the questions were related to student efficacy to do things such as light a Bunsen burner or use a microscope (Smist, 1993). None of these instruments exhibited the specificity that is generally recommended for self-efficacy surveys (Bandura, 1997, 2006b) and none of them addressed student efficacy specifically related to the process and products of scientific inquiry. One of the outcomes of this study will be an instrument that measures student self-efficacy as is relates to high school chemistry and to the more general process of scientific inquiry.

Chapter 2: Review of the Literature

Introduction

This chapter looks at the current state of research as it relates self-regulation, self-efficacy, and science education. An initial search was done using educational research databases looking for articles using the terms: self-efficacy, self-regulation, and science education. This search yielded no results. The search was expanded by dropping the science education limiter. This yielded four articles of which only one was suitable. Combinations of the other search terms were used as well with similar results. Instead of using the database, I went through the reference sections for each chapter of nine different handbooks that had been published within the past twelve years (from 1995 – 2008) on self-efficacy, cognition, self-regulation, or science education. Once papers of interest were identified, the reference sections of these papers were also reviewed for research that might relate to this study. This search yielded a total of fifty-three articles. Seventeen of these articles were most relevant to this study and are reviewed in this chapter.

Self Regulation in Science Education

Guided Inquiry

The model of guided inquiry has been studied extensively within elementary and middle school contexts (Magnusson, 1996; Magnusson & Palincsar, 2005; Palinscar, Collins, Marano, Magnusson, 2000; Palincsar, Magnusson, Collins, &

Cutter, 2001). One of the first studies to employ the guided inquiry methodology was by Magnusson (1996). In this study Magnusson sought to use the guided inquiry model to help 4th grade students investigate the nature of sound. Using computers with microphones, students were directed to explore how the shapes of sound waves correlated with the sounds that were heard. The goal of the instruction was to allow students to discover the relationship between wavelength and pitch. This was followed with a computer based activity where students had the chance to explore the relationship between notes on a piano and the resulting sound waves that form. As a culminating activity, the students were tasked with building a musical instrument that could play one octave of a major scale.

The experimental and comparison groups in this study were small with only eight students each. The students in the control group were from a separate 5th grade class and they did not have access to the computer equipment. In addition, the students in the control group were given instruction in how music is used in human communication, but not in the scientific nature of sound. The data from the study was collected through interviews in order to determine the extent to which students had developed scientifically valid models of the relationship between the way a sound wave looks and how it sounds.

The results were mixed. Magnusson (1996) found that students in the experimental group were able to accurately describe the differences between wavelength and amplitude when interviewed during the study, but struggled to maintain this knowledge when interviewed in a more formal way at the conclusion of the study. The control group of students also seemed to struggle with the concept of amplitude and wavelength as being separate and independent ideas when they were interviewed at the end of the study. Interestingly, in these final interviews, the students from the experimental group were quite resistant to changing their ideas about how to define wavelength and amplitude. In contrast, students in the control group were much faster at recognizing the difference between amplitude and wavelength when encouraged to do so. Magnusson (1996) also noted that the students in the experimental group struggled with the ability to differentiate musical pitch and this significantly interfered with the students' ability to create the final product of a musical instrument.

Magnusson's (1996) conclusions from this study were primarily directed towards the use of computer technology as a teaching aid. While guided inquiry instruction was mentioned numerous times within the article, it was unclear how the guided inquiry process was used to direct and inform instruction. Since the model of guided inquiry was still in the process of development, this may have played a role. One point that Magnusson (1996) did make as a part of the conclusion, though, was an emphasis on the necessity of authentic experience to foster student learning.

While the Magnusson (1996) study was directed towards mainstream students, the next two articles (Palincsar, Collins, Marano, & Magnusson, 2000; Palincsar, Magnusson, Collins, & Cutter, 2001) looked specifically at the application of the guided inquiry model to instruction for students with learning disabilities in the mainstream classroom. These studies used an earlier conception of the guided inquiry model. Many of the components of the models are the same. Guided inquiry is a cyclical process where planning precedes both investigation and reporting. In addition, the outcome of an investigation provides the starting point for another cycle.

In the first of these two studies, Palincsar et al (2000) did a case study of one 4th grade learning-disabled student. The recursive nature of the guided inquiry cycle was shown to be effective in helping this student develop ideas specifically related to the concepts of sinking and floating using a Cartesian Diver. The learning logs that were kept over the course of the study demonstrated significant growth in this student's ability to make careful observation and even propose explanations for what was observed. The structured nature of the guided inquiry cycle appeared to have a positive impact on this student's learning.

One of the key phases of the guided inquiry model is that of reporting out findings. Interestingly, when it came to reporting results, this student's ideas were largely ignored by his group as they were making a poster to present their ideas. Even more ironic, this learning disabled student's ideas were more scientifically correct than what his group had proposed. Palincsar et al (2000) noted that were it not for the learning logs and frequent interviews by the researchers, the classroom teacher might not have recognized the learning gains that this learning disabled student had made. The point was not to fault the teacher, rather the point was to recognize that teachers face complex demands on their time and so it is essential to find more ways of determining what students really know.

The follow up study by Palincsar et al (2001) was conducted in four different classes, three of which were 4th grade classes and one was a 5th grade class. The four teachers had all been part of a community of elementary educators that had been meeting for about a year to discuss how to incorporate science inquiry into their lessons. All of the classes in the study contained a mix of high achieving students, low achieving students, and students that were on individualized education plans (IEP).

The study was conducted in two phases over two school years. The content for both years was either light or flotation. The first phase was conducted in year one of the study. In this phase, the teachers practiced developing and teaching science inquiry lessons using the guided inquiry model. While part of the focus in this first phase was on developing strategies for teaching, the authors also wanted to see how students with special needs responded to the opportunities and challenges presented by the guided inquiry model. Assessment of student learning was measured through pre- and post- assessments and also through data from observations and interviews. The data collected in the initial phase of the experiment was provided to the teachers so that the teachers and researchers could examine which teaching practices seemed most promising within a guided inquiry framework.

The second phase of the experiment took place over the second year of the experiment. In this second phase, the instruction focused on three practices that emerged from an analysis of the initial teaching phase: monitoring and facilitating student thinking, supporting print literacy, and improving students' abilities to work in groups. Even though the teachers in this second phase were the same as the teachers in the first phase, the students were new. The classes in this second phase contained an equivalent mix of students by ability types in comparison to the first phase. Pre-test measures compared using a Mann-Whitney U test (two-tailed) confirmed that the classes in phases one and two were statistically identical.

A focus on the three instructional practices as a part of the guided inquiry process in the second phase of the study produced positive results. In comparison to the initial phase, students engaged in the second phase of the research project had much greater learning gains between the pre- and post- test assessments. Whereas the learning gains from the first phase of the study only tended to be significant for the normal achieving students, significant learning gains were observed for all three student groups (normal achieving, low achieving, and those on an IEP) in the second phase of the study.

Another interesting finding that emerged from this study was that the interventions that were a part of this study placed a significant load on the practicing teacher. Palincsar et al (2001) commented that, "the interventions identified in this research as advanced teaching practices are, in many respects, simply part of exemplary teaching. Nevertheless, they place significant demands upon classroom teachers in terms of time, energy and cognitive space." (pg 30) While the teachers in the study agreed that the guided inquiry model of teaching was more demanding, they also noted that they felt more empowered in their work.

These studies indicated that a guided inquiry framework was beneficial to students of all ability levels. In addition, in all three of the studies, the guided inquiry model helped inform teaching that occurred over a period of weeks. These studies, though, only included students at the elementary level and so it will be informative to determine the extent to which a guided inquiry model of teaching is beneficial for high school aged students.

Metacognition, Reflection and Achievement

Similar to the guided inquiry framework, White and Frederiksen (1998) also proposed a cyclical teaching model in order to investigate the role of metacognition in science education. Their primary goal was to teach physics content to middle school students. They hypothesized that students' difficulty in subjects such as physics was not due to a lack of intellectual ability; rather students struggled because they did not know how to construct scientific conceptual models nor how to monitor their learning. To this end, White and Frederiksen (1998) proposed that science curricula be structured around a recursive loop that they called "The Inquiry Cycle" (see Figure 9). The "Inquiry Cycle" was intended to mimic the process of real science and thus aid students as they develop accurate conceptual models.

There were a number of ideas embedded within the inquiry cycle, although they are not explicitly stated in the model. Students going through this "Inquiry Cycle" were prompted to pause at designated times in order to reflect on what they had learned. To better model the process of science, the authors stated that they wanted this curriculum to be general enough so that students could explore physical phenomena without feeling constrained to follow specific experimental protocols. They also wanted the learning environment to approximate the social environment that often accompanies researchers in scientific settings.

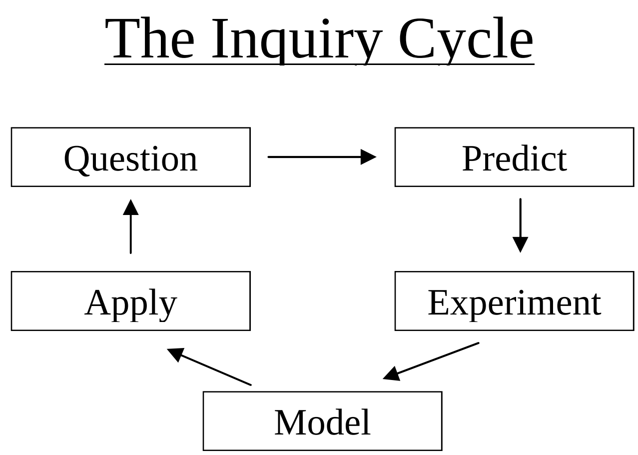


Figure 9: Stages of the Inquiry Cycle by White & Frederiksen (1998). (Note. From "Inquiry, modeling, and metacognition: making science accessible to all students", by B. White and J. Frederiksen, 1998, *Cognition and Instruction*, 16(1), p. 5. Copyright 1998 by Routledge. Reprinted with permission)

In order to determine the extent to which this inquiry cycle assisted students in the development of conceptual models, White and Frederiksen (1998) developed The ThinkerTools Inquiry Curriculum. This curriculum was composed of computer based simulations that allowed students to explore relationships between concepts such as force and motion. Students manipulated variables of their choosing and then attempted to elucidate the physical laws that governed the motion they observed. As students moved through the "Inquiry Cycle", computer prompts elicited feedback as students transitioned from stage to stage. Students were also prompted by the computer program to assess and critique other students' work. While initially, student work was fairly structured, the intent was that by the end of the curriculum, the students would be using these simulated environments to research questions of their own design

This study took place across twelve different classrooms in two urban schools with approximately 340 students from grades seven to nine. Every class contained a broad distribution of student abilities. The researchers noted, "This wide distribution is ideal for research purposes but is challenging for teachers" (White & Frederiksen, 1998, pg 29). Assessment of student abilities was based on prior achievement scores. Students were classified as either high achieving or low achieving in order to study whether academic ability affected the outcomes. While all of the students took part in the content part of The ThinkerTools Inquiry Curriculum, only half of the classes in the study were randomly selected to participate in the reflective assessment component of the curriculum.

A number of results from this study are especially intriguing. In looking at the impact of reflective assessment on written work, students originally categorized as high achieving only received marginal benefit from these reflective exercises in comparison to the control group (effect size of 0.4σ). For those students originally categorized as low achieving, though, their written reports were significantly better than their control group counterparts (effect size of 1.44σ) and actually approximated the work that was completed by the high achieving students. This seemed to imply that high achieving students are naturally reflective about their work and that this type of reflective self-assessment can be taught.

Since the students worked in groups, White and Frederiksen (1998) also looked at how group composition impacted the projects that students completed. They noted that students from heterogeneous groups outperformed their peers in homogenous groups only when the students were a part of the experimental group. When the authors looked at who benefitted the most, they found that the low achieving students benefitted the most when paired with a high achieving partner. Similar to the findings related to reflective assessment, the high achieving students did well whether paired with students of similar ability or with students of lower ability. In contrast, group composition made no difference for students in the control group. Again, these results highlighted the importance of reflection on one's work and also supported the idea that high achieving students were high achieving because they have developed the metacognitive ability to reflect on their own work.

One major limitation of this study as it relates to scientific inquiry in the classroom was the use of computers to create a simulated environment. White and Frederiksen (1998) noted that,

> The ThinkerTools software enables students to create experimental situations that are difficult or impossible to create in the real world…This is more straightforward than the corresponding real-world inquiry task. After all, objects in the real world are not driven by laws, rather laws simply characterize their behavior" (pg 14 & 15).

The limitation of this simulated world is that students are given "clean" sets of data with which to work. These simulated data lack the often chaotic and "noisy" data that is typically the real data of science. How will these students fare when given the freedom to explore the real, physical word? Error is a natural and inevitable part of the scientific enterprise (Rutherford & Ahlgren, 1989). Reflective-assessment was effective in helping to promote students' understanding of inquiry in the fairly controlled world of simulated physical events. The study in this proposal sought to build on this base by looking at how reflection can be used to enhance student understanding of scientific inquiry conducted in the real world.

The authors also noted a caution concerning student reflection in their conclusion. Too much student reflection may be counter-productive. One student wrote, "Too much self-assessment!...Don't you think that's a bit much?" (White & Frederiksen, 1998, pg 86). While the White & Frederiksen (1998) study supports the conclusion that reflection helps raise student achievement, reflection is not an end in and of itself. Reflection is a means to understanding the inquiry process and is only helpful to the extent that it facilitates student learning.

Davis (2003) provided an alternative perspective on the nature of prompts to promote student metacognition. She specifically looked at the nature of the prompts that were given to students in an attempt to promote metacognitive reflection as they worked their way through a science inquiry project. In her study, students were given two types of prompts. One type of prompt was called a "directed prompt". The "directed prompt" asked the students to consider a specific idea as it related to their particular stage within their science project. The opposite of the "directed prompt" was the "generic prompt". The "generic prompt" occurred in the experimental group at the same places as the “directed prompt” occurred for the comparison group. Instead of asking students to reflect about a specific idea, though, the “generic prompt” merely asked students to respond to something such as “Right now, I am thinking…”

The Davis (2003) study was also conducted with middle school students similar to the White and Frederiksen (1998) study. Davis (2003) stated that about 180 students from one school were the subjects studied and that the school was “somewhat diverse” (pg. 104). This particular study took place at the end of unit on heat flow and energy conversion. The students were asked to use their knowledge from the preceding unit to analyze a fabricated news article that was meant to read like a science tabloid. The students were to read the article and provide a critique of the science content that the article presented. They were then asked to write a letter back to the editor, identify scientific errors they encountered, and then explain evidence that they used to back up their claims. The students worked through the project using a computer program called Computer as Learning Partner (CLP). Although it was unclear exactly how the students were prompted, the students were prompted to reflect (directed or generic) a total of eleven times. Davis (2003) then compared the letters that the students wrote to their responses on the eleven reflection prompts.

There were two results from the Davis (2003) study that were particularly relevant to this proposal. Regardless of the experimental condition, students whose reflections were judged as poor or unproductive were generally less successful on their final project. Poor reflections were those prompts to which students either did not respond or responded with something akin to “we are not having any problems”. This result was similar to what has been found in other studies related to reflection and achievement (Bransford, Brown, & Cocking, 2000; Flick & Tomlinson, 2006; White & Frederiksen, 1998).

The relationship between prompting type and quality of reflection was the second result of interest. Davis (2003) stated that “generic responses appear to promote [more] productive reflection” (pg 116) and attributes this difference to “cognitive economics”. She stated that students tended to move towards work that required the least amount of cognitive effort and thus avoided work that carried large cognitive demand. The shortcoming, then, of directed prompts was that they allowed students to be mentally “lazy” and get away with reflection responses such as “we are not having any problems”. Generic prompts, in contrast, did not allow students the cognitive luxury of a “no problems” type of responses because they asked students to reflect widely on what they were currently thinking as opposed to narrowly focusing on a specific, predetermined (directed) idea.

While cognitive economics may play a role, directed responses may also be less effective due to the nature of how they are interpreted. Students often seemed to be concerned with the “correct answer” as opposed to developing the correct understanding. It may be that students interpreted directed prompts as prompts that were looking for the “correct answer” as opposed to an invitation to reflect on their current understanding and thinking. The study in this proposal will investigate the extent to which prompting type influences the quality of student inquiry reports.

In addition, the extent to which the type of prompt is beneficial to students may vary based on prior achievement levels. It could be argued that students who are already high achievers may find generic prompts to be more useful because they have already developed the necessary skills to reflect productively. In contrast, lower achieving students may benefit from the scaffolding that directed prompts may provide. One focus of this study proposal will be to examine the extent to which different types of metacognitive prompting aided high achieving and low achieving students.

The other limitation of the Davis (2003) study, similar to the White & Frederiksen (1998) study, is that students responded to a hypothetical situation. How would students responses differ if presented with a real life situation or with real life data? Granted, part of the use of hypothetical or simulated data may be due to the fact that both studies were conducted at the middle school level, but, again, the focus of this dissertation proposal was on helping provide students with the necessary tools to understand, interpret, and communicate ideas in relation to data they have collected in a lab setting.

Prompting and Prior Student Achievement

One study that looked at this relationship between prior student achievement and the use of reflective prompts was conducted by Coleman (1998) with a group of 48 fourth and fifth grade students spread over two schools. The purpose of the study was to look at the ways in which collaborative explanation led to gains in conceptual understanding of scientific concepts. In this case, the students were studying the process of photosynthesis. The students in the study were divided into three groups. Through pre-testing, one group of students was identified as "high achieving". The "high achieving" group did not receive any special training or instruction. All of the remaining students were identified as "average achieving". The "average achieving" group was then split in two. One "average" group acted as a control while the other "average" group received extra instruction in prompts to facilitate group discussion.

While the Coleman (1998) study looked at scientific inquiry through the lens of collaborative discussion as opposed to through the lens of written response, this study is of interest because of its foundation in the writing process. Coleman (1998) stated that her collaboration prompts were adapted from prior work by Scardmalia & Bereiter (1985) and Scardmalia, Bereiter & Steinbach (1984) that were originally developed as interventions to help young writers. In these collaboration groups, students were asked to:

1. Evaluate their own thinking and understanding.
2. Justify their responses based on prior experience or information learned in the class.
3. Compare and contrast ideas in relation to scientific thinking.

At the conclusion of the study, students were assessed in three ways. First, groups were rated on the scientific accuracy of their discussions. These discussions were in response to questions that had been provided by the author (eg. What does soil provide to a plant if all plants need for photosynthesis is CO_2, water and sunlight?) .

Second, student learning was assessed through a knowledge post-test. Finally, students were assessed via a concept map activity where they drew relationships between key ideas of photosynthesis.

Similar to White & Frederiksen (1998), Coleman (1998) found that the final products produced by the students that had participated in the experimental condition were in many ways undistinguishable from the final products of the "high achieving" students. These products were significantly better than the final products produced by the control group. Again, these results indicated that higher achieving students already possess the metacognitive skills to monitor and assess their own learning.

One shortcoming of the study was that none of the "higher achieving" students actually took part in any of the experimental conditions of the study. Coleman (1998) noted this shortcoming and states, "It would have been interesting to know whether the students in the HIL ("higher achieving") condition would have benefited from the prompting procedure in the same way as the AI ("average achieving") students" (pg 417). One of the aims of this study proposal was to investigate if high achieving students benefit (if at all) from metacognitive prompts related to science inquiry projects.

Learning Logs and Metacognitive Prompting

The final paper reviewed in this section on self-regulation in science education was a study by Flick and Tomlinson (2006) that took place in a fourth grade classroom. Building on work done Brown & Campion (1998), Flick and Tomlinson

(2006) adapted cognitive strategies originally developed to assist struggling readers. These cognitive strategies were tailored to issues specific to scientific inquiry. Flick and Tomlinson (2006) stated their goal was to “evaluate how the teaching of cognitive strategies improved student performance on science inquiry assessment tasks” (pg 183). The strategies specifically employed in this study were:

1. Setting a purpose: This strategy involved helping students set daily learning goals and developing activity purpose statements in order to help students focus on key ideas / learning goals.
2. Using prior knowledge: In this strategy, students were asked to make graphic organizers such as concepts maps, Venn diagrams, “webs” of knowledge, and K-W-L charts. These explicit prompts came prior to learning activities in order to activate student prior knowledge related to the task at hand.
3. Looking for patterns: While pattern identification is most often thought of as a math / science activity, the strategy of pattern identification is also extensively used in reading comprehension.
4. Metacognition: Students were given instruction on how to monitor their own learning through the use of graphic organizers, questioning routines, and Learning Logs.

While the focus of these strategies was on facilitation of the process of scientific inquiry, Flick and Tomlinson (2006) pointed out that these cognitive strategies for scientific inquiry were closely related to the cognitive strategies employed in reading comprehension. For example, both science inquiry tasks and reading comprehension activities often ask students to draw inferences based on the given data. In science, this data comes from experimentation whereas in reading, this data comes from the text itself. The related nature of these science inquiry and reading comprehension activities was highlighted to the students in this study.

The study took place over the course of an entire school year. Also, even though the focus of this study was on cognitive strategies as they related to scientific inquiry, these four strategies were adapted and employed in other content areas as well within the classroom. Examples of other content areas where these strategies where implemented were reading, math problem solving, and essay writing. Changes in student cognition were measured by the following:

1. A tool developed by Flick and Tomlinson (2006) called the Cognitive Strategies Inventory (CSI).
2. Student Learning Logs.
3. Graphic organizers that students produced for their various learning activities.
4. Student interviews.

A number of Flick and Tomlinson's (2006) conclusions are relevant to this study. First, from student interviews, they noted that their fourth grade students tended to focus on simply completing a task as opposed to understanding what the task was all about. Research in self-regulation has termed this difference "process" versus "product" orientation (Zimmerman & Kitsantas 1997, 1999). The impact of goal orientations on self-regulation and self-efficacy will be covered in more detail in a later section of this literature review.

Also of interest, Flick and Tomlinson (2006) noted that as students progressed throughout the year, they typically made greater use of the cognitive strategies. For example, two science projects were undertaken during the year. Students indicated that they used the cognitive strategies to a greater extent on the second project in comparison to the first. There was great variability in the number of students that responded to each prompt after the first (n = 1 to 9) versus the second science project (n = 7 to 12) and so these results are tentative at best. Finally, student learning logs showed an increased ability to reflect on their own learning as the year progressed. At the beginning of the year, students' comments were very general and lacked specificity. At the conclusion of the year, students were much better at articulating what they were thinking and areas in which they were struggling. The authors did not state whether they felt that these improvements in student metacognition were due to changes in student cognition or simply the result of students being more able to articulate what they had been thinking all along.

The significant limitation of this study was that final student work was not analyzed. It would have been interesting to see how the products of student inquiry were related to the quality of student reflection about their own learning. Work from the previous two studies that were reviewed in this section (Coleman, 1998; Davis 2003) indicated that the quality of student reflection was correlated to the quality of student work produced. Again, the study in this proposal sought to expand on this work

Section Summary

The work that has been reviewed in this section strongly supported the idea that self-reflection aids in the successful completion of science inquiry tasks. While only two of these studies provided evidence of how student work was evaluated, (Davis 2003 ;White & Frederiksen, 1998), the limitation of both of these studies was that student inquiry work was collected via computer simulations or hypothetical science events. The aim of the research in this study was to investigate how self-reflection aided student completion of science inquiry tasks with real world data.

In addition, all seven of the studies reviewed in this section (Coleman, 1998; Davis, 2003; Flick & Tomlinson, 2006; Magnusson, 1996; Palinscar, Collins, Marano, Magnusson, 2000; Palincsar, Magnusson, Collins, & Cutter, 2001; White & Frederiksen, 1998) were conducted with students that were either in elementary or middle school. In contrast, the aim of this study was to investigate how high school students were able to use self-reflection as a means of monitoring their own learning and then looked to see if they were able use this feedback to make the appropriate adjustments. The research reviewed in this section indicated that higher achieving students already had the skills necessary to be metacognitive and to self-regulate their learning. As such, it was expected that higher achieving students would see little difference in the quality of their work as a result of their prompting in self-reflection.

In contrast, it was expected that lower achieving students would see a significant difference in their abilities to conduct inquiry investigations and communicate their ideas as a result of prompts for reflective thought.

Self-Regulation in Writing

Section Overview

While there are a number of skills that are part of a scientific inquiry process, one fundamental skill that students must learn is the ability to communicate their ideas and findings through writing. Having high self-efficacy related to the ability carry out scientific inquiry or having excellent self-regulatory strategies are only of so much help, though, if a student lacks the skills necessary to communicate their ideas effectively. While it is true that writing is not the only form of communication available to students, the fact that writing is such a common form of communication merited a look at studies which addressed self-regulation of the writing process. In addition, it may be that students understand how to carry out a scientific investigation, yet lack the skills to adequately communicate their ideas. A number of studies (Page-Voth & Graham, 1999; Scardamalia, Bereiter, & Steinbach, 1984; Zimmerman & Risemberg, 1997) have observed that novice writers tend to view writing as merely knowledge-telling. In this knowledge-telling process, novice writers tended to write down ideas as they recalled them from memory with little thought given to what came before or what will come after. Also, little effort was given to the task of revision so initial ideas become a final product. From personal experience, this description of writing as a “knowledge-telling” event described many of the science inquiry projects that I received from my students.

Another perspective on the problems that students face when it comes to writing has been proposed by Scardamalia, Bereiter, & Steinbach (1984). They viewed the task of writing, itself, as a form of problem solving. If this is true, then students who are working to complete science inquiry tasks are not only faced with the problem solving aspects inherent specifically to developing tools and methods necessary to carry out the investigation, but they are also subject to the problem solving aspects related specifically to the process of writing. Therefore it is important to not only understand the types of metacognitive prompts which aid in the experimentation phase of a scientific inquiry, it is also necessary to understand the metacognitive aspects of expert writers so that a student’s writing ability does not hinder their ability to clearly communicate their ideas. As such, this section of the literature review covers studies which investigated self-regulation as it related specifically to the task of writing.

Writing as a Reflective Process

The earliest study that this section of the literature review covered was work done by Scardamalia, Bereiter, and Steinbach (1984). Their goal was to see if sixth grade students were able to “sustain reflective processes in composition independently” (pg. 174). This sustained reflective process is exemplified by expert writers as they do such things as formulate goals, anticipate difficulties, reconcile divergent ideas, etc. Novice writers, in contrast, rarely demonstrated any of this reflective thinking in their own writing. Scardamalia, Bereiter, and Steinbach (1984), noted that expert writers' reflective thinking resembles that of an internal dialogue that looks very similar to a soliloquy.

This soliloquy is the result of interaction that takes place between two different cognitive "spaces". Scardamalia et al (1984) referred to one of these "spaces" as the "content space". This space is composed primarily of beliefs and ideas. It is this space in which one's opinions are determined, moral decisions are made, formulas and explanations are elucidated, etc. The second "space" is called the "rhetorical space". This space is specifically oriented toward producing text. These two spaces are interdependent where the ideas in one space become problems for the other and vice versa. For example, once one generates an idea in the "content space", the issue, then, is how to accurately represent this idea as text. Or when one is composing text and realizes that the limitations of text make an idea unclear or incomprehensible, this idea is then sent from the "rhetorical space" back to the "content space" for processing. They stated, "Our contention is that this interaction between the two problem spaces constitutes the essence of reflection in writing" (pg. 176). In order to effectively promote student metacognition in writing, Scardamalia et al (1984) contended that teachers must think about ways to activate a student's "conversation" between these two spaces.

As was mentioned before, the Scardamalia et al (1984) study took place in two sixth grade, public school classrooms. One of the classes was the control and the other was the experimental group. The instruction was given twice a week for a period of fifteen weeks. Part of this time was spent on how to write an opinion essay while the remainder of the time was spent on how to write a factual exposition. All of the students completed a pre- and post-test essay of each type (opinion and factual exposition). Six students from each class, though, were randomly selected for comparison between the two groups. Instruction with the experimental group consisted of three components:

1. Modeling Thought: This component consisted of the both the teacher and students actively modeling the process of reflective thinking. Following these modeling activities were discussions that highlighted the thinking strategies that were demonstrated.
2. Direct Strategy Instruction: It was noted that students often have difficulty trying to reconcile opposing viewpoints. This strategy directed students to "rise above" the conflict and attempt to create a position that acknowledged what was valid in each perspective.
3. Procedural Facilitation: This instructional component presented students with sentence stems to be used in their papers. Of interest to this study was the list of sentence stems that were used to help students construct their factual exposition papers. Examples of these sentence stems were prompts such as "A consequence of this is…", "I could describe this in more detail by adding…", "This results in…", or "My main point is…". These prompts were also divided into five

> categories: New idea, Improve, Elaborate, Goals, and Putting It Together. A modified version of these prompts was used in this study.

As was expected, when the groups were compared in a think aloud protocol, the subjects in the experimental group were significantly more reflective as it related to their writing. Also, when student work was compared, those who were in the experimental group produced work that showed much more personal involvement through examples and illustrations than students in the control group. In this sense, these results agreed with the work done by Coleman (1998) and Davis (2003) which pointed to the quality of reflection as related to the quality of student work. Scardamalia et al (1984) cautioned that this study was only intended to address whether elementary students could be engaged in reflective processing. While these activities demonstrated that sixth grade students were capable of some elements of reflective processing, this reflection was mostly self-focused and was still a far cry from the types of reflective processing such as elaboration and reshaping that are associated with expert writers. The study in this dissertation proposal sought to provide insight into high school students' abilities to reflectively process their own writing.

Goal Setting and Writing Performance

While reflective processing is one strategy that expert writers employ, a number of researchers have highlighted the goal oriented nature of expert writers (Page-Voth & Graham, 1999; Schunk & Zimmerman, 2007; Zimmerman & Kitsantas, 1999; Zimmerman & Risemberg 1997). Zimmerman and Risemberg (1997) noted that the British novelist, Anthony Trollope would set daily writing goals that he would dutifully record in his personal diary. They also related how Hemmingway had daily output goals and he would not allow himself time off unless he was at least a day ahead of his planned goal. In terms of self-efficacy, Bandura (1997) would describe these day by day assessments of progress as proximal goals. The daily, proximal type, goals were only half of the picture. These authors also had a goal of completing a piece of literature. This long term goal would be termed a distal goal because there was a significant period of time that would need to take place between starting and then eventually finishing the project. In order to accomplish these distal goals, proximal goals were put in place that provided more immediate feedback as to whether these authors were, or were not, progressing toward their desired endpoint.

In contrast to the patterns of expert writers, Flick & Tomlinson (2006) noted that their students tended to focus on project completion. By their statement, they implied that their students had adopted distal goals but failed to provide the proximal goals that would facilitate their attainment of these distal goals. Instead of using the terms proximal and distal, Zimmerman & Kitsantas (1997, 1999) differentiated these as process versus outcome goals. Process goals focus on developing an understanding of heuristics or methodologies that can be used to complete a task whereas outcome goals are only focused on producing a product.

In one study, Zimmerman and Kitsantas (1999) looked at how having students shift from process to outcome goals affected their ability to acquire writing revision skills. In contrast to the other studies that have been reviewed so far, this study was done with 84 high school students from grades 9 – 11. Interestingly, this study was done at an all-girls parochial school. Also of note, while students in this study scored between the 5^{th} and 99^{th} percentile on a standardized test of English, the mean combined score of these students would have placed them in the 75^{th} percentile. The authors noted that this indicated that they were working with a student population that would have been considered above average. While caution must be exercised in trying to broadly apply these ideas to all students, the beneficial nature of process goals is predicted by social cognitive theory (Bandura, 1997).

Students in this study were assigned a writing revision task where they were given a number of very simple sentences on a particular topic. The students were then asked to combine all the individual sentences into one elaborate sentence that still captured the key ideas from each of the simple sentences. In order to look at the effects of outcome versus process goals, the students were split into seven groups. Because the Zimmerman and Kitsantas (1999) were also interested in how reflection aided student comprehension, this variable was also included. The seven groups were differentiated in the following way:

1. Control group: These students were given an initial practice session and the final assessment.
2. Process Goal group: These students were asked to adopt a learning goal related to effectively implementing the steps that were provided in the initial practice session. There were actually two Process Goal groups.

Only one of these groups was asked to reflect on their learning in this study.

3. Outcome Goal group: These students were asked to adopt a goal related to the successful completion of the writing revision task. Again there were two Outcome Goal groups, only one of which was asked to reflect on their leaning throughout the study
4. Process Goal to Outcome Goal group: These students were originally given the same goal as the Process Goal group. Half-way through the study, though, once the authors felt that the students in this group had sufficiently internalized these heuristics, the students were asked to switch their goal to match that of the Outcome Goal group. Similar to the other conditions, there were two student groups that were assigned to this Goal Switching condition, only one of which was asked to reflect on their learning.

When analyzing the results, Zimmerman and Kitsantas (1999) found that the most effective condition for student learning was found for students in the Goal Switching group. Student in this group produced better results, were more efficacious about their ability to complete these writing revision tasks, and indicated a greater intrinsic interest in this writing task when compared to the other groups. In addition, the students in the Goal Switching group were more apt to attribute their lack of success to deficient strategy use as opposed to attributing lack of success to uncontrollable factors such as lack of ability. While not as pronounced, the same beneficial effects were noted for the students who had been part of the Process Goal group. The authors noted that there was not much of a difference between the students in the Outcome Goal group and the control group. The authors stated that these results mirrored a prior study (Zimmerman & Kitsantas, 1997) where students in the control group and Outcome Goal group were virtually indistinguishable. They assumed that this was due to human nature to set goals and that these students had automatically set outcome goals for themselves. It is also worth noting that within each of these Goal conditions, the students who were involved in reflection outperformed their non-reflecting counterparts in every measure.

While there were some significant limitations to the Zimmerman & Kitsantas (1999) study especially in relation to gender, and the limitation of the actual writing that students were asked to do, there were a number of key ideas that their work highlighted. First, as has already been noted, individual goals were instrumental for successful writing. Similar to that of expert writers, process goals or proximal goals seemed to provide the greatest benefit. In addition, as has already been highlighted in a number of other studies, self-reflection tended to lead to better final products.

One of the difficulties with attempting to apply these results (Zimmerman & Kitsantas, 1999) centers on the question of competence. Zimmerman & Kitsantas found the greatest effect for students who initially adopted a process goal and then changed their focus to an outcome goal. Students were instructed to change their goal focus, apparently, when the authors felt that the students had become competent in the strategies that they were teaching. While this is fine for processes that are discrete and well defined such as combining short sentences into longer ones, it is much more problematic for tasks which in which the discrete steps are more ambiguous such as writing an essay. How is competence determined in these more complicated scenarios and how do we know when competence has been achieved in order to help students transition from a process goal orientation to that of an outcome focus? There is no definitive answer at this point, but this study sought to build on the work that has already been done in this area.

The results of the Zimmerman & Kitsantas (1999) study related to the importance of goal orientations are supported by work done with learning disabled (LD) students by Page-Voth & Graham (1999). In the study, Page-Voth & Graham (1999) were interested in ways to help LD students become more effective at essay writing. They focused their efforts specifically around two writing strategies: increasing the number of supporting reasons, and increasing the number of refutations to counterarguments. The study included a total of thirty 7th and 8th grade students who all attended schools within the same district. In their study, the students wrote a total of four essays. The initial essay was used to establish a baseline for each student whereas the subsequent three essays were used to assess student progress. Students were assigned to one of three conditions. In one of the experimental conditions, the students were asked to adopt a goal related to the two writing strategies and then were also provided with specific strategy instruction. Another experimental condition only asked that students adopt a writing goal related to the two writing strategies. The final condition was the control group where the students simply completed the essays as prompted.

When student essays were examined, Page-Voth & Graham (1999) found that students from the goal setting groups wrote essays that contained more supporting reasons for their ideas, had more refutations of counterarguments, were longer in length, and were independently judged to be of higher quality when compared to the essays produced by students in the control group. Surprisingly, there was not any significant difference between the essays produced by the students in the two different goal setting conditions. While the authors posed a number of explanations as to why this may have been the case, these results were clearly in line with other studies which supported the importance of goal planning in effective writing.

One other result of note from the Page-Voth & Graham (1999) study related to measures of self-efficacy. In their study, the authors also measured student efficacy for writing essays. Regardless of experimental condition, the authors noted that for the self-efficacy measure, “there was no statistically significant difference for group membership, the repeated measure, or the interaction between the two” (pg 236). Student estimates of their efficacy as it related to their ability to successfully write essays was generally neutral. Part of this may be due to the self-efficacy instrument itself. The students were asked to respond on a 5-point Likert-type scale. Bandura (1997, 2006b) recommended at least a 10-point Likert-type scale in order to provide enough specificity to measure changes over time. In addition, there was a question about the extent to which these students readily understood what was being asked of them in terms of self-efficacy. These students were labeled as learning disabled, and this may have affected their ability to accurately self assess their beliefs about their own abilities. Also, the study did not provide any statistical measure that assessed the internal consistency of the scale. This does not mean that their instrument did not measure what the authors stated that it measured, it only means that an extra level of caution must be taken with the results. Since the results were ambiguous, this again highlighted the necessity of ensuring that measurement tools are valid and accurate.

Section Summary

In reviewing the studies presented here, it is interesting to note how similar the task of writing is to the task of scientific inquiry. Similar to what Flick and Tomlinson (2006) proposed about the relation between the cognitive strategies necessary to do science inquiry and the cognitive strategies that aid in reading comprehension, it appears that the same sorts of parallels apply to the process of writing as well. Science inquiry is goal-directed (Flick & Tomlinson, 2006) as is the process of writing (Page-Voth & Graham, 1999; Schunk & Zimmerman, 2007; Zimmerman & Kitsantas, 1999; Zimmerman & Risemberg 1997). Scientific inquiry involves pre-planning in order to carry out the task (Flick & Tomlinson, 2006), and pre-planning was also one of the traits that appeared to differentiate novice and expert writers (Scardamalia, Bereiter, & Steinbach, 1984; Zimmerman & Risemberg 1997). Also, in the same way that a reflective disposition towards writing led to a better product (Scardamalia, Bereiter, &

Steinbach, 1984), the same appeared to be true in relation to the process of scientific inquiry (Coleman, 1998; Davis 2003).

Self-Regulation and Self-Efficacy

Section Overview

While the literature review up to this point has focused primarily on the topic of self-regulation, it is important to note that self-regulation often operates in close connection with self-efficacy (Bandura, 1997; Gaskill & Woolfolk Hoy, 2002; Pajares, 2008; Pajares & Urdan, 2006; Schunk, 2001; Schunk & Ertmer, 2000; Zimmerman, 1995). Self-efficacy is concerned with an individual's judgment about their ability to carry out a specific task whereas self-regulation is the development of an individual's metacognition to actually carry out these specific tasks. While numerous studies have pointed to the predictive value of self-efficacy in relation to future student achievement (Bandura, 1995, 2006a; Pajares, 2006; Zimmerman 1995, 2006), Bandura (1997) also noted the importance of self-regulation as it related to self-efficacy. He stated,

> Neither cognitive processing skills nor metacognitive skills will accomplish much if students cannot get themselves to do academic assignments. A strong sense of efficacy to regulate one's motivation and instructional activities undergirds belief in one's academic efficacy aspirations (pg 231).

In other words, if a student feels highly efficacious about a specific learning task, yet lacks the metacognitive skill to effectively complete this task, then the student's chances of success are quite small. The opposite scenario, though, is also true. If a student possesses the metacognitive skills necessary to complete a task, yet feels inefficacious about using these skills, the end result will still be the same. The student's chances of successfully completing the task will be quite small. In this sense there is a dependent and reciprocal relationship between an individual's self-regulatory skills and their sense of self-efficacy. The papers in this section of the literature review looked at the changeable nature of self-efficacy and the interdependent nature of self-efficacy and self-regulation.

Goal Orientation and Self-Efficacy

While process goal orientation has been shown to have a positive impact on student self-regulation of writing (Page-Voth & Graham, 1999; Zimmerman & Kitsantas, 1999), process goal orientation has also been shown to play a significant role in self-efficacy development. Bandura and Schunk (1981) looked at the relationship between goal orientation and student self-efficacy, intrinsic interest, and subsequent performance on math subtraction problems. In their study, Bandura & Schunk (1981) focused on the differential effects of students who adopted proximal versus distal goals. As has already been stated, proximal goals have a number of advantages over distal goals in relation to the development of self-efficacy and individual performance. First, since they are more temporal in nature, they provide more immediate feedback as to whether or not a certain course of action is actually successful. Second, individuals often find success satisfying and motivating. The immediate feedback often provided by one's assessment of proximal goals acts both to create incentive to continue and to provide a basis for continually reassessing one's self-efficacy to successfully continue.

In order to look at the relationship between proximal goal setting and subsequent changes in interest and self-efficacy, Bandura & Schunk (1981) worked with 40 elementary students, ages 7 – 10, which were from six different elementary schools. These students were chosen because they all demonstrated significant deficits in arithmetic skills accompanied by low interest in math. Subsequent pre-testing confirmed these observations. These students were assigned to one of four experimental conditions: proximal goals, distal goals, no goals, or no treatment. Students in the both of the goal conditions were given suggestions as to what kind of goals they should adopt. The authors stated that these goal orientations were merely offered as suggestions in order to allow the students to feel that they had some ownership in their own goal creation.

The treatment for the proximal, distal, and no goal conditions was a sequence of seven lessons that were given as self-directed learning experiences that covered various ideas related to subtraction. Each lesson took approximately 30 minutes to complete. To go through these lessons, students were released from their regular classrooms at staggered times and worked in isolation from other students. Since each lesson consisted of six pages, the proximal goal students were urged to focus on making it through the six pages each session whereas the distal goal students were urged to focus on completing the entire seven lessons. Students in the no goal group received the same seven lessons, but without any recommendations regarding goals.

The no treatment group was only given the assessments that the students in the other groups received.

Bandura & Schunk (1981) found that students in the proximal goal condition exhibited greater self-efficacy related to their ability to do subtraction and greater intrinsic interest in these types of math problems than students in the other experimental conditions. In addition, the students in the proximal goal group had significantly higher scores on the post-assessment and their self-efficacy assessments were also better predictors of success on the post-assessment as compared to the predictive effects of student self-efficacy for those from the other groups. In contrast, students in the distal and no goal conditions were in many ways indistinguishable from each other in terms of self-efficacy judgments, intrinsic interest in subtraction problems, and the predictive value of their self-efficacy judgments. While the distal and no goal students performed significantly better than the students who received no treatment at all, their performance was significantly below that of their peers who had developed proximal goals.

The work of Bandura & Schunk (1981) again highlighted the value of proximal goals in the development of learning targets. These results also provided support for the idea that people naturally develop distal goals whether directed to do so or not. Also worth noting is that with increased self-efficacy comes increased intrinsic value. The authors stated that, “Young children are not innately interested in singing operatic arias, playing tubas…or propelling heavy shot-put balls through the air. However, through favorable continued involvement, almost any activity can become imbued with consuming significance" (Bandura & Schunk, 1981, pg 587). In relation to the study in this proposal, many students seem to avoid and even dread scientific inquiry experiences. It will be interesting to note if student intrinsic interest in science inquiry changes as they develop the self-regulatory strategies and self-efficacy to be successful. Also, whereas this study was conducted in a laboratory type environment, the study that is presented in this proposal took place in a classroom as an integrated part of the daily instructional activities.

Two other follow up studies done by Schunk & Swartz (1993) and Schunk (1996) also highlighted the central role that goals play in the development of self-efficacy. Similar to the Bandura & Schunk (1981) study, these studies were not done in a classroom setting but rather took place in a laboratory type environment. The Schunk & Swartz (1993) was composed of two parts. The first part of the study investigated the role that feedback played in the development of student self-efficacy as it related to student goals. Similar to Zimmerman & Kitsantas (1997, 1999), Schunk & Swartz (1993) had 60 fifth-grade students adopt a process goal (learn to use the correct strategy), a product goal (strive to write a complete paragraph), or a general goal (do your best). Only half of the students in the process goal condition received any form of feedback about their progress as they were learning. The students received instruction over 20 days related to writing specific types of paragraphs. Tests measuring writing skill and self-efficacy for writing were given before and after the 20 day treatment.

Schunk & Swartz (1993) found that self-efficacy for writing was the greatest for students in the process goal condition regardless of whether the students received feedback or not. In addition, the students in both process goal conditions felt more capable of learning to write than the students in the other conditions. Self-efficacy for writing and the quality of the writing that the students produced were also highly and significantly correlated. When the data was assessed, self-efficacy for writing accounted for 69% of the variation in student writing performance. That self-efficacy was greatest for those students in the both progress goal conditions indicated the effectiveness of progress type goals in promoting and developing student efficacy.

The second part of the Schunk & Swartz (1993) study was almost identical except two measures were added at the conclusion of the experiment to see the extent to which students were able to apply the writing concepts they learned to new and novel settings. The authors also wanted to see if these learning gains were actually maintained over a period of time. To test this, the students were given a final post-test six weeks after the instruction took place. It is also worth noting that this second study took place with 40 fourth-grade students. This grade level was chosen intentionally in order to assess how well the previous findings were transferable to younger grades.

As before, students in both process goal treatments developed greater senses of self-efficacy towards writing and displayed greater skill in their writing products. These students were also able to retain the strategies over a longer period of time and indicated that they used these strategies more often. While Schunk & Swartz (1993) noted that the students in the process goal with feedback condition scored consistently higher on all post-test measures than students in any other condition, they noted that there was no significant difference between the students in the different process goal conditions. These results confirmed the benefit of a process goal orientation to learning, but they remained ambiguous about the extent to which feedback is necessary and useful to enhance student self-efficacy and thereby increase student achievement.

The Schunk & Swartz (1993) study on writing was followed up by Schunk (1996) where he looked at the relationship between goal orientation and achievement in math. Goals in the Schunk (1996) study were termed learning goals and product goals. Learning goals were analogous to process goals. As with Schunk & Swartz (1993), the Schunk (1996) study was composed of two related experiments.

The first experiment was conducted with 44 fourth-grade students from two classes in one elementary school. As before, these students were taught in a laboratory type setting that was removed from their regular classroom. These students received seven lessons on fractions. The students were divided up into four groups: learning goals versus product goals and self-evaluation versus no self-evaluation. For the self evaluation condition, the students were asked to reflect on what they had learned at the end of the first six lessons. For the students who were not assigned to the self-evaluation group, they were asked instead to evaluate how they felt about math at the end of each lesson.

Similar to the Schunk & Swatrz (1993) study, students in both learning goal conditions had significantly higher math self-efficacy and skill in comparison to the students in the product goal without any self-evaluation. Students in the product goal with evaluation condition were statistically indistinguishable from students in both learning goal conditions. Also in their study, self-efficacy for math, math skill, and persistence (time spent solving) on difficult problems were all highly correlated (r = .63 to .89). Again, these results highlighted the benefits of adopting process-type goals in learning contexts.

The second half of the Schunk (1996) study mirrored the first part with the following changes. At the beginning of the study, the students were not only asked to assess their efficacy for math, but also their perceived efficacy for learning math. Also at the end of the experiment, students were asked to evaluate how satisfied they were with their learning progress. A total of 40 fourth grade students from two different classes at one elementary school were involved in this study.

Again, students who were assigned a learning goal orientation displayed greater self-efficacy for math and were also more successful at completing problems than their counterparts who were assigned a product goal orientation. Students in the learning goal group also displayed higher interest in the task whereas the comparison students were very concerned about how their work compared to other students and students in the comparison group also tended to be work avoidant. Not surprisingly, self-efficacy for learning was positively correlated to the number of problems completed and the sense of satisfaction that students felt about their learning.

All of these studies, taken together, highlighted the role that goal setting can play in academic achievement. Goals that emphasized the learning process aided in the development of student efficacy by allowing students to see that their efforts were actually resulting in learning gains. These learning gains promoted self-satisfaction and fostered a greater sense of self-efficacy. This greater sense of self-efficacy affected how persistent students were in the face of difficulties and ultimately led to gains in achievement. In contrast, goals that focused only on product production did not allow students to assess if they were improving. These product production orientations, then, resulted in small changes in self-efficacy and little change in relation to achievement.

Self-efficacy and Self-regulation

While goal-setting can play a key role in enhancing student self-efficacy, self-efficacy is also significantly correlated with an individual's ability to self-regulate their own learning. Zimmerman & Martinez-Pons (1990) investigated the relationship between self-regulation and self-efficacy looking at both the effects of gender and schooling. They noted that, "little attention has been devoted to the relation between efficacy perceptions and students' use of self-regulated learning strategies" (pg 51). The authors also noted that this study was correlational and no attempt was made to determine causality between these two variables.

The study was conducted with students from 5^{th}, 8^{th}, and 11^{th} grade all within a large city on the east coast. Half of the student population was enrolled in a school for gifted children and the other half came from local public schools. To measure self-regulated learning, a tool was used that had been developed previously by Zimmerman and Martinez-Pons (1986) where students were asked to indicate how they would respond to various academic scenarios. Measurements of student self-efficacy were specific to verbal and mathematical self-efficacy. Each of these scales consisted of ten words or problems in which students were asked to assess their confidence in defining a word or solving a math problem. Students assessed their confidence using a scale that ranged from 0% confident to 100% confident.

When analyzing the results, Zimmerman & Martinez-Pons (1990) split students based on grade, gender, and whether or not they attended the gifted school. In terms of efficacy, not surprisingly, students' mathematical and verbal efficacy increased from grades 5 through ll. Also, gifted students displayed significantly higher self-efficacy in both math and verbal ability. For verbal self-efficacy, each increase between grades was significant. In contrast, the only gains that achieved statistical significance for mathematical efficacy occurred for the efficacy gains for students between the 5th and 8th grade. Looking again at the data for verbal self-efficacy, for gifted students, their verbal self-efficacy gains between the 5th and 8th grade were significant, whereas their gains from 8th to 11th grade were statistically insignificant. In contrast, the significant gains in verbal self-efficacy occurred between grades 8 and 11 for the public school students.

Self-efficacy in relation to gender revealed interesting results as well. The boys in this study possessed greater verbal self-efficacy than the girls. In contrast, gender did not lead to any differences in mathematical self-efficacy. Unfortunately, these results were not analyzed to see if this effect was more pronounced at different grade levels. These results were intriguing because they seemed to contradict more recent work that has investigated the interaction of gender and self-efficacy. For example, Jacobs et al (2002), noted that boys' self-efficacy in math tended to be greater than that of girls from kindergarten all the way through grade 10. At 10^{th} grade, however, self-efficacy for mathematics was about the same for boys and girls and remained the same through grade 12. In language arts, Jacobs et al (2002), found that girls' self-efficacy in language arts was greater than that of boys at the same grade. While the size of this self-efficacy gap varied from grade to grade, this gap was constant through all grades measured. In addition, whereas Zimmerman & Martinez-Pons (1990) found that student efficacy increased from grades 5 through 11 in their measures, Jacobs et al (2002) found that student efficacy generally decreased from grades 5 though 11 in their measures. While these apparently contradictory findings were most likely due to differences in the way that self-efficacy was measured, the results were intriguing nonetheless.

A literature review by Pajares (2003) only added to the confusion. Pajares (2003) noted that at grade 9, for example, boys had a greater sense of self-efficacy than girls as it related to writing. Yet when student writing was scored, girls tended to receive better marks on their writing. Also when asked to compare themselves to their peers, even though girls indicated lower language arts self-efficacy, they consistently stated that they thought that they were better writers than their peers who were boys. This was not as much the case with the boys. Boys tended to think that their writing skills approximated those of their peers who were girls. Pajares (2003) stated that part of this apparent contradiction might be found in what the students were using as their "frames of reference" when estimating their self-efficacy judgments. Boys may have been using a more external / comparative standard whereas the girls may have been judging themselves against a higher internal standard. Since science inquiry involves skills related to both math and writing, it will be intriguing to investigate the role that gender may play in relation to student self-efficacy.

In addition to self-efficacy, the Zimmerman & Martinez-Pons (1990) study also confirmed conclusions from other studies that have been reviewed in this proposal. Gifted students indicated that they used significantly more self-regulation strategies as a part of their learning than regular students. These results agreed with prior studies on writing in this review (Coleman, 1998; Scardamalia, Bereiter, & Steinbach, 1984; White & Frederiksen, 1998) which indicated that academically advanced students did not seem to find as much benefit from instruction in self-regulatory strategies presumably because they already currently used these strategies.

One of the limitations of the Zimmerman & Martinez-Pons (1990) study was that it only assessed student perceptions of self-regulation and student perceptions of self-efficacy. The study did not contain any attempt to verify that what the students perceived about themselves was indeed actually true. Zimmerman & Martinez-Pons (1986) did test to see if student perceptions matched their actions within a classroom context as a part of the verification study of the self-regulation instrument. Their results showed a high correlation between student perception and actual student activity. It would have been beneficial to see the same sort of validation take place with student self-efficacy in order to verify that student perception of efficacy did indeed match the actual ability to perform. This study proposal will seek to measure student self-efficacy in relation to scientific inquiry and evaluate the extent to which these self-efficacy measurements match the products that the students are able to produce.

Self-efficacy for Self-Regulation & Academic Achievement

The final two articles to be discussed in this literature review will be treated together because the studies share similar methodologies, conclusions, and authors. Both Zimmerman, Bandura, & Martinez-Pons (1992), and Zimmermand & Bandura (1994) highlighted the role that self-regulation played in influencing an individual's sense of self-efficacy for high academic achievement. To be more specific, both of the studies highlighted the importance of an individual's sense of their own self-efficacy to self-regulate their learning. It was this sense of self-efficacy to self-regulate that appeared to have a causal influence on final achievement that was mediated through a student's sense of self-efficacy to be academically successful.

The initial study by Zimmerman, Bandura, & Martinez-Pons (1992) took place with 102 students from two different high schools in a large city on the east coast. Self-efficacy for self-regulation was measured with a tool that was based on prior work related to self-regulation (Zimmerman & Martinez-Pons, 1986, 1990). Students' academic self-efficacy was measured by their responses to the question "How well can you learn: (biology, geometry, history…)" as it related to various academic areas. The students were also asked what grade they expected to receive in various courses. In addition, students were asked to indicate what would be the lowest grade that they could receive in these courses and still be satisfied. This study was exploratory in nature and so final achievement was measured by the grades that the students actually received in their courses.

Using path analysis, Zimmerman, Bandura, & Martinez-Pons (1992) found that the variables that had the most direct impact on student achievement were self-efficacy for achievement, and goals that students had set for themselves (ie. “what grade do you expect to receive”?). The significant role that student self-efficacy for academic achievement played was not surprising. The significant impact of self-efficacy for academic achievement on actual achievement has been well documented (Bandura, 1995, 2006a; Pajares, 2006; Zimmerman 1995, 2006). It was interesting that a student’s sense of self-efficacy for achievement not only affected the final grade, but this sense of efficacy also played a role in the goals that students set for themselves. Students who felt that they were capable of accomplishing an academic task also set higher goals for themselves. As has already been noted in this review, student goals and student goal orientation have played a significant role in student achievement. These results again highlighted the importance of developing a student’s sense of their own academic self-efficacy.

In addition to academic self-efficacy, self-efficacy for self-regulation also played a role. Even though a student may already possess the self-regulatory skills necessary to be successful, if they do not feel like they have the ability to use these skills in the specific tasks of various academic disciplines this significantly affects their self-efficacy for achievement. These results were mirrored in a follow up study by Zimmerman & Bandura (1994). This study was conducted with students at the collegiate level. The total sample contained 95 university students who were enrolled in various writing classes. Self-efficacy for self-regulation was assessed specific to a writing context whereas self-efficacy for self-regulation in the Zimmerman, Bandura, & Martinez-Pons (1992) study was measured in a more global sense. Similar to the prior study, self-efficacy for achievement was measured in terms of what grade in the course students thought that they could achieve. Students were then asked what grade they expected to receive from the course (grade goal) in addition to what was the lowest grade that they could receive and still be satisfied. Again, Zimmerman & Bandura (1994) used final course grades that each student received as the measure of academic achievement.

Using path analysis, Zimmerman & Bandura (1994) noted that self-regulatory efficacy played a significant role in influencing final grades through self-efficacy for academic achievement. In this study, Zimmerman & Bandura (1994) used verbal scores from SAT exams to assess verbal aptitude. Generally speaking, they found that students with higher verbal aptitude also had higher self-evaluative standards. The influence of self-efficacy for academic achievement on grade goals and on final grades also mirrored that of the previous study. In assessing the strength of this model, Zimmerman & Bandura (1994) pointed out that the high degree of agreement between the paths and the relative strength of each path gave credence to this particular model.

In relation to this research study proposal, both of the previous studies highlighted the importance of developing student confidence in the use of the self-regulatory skills. Development of these self-regulation skills led to greater efficacy to use these skills which in turn promoted higher efficacy for academic success.

Section Summary

A number of key ideas were discussed in this section. Foremost, the connections between self-efficacy, goal orientation, and student achievement were again highlighted (Bandura & Schunk, 1981; Schunk, 1996; Schunk & Swartz, 1993). Secondly, it was unclear the role that gender played in relation to self-efficacy assessments (Jacobs et al, 2002; Pajares, 2003; Zimmerman & Martinez-Pons, 1990). This dilemma becomes even more intriguing as gender in science education has become more of a central issue (Bell, 2001; Dawson, 2000; Greenfield, 1997). Finally, the causal relationships between self-efficacy for self-regulation, self-efficacy for academic achievement, student goals, and achievement outcomes (Zimmerman & Bandura, 1994; Zimmerman, Bandura, & Martinez-Pons, 1992) provided even more reason to focus classroom time teaching self-regulatory skills to students in order to bring about positive changes in student achievement.

Chapter Conclusion

The studies that have been a part of this literature review highlighted a number of ideas in relation to the processes of writing and scientific inquiry. Student goal orientation towards learning of processes as opposed to completion of product is key in helping students move from novice understandings of the writing process towards more expert orientations. As such, the ideas that will be explored in this study will most likely be of greater benefit to low and average achieving students. Many high achieving students are successful because they naturally adopt these more process focused orientations. The adoption of process oriented goals also implies that time and space must be set aside in order to allow students to reflect on their progress and movement towards goal attainment. Students who were prompted to reflect tended to produced significantly better work than those who did not. Also, as students spent time self-reflecting on their goals, this process also had a positive effect on their self-efficacy judgments. A student’s sense of self-efficacy was positively related to gains in achievement, higher personal learning goals, and even greater intrinsic interest in the specific activity at hand. Increasing student achievement is the goal of this study.

The literature that has been reviewed in this chapter also indicated that much of the work that has been done in relation to scientific inquiry has been directed towards elementary and middle-school aged students. This study seeks to build on this base of knowledge and explore the extent to which the strategies employed with these younger students can also be used to enhance the learning of high school students.

Chapter 3: Design & Methods

Proposed Study Model

The goal of this study was to build on current understandings of the cognitive and metacognitive skills necessary for students to be successful in science inquiry tasks. The model of scientific inquiry used in this study was based on the guided-inquiry cycle of Magnusson & Palincsar (2005). While Magnusson & Palincsar (2005) discussed the self-regulatory metacognitive monitoring that students must employ in this cycle, the self-regulation framework of Zimmerman & Clary (2006) provided a richer model of this process. A combination of these two frameworks yielded the model used to inform the work in this study (see figure 10).

In the guided-inquiry model, science inquiry was conceived of as two complementary processes that worked in concert to produce a final product. In the initial phase, students design and then carry out an experiment to address a question. In the subsequent phase, students determine how to best communicate the results of the experiment. This reporting often takes the form of a written project in high schools. The model of writing proposed by Bereiter & Scardamalia (1987) complements the guided inquiry model of Magnusson & Palincsar (2005) in that both models propose that students engage in metacognitive thinking about their actions. The Bereiter & Scardamalia (1987) construct of knowledge-transformative writing informed the writing instruction in this study.

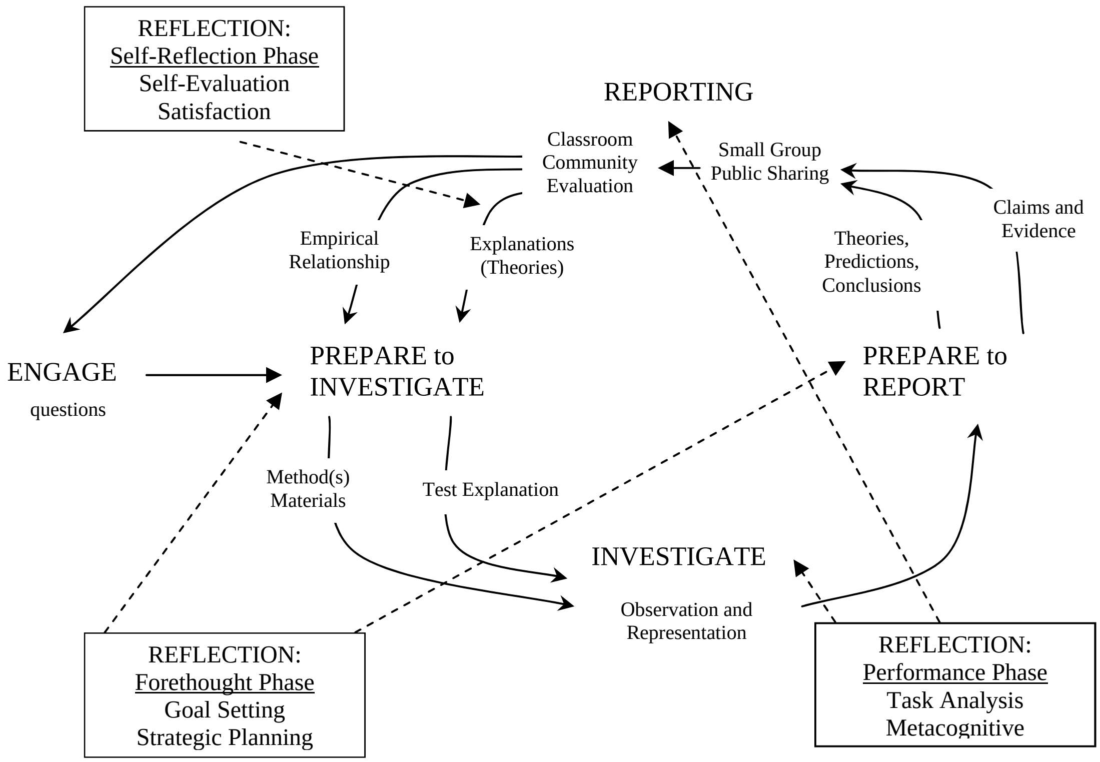


Figure 10: Guided-Inquiry and the Self-Regulation cycle adapted from work by Magnusson & Palinscar (2005); Zimmerman & Cleary (2006)

In addition to the focus on the writing process, this study also sought to investigate the nature of reflective prompting and the extent to which directed versus generic prompting would enhance both student success and student self-efficacy. Generic prompts asked students to respond to the sentence prompt, “Right now I am thinking about…”. In contrast, directed prompts asked students to reflect on a specific action they are doing such as, “Does the data you have collected so far make sense? Explain”. As students moved through the guided-inquiry cycle, some of the students in the treatment groups were prompted to reflect. These students were prompted during the Investigation phase, the Reporting phase and then finally, one time at the end of the investigation. The nature of these prompts varied based on treatment group membership.

Demographic Data

The study took place at Suburban High School (not the real name). At the time of the study the school enrolled approximately 1700 students, the majority (73%) of which were Caucasian. Other ethnicities present within the high school were Hispanic, Asian, African, and Native American who composed 16%, 6%, 3%, and 1% of the student population respectively. In this study, the demographic breakdown was similar. The majority of the students in the study sample were Caucasian (82%). The other ethnicities in order of percentage were Asian (8%), Hispanic (7%), and African American (3%).

The school also had approximately 100 students (6%) enrolled in English Language Learner (ELL) programs; the majority of these ELL students were Hispanic (85%). The percentage of Hispanic students in ELL programs had increased by about 10% over the past three years. The number of students in special education programs was slightly higher with about 150 students enrolled (9%). The majority of these students were Caucasian (67%) while the rest were predominately Hispanic (16%). There were approximately equal numbers of male and female students at the school and this trend was mirrored in these special programs as well. There were approximately equal numbers of males and females who chose to participate in this study. In addition, none of the ELL students or those in special education programs chose to participate in this study.

Approximately 33% of the students at Suburban High School qualified for free and reduced lunch. This number had increased every year since 2001-2002. That only 33% of the students qualified was probably too low as both middle schools that fed into Suburban High School both had over 50% of their students eligible for free and reduced lunch. It was not known how many of these students from low income backgrounds were a part of this study.

Achievement data from the state tests taken over the past five years indicated that Suburban High School's passing rates tended to be slightly higher than the state average. This was true for the tests in reading, math and writing. The historic trend lines for both Suburban HS and State of Oregon passing rates generally mirrored each other. The data indicated that Suburban HS students tended to do better on the writing and reading assessments, but struggled more so with the math assessment. Also of note, passing rates on the reading and writing tests had increased over the past five years whereas the math passing rates were basically flat. This data indicated that Suburban High School appeared to be an average high school in terms of student academic achievement.

Table 1: Number of Suburban High School students meeting or exceeding the state benchmark in Reading.

	2005	2006	2007	2008	2009
Students in this study					90 %
Suburban HS	43 %	58 %	67 %	63 %	68 %
State	54 %	55 %	65 %	65 %	66 %

Table 2: Number of Suburban High School students meeting or exceeding the state benchmark in Math

	2005	2006	2007	2008	2009
Students in this study					83 %
Suburban HS	39 %	43 %	63 %	60 %	51 %
State	47 %	45 %	55 %	52 %	54 %

Table 3: Number of Suburban High School students meeting or exceeding the state benchmark in Writing

	2005	2006	2007	2008	2009
Students in this study					93 %
Suburban HS	50 %	59 %	64 %	41 %	68 %
State	52 %	57 %	59 %	44 %	55 %

Table 4: Average Suburban High School student SAT scores from the 2008-09 AYP Report Card

	Suburban HS	State	Nation
Critical Reading	509	523	501
Math	512	525	515
Writing	485	499	493
Percentage Tested	30 %	52 %	45 %
Number Tested	135	18,016	1,518,859

While the state testing data indicated that Suburban High School was average to slightly above average when compared to the rest of the state, Table 4 painted a slightly different picture. Suburban High School students scored above the national average on the SAT trait of Critical Reading, but fell below both State and National averages on the SAT traits of Math and Writing. The state tests compared sophomore students against sophomore students whereas the SAT compared junior and senior students against each other. Also, only 30 % of Suburban HS students actually took the SAT, which means that the SAT data is not necessarily representative of the entire Suburban HS population.

Table 5: Attendance, Graduation and Dropout rates from the 2008 – 09 AYP Report Card

	Suburban HS	State
Attendance Rate	89 %	91 %
Graduation Rate	87 %	84 %
Dropout Rate	3.2 %	3.6 %

Table 5 lists information gleaned from the 2008 – 2009 Report Card issued by the State of Oregon. The attendance rate at Suburban HS was slightly lower than that of the state on average. In spite of slightly lower attendance, Suburban High School graduated slightly more of its students than an average high school in Oregon. In addition, the dropout rate at Suburban High School was slightly lower than the state average as well. Suburban High School was rated as Satisfactory by the state on 2008

- 2009 school report card. Taken together, this data indicated that Suburban HS was an average high school.

The students involved in this study, though, were not the average students at Suburban HS (see Tables 1, 2, & 3). When compared to the general student population, the students involved in this study scored much higher on measures of math, reading, and writing ability as assessed by state testing. Generally speaking, the students in this study were high performing students.

Participants

A total of four chemistry classes were involved in this study with class sizes ranging between 25 – 32 students. There were a total of 121 students in these four classes. Of these 121 students, only 61 volunteered to take part in the study. The identity of the students in the study was not known until after all of the data had been collected. Only the work from these 61 students was used in this study.

All of these students were either sophomores or juniors and all had chosen to take the course as an elective science offering. General chemistry was an introductory chemistry class designed to provide students the skills and experiences necessary to be successful in future college science classes. While the course was open to all Suburban High School students, more than half of the students indicated on a class survey that they were taking chemistry because they planned to go on to college after graduation. Each of the four classes was randomly assigned one of four experimental conditions. Student placement in classes was not completely random due the nature of scheduling constraints due to the master schedule; therefore one-way ANOVA's were used to assess the initial homogeneity of the four chemistry classes involved in this study.

Part of the design of this study required that students be identified as either high achieving or low achieving students following the work of White & Frederiksen (1998). White & Frederiksen (1998) chose a cut score to differentiate between high and low achieving in order to divide their sample "into halves as evenly as possible with regards to all the different factors (ie. treatment, grade level, etc.)" (p 28). In reviewing the first semester grades for the students in this study, approximately half of the students in each of the classes had first semester grades in their chemistry classes of 85% or higher. The 85% mark seemed to be a natural break point across all of the classes (lowest percent in the high achieving group: 86%, highest percent in the low achieving group: 82%). First semester grades were used to determine achievement grouping because no interventions were introduced in the first semester. For this reason, students who had earned an 86% or higher in their first semester of Chemistry were considered to be high achieving students for the purposes of this study. Low achieving students were those who had earned an 82% or less in their Chemistry class in the first semester.

Student grades from the first semester were based predominately on test scores. Three tests were responsible for 50% of the overall grade. Labs and daily assignments were responsible for the other 50%. Only one inquiry task was completed during the first semester and the score from this task was included in the first semester grades. The inquiry task represented approximately 4% of the semester grade and was intentionally included to help differentiate high and low achieving students. This study was designed to assist students who struggled with science inquiry tasks and so the inclusion of this initial inquiry task score helped to ensure that students were assigned to the appropriate achievement group. This inquiry score, though, was not the sole basis for determining achievement. Knowledge of content was also an important outcome of the course. This initial inquiry task was an integral part of the curriculum. Student work on the inquiry task was one means of assessing student knowledge of course content. The initial inquiry task was a summative assessment and needed to be evaluated in light of other assessments of student knowledge in order to determine a student's prior achievement. For this reason, the student's first semester grades were used to assign individuals to the two achievement levels used in this study.

Study Instruments

Chemistry Self-Efficacy Survey (CSES)

The CSES was designed to contain two subscales where students responded to questions that had been framed in a Likert-type format. Ten of the questions dealt with chemistry content knowledge. The content questions were based on material that would be covered in a general chemistry course. The content of these questions was based on recommendations from the report, *Understanding University Success* (Center for Educational Policy Research, 2003). The report was the result of a two-year collaborative effort that included more than 400 faculty and staff from twenty US research universities that outlined skills and understandings deemed essential to college success. Recommendations were broken down by content area. These questions were reviewed by two other high school chemistry teachers who had between 12 and 27 years of experience teaching science. The questions were also reviewed by two science education experts who taught in the Science and Mathematics Education department at a local university. These experts both had prior science classroom experience and have been conducting research in science education for over 25 years. Survey questions were modified based on their recommendations.

In addition to content, eleven questions on the CSES addressed issues related to science inquiry. These eleven questions were based on the Science Inquiry Scoring Guide used by the state of Oregon (Oregon Department of Education, 2008). Oregon subdivided inquiry into four skills: Forming a Question or Hypothesis (Framing), Designing an Investigation (Designing), Collecting and Presenting Data (Collecting), Analyzing and Interpreting Results (Analyzing). The Science Inquiry Scoring Guide provided a rubric to help teachers assign students a score ranging from 1 to 6 in each of the four dimensions. In order to be considered proficient, a student needed to receive a score of a least a 4 on the dimensions of the inquiry which were assessed. The eleven questions in this section were reviewed by the same two science education experts that reviewed the content section of the CSES and were modified based on their recommendations.

The CSES was given in an electronic format. The students were given a paper copy of the survey but were asked to provide their responses via an electronic student response unit. All copies of the CSES were collected after each administration. The students had ten seconds to respond to the question before the computer program automatically advanced to the next question. The time frame was chosen based on prior work by Bandua & Schunk (1981) which recommended that students should have enough time to read and respond to a prompt, yet should have too little time to actually solve any of the problems. Self-efficacy is a measure of belief about ability and so this short time-frame helped to ensure that student responses were beliefs about their ability not their actual skill level.

Every student had their own response pad that they used to send their response to a central receiving device. The response pad keys were labeled “a” through “h”. Because the response pads limited students to eight response options, students were asked to respond on a 7-point scale. Students also had the option responded with an “h” if they choose not to respond to a particular question. The computer program logged each student’s responses. Student responses to each statement were translated to a number value where “a” = 1 up to “g” = 7. Student responses on the two scales were averaged to create a single value for each scale on the instrument. A low average indicated a low sense of self-efficacy whereas a high value indicated a strong sense of self efficacy. At the time of the study, this was the third year that students had been using these electronic response units in science classes at the high school.

The results from the CSES survey were checked for both validity and reliability. Confirmatory Factor Analysis from the results of the survey provided strong evidence that the instrument only contained two factors. The analysis of the questions revealed that one factor contained questions related specifically to chemistry content whereas the other factor contained questions related to science inquiry. The Cronbach's α values for the chemistry content self-efficacy factor ranged from 0.77 to 0.92. Cronbach's α values for the science inquiry self-efficacy factor was slightly higher and ranged from 0.85 to 0.95.

Science Inquiry Scoring

When students handed in their science inquiry samples they were handed in with student ID numbers being the only identifying marker. A colleague randomized the student samples and the researcher initially assessed all of the samples using the State of Oregon Science Inquiry Scoring Guide. The researcher had attended multiple training sessions offered by the state on the use of the scoring guide. Randomization and Student ID's ensured that the researcher was unaware which class a particular sample was from. The scoring guide used a scale of 1 to 6 to rate students on four dimensions of inquiry. The researcher scored each student on each dimension and also recorded the sum of the student scores across the four dimensions. The state stipulated that a score of 4 or higher on a particular dimension was needed for that dimension to be considered passing. Once the papers were assessed, the papers were re-sorted by class and returned to the students so that they could see their scores.

Copies of the student work that contained only student ID numbers were retained for further analysis in this study.

To assess the scoring reliability of the researcher in this study, a group of four science teachers was assembled at the conclusion of the 2008 – 2009 school year. These four teachers had between 6 and 27 years of experience teaching science with the average being 16 years of experience. Before any samples from the study were scored, the team took some time to cross score two anchor papers provided on the Oregon Department of Education's website. The teachers on this scoring team had not seen these anchor papers before. After scoring each anchor paper, the team discussed their scores relative to each other and relative to the score that the State of Oregon had assigned. The goal was to gain consensus around the language and intent of the scoring guide. This initial process took about forty minutes to complete and was facilitated by the researcher who had been trained to use the scoring guide.

This initial training was followed by three scoring rounds. The inquiry samples were divided into three groups based on the whether the samples came from the first, second, or third phase of the study. This allowed the teachers to compare samples that were written about similar topics. The team was not informed which samples were collected at the beginning of the study and which samples were collected at the end. In addition, the teachers were not told which papers were from experimental conditions and which ones were the from the control group. Each round began with a cross-scoring activity that lasted about forty-five minutes. The team picked eight random samples to cross-score from a stack of inquiries from one of the study phases. Then, they each scored all eight samples. Once everyone on the team had scored all eight samples, they compared their scores and came to a consensus for each of the samples. Once the initial samples were moderated, the team divided up the remaining samples and scored them individually. Inter-rater reliability was compared against the marks that the researcher had given and was calculated using the Kappa Measure of Agreement. Values ranged from 0.04 to 0.67 (see table 6). This statistic was run again, except, agreement was defined as each rater being within one point of each other (ie. if one rater scored a section as a 3 and another scored the same section as a 4, these raters were considered to be in agreement). With this adjustment, the inter-rater reliability rating was much higher and ranged from 0.43 to 1.0 (see table 7). Because the inter-rater reliability was much higher with the adjusted score agreement, inquiry scores used in this study were the average of the scores assigned by the raters.

Table 6: Inter-rater reliability for score agreement

	Rater 1	Rater 2	Rater 3	Rater 4
Inquiry 1 – Framing Section	0.25	0.32	0.27	0.21
Inquiry 1 – Analysis Section	0.04	0.28	0.44	0.32
Inquiry 2 – Framing Section	0.55	0.57	0.30	0.40
Inquiry 2 – Analysis Section	0.30	0.67	0.67	0.13
Inquiry 3 – Framing Section	0.41	0.06	0.18	0.34
Inquiry 3 – Analysis Section	0.10	0.17	0.22	0.08

Table 7: Inter-rater reliability for adjusted score agreement

	Rater 1	Rater 2	Rater 3	Rater 4
Inquiry 1 – Framing Section	1.00	1.00	1.00	0.43
Inquiry 1 – Analysis Section	1.00	1.00	1.00	0.63
Inquiry 2 – Framing Section	1.00	1.00	1.00	1.00
Inquiry 2 – Analysis Section	0.64	1.00	1.00	1.00
Inquiry 3 – Framing Section	1.00	1.00	1.00	0.64
Inquiry 3 – Analysis Section	1.00	1.00	1.00	1.00

Science Inquiry as Knowledge Transformation Study Design

Phase I Mass Conservation Inquiry	**Phase II** Gas Law Inquiry	**Phase II** Acid / Base Inquiry	**Phase IV** Student Feedback
Phase I began approximately five weeks into the school year. Self-Efficacy Survey: Phase I began when all classes in this study took the Chemistry Self-Efficacy Survey (CSES) Science Inquiry Task: Following the Self-Efficacy Survey, all classes in this study completed an inquiry task that utilized student knowledge related to mass conservation in an chemical reaction. There was no differentiation of instruction by experimental condition in Phase I .	Phase II began approximately nineteen weeks into the school year. Self-Efficacy Survey: Phase II began when all classes in this study took the Chemistry Self-Efficacy Survey (CSES) Science Inquiry Task: Following the Self-Efficacy Survey, all classes in this study completed an inquiry task that utilized student knowledge related to gas laws. There were a total of four chemistry classes and each class was a different experimental condition. The descriptions of each experimental condition are described below.	Phase III began approximately twenty-six weeks into the school year. Self-Efficacy Survey: Phase III began when all classes in this study took the Chemistry Self-Efficacy Survey (CSES) Science Inquiry Task: Following the Self-Efficacy Survey, all classes in this study completed an inquiry task that utilized student knowledge related to acids/bases. There were a total of four chemistry classes and each class was a different experimental condition. The descriptions of each experimental condition are described below.	Phase IV began approximately Thirty-four weeks into the school year. Self-Efficacy Survey: Phase IV began when all classes in this study took the Chemistry Self-Efficacy Survey (CSES) Science Inquiry Task: There were no inquiry tasks in this phase although students were asked to turn in a sheet where they reflected on what they had leaned over the school year related to inquiry.

Class A (Control): This class was randomly chosen to act as the control for this study. No modifications were made to the instruction in this class.

Class B (Proximal / Process Goal; Sentence Stems) : This class was asked to adopt a proximal / process goal for this study and was given instruction related to the guided inquiry science process and also to the use sentence stems.

Class C (Proximal / Process Goal; Sentence Stems; Directed Reflection Prompts) : This class experienced the same instruction as Class B with the addition of directed reflection prompts as they engaged in scientific inquiry.

Class D (Proximal / Process Goal; Sentence Stems; Generic Reflection Prompts) : This class experienced the same instruction as Class B with the addition of generic reflection prompts as they engaged in scientific inquiry.

Figure 11: Diagram of the design of the study

Study Design

This study took place in four phases (see Figure 10). The initial phase was the control which established a baseline for the rest of the study. The second and third phases introduced the various experimental conditions. These two phases were identical except for the inquiry prompts that were used. In the final phase, students were assessed one last time on their self-efficacy related both to chemistry content and science inquiry. This study took place over the course of the 2008 – 2009 school year.

Phase I – Establishing a Baseline

Phase I was the control phase of this study. This phase began as students completed the Chemistry Self-Efficacy Survey (CSES). At that point, the students had not completed any inquiry tasks as a part of the chemistry course, so their responses were solely related to whatever their experiences in inquiry had been during prior school years. It was anticipated that student scores for the Chemistry Content Self-Efficacy strand would be low as much of the content on the survey had not yet been covered. It was also anticipated that the student scores for Science Inquiry Self-Efficacy would also be low as the science inquiry passing rate was low for the school.

Once the CSES was administered, all four classes were given instruction on the concept of the mole in addition to instruction on how to calculate percent composition and molar mass. The concept of molar ratios was emphasized with a lab where various chloride salts reacted with silver nitrate. The concept of conservation of mass was also reviewed at that point and was verified via in-class demonstrations. Leading up to the inquiry task, the students were shown how chlorate compounds tend to decompose rapidly when heated. A gummy bear was added to this decomposition reaction which caused the reaction vessel to flame and hiss. Using the chemical equation for a combustion reaction, the students were shown how this provided evidence that the gas given off was oxygen and that only a binary salt was left in the reaction vessel. The students were asked to design an experiment in which they would determine the identity of two unknown chlorates that they would be given. The students were given three class periods in which to plan, experiment, and then write their science inquiry projects. Students were encouraged to consult the internet, use their textbooks, or confer with other student groups as they planned their experiment. Supplies and materials in the stockroom were available to use as needed. Students were asked to turn in their projects within two weeks. A copy of the inquiry prompt that the students received has been placed in the appendix.

This task had both teacher directed and student directed elements (NRC, 2000). The exploration was guided by a question that was given by the teacher. The students, then, determined what type of data to collect and how this would be done. The students were then to formulate explanations for their data on their own and connect their findings back to the classroom content. Students were given generic questions to consider as they processed their results, but that was the extent to which they were assisted in interpreting their results. While the initial part of the inquiry task was more teacher directed, the majority of the task was a student directed endeavor.

Scores from this inquiry task were assessed using the State of Oregon Science Inquiry scoring guide. All student work was assessed by the researcher and then returned to the students. The students had not received any instruction in the use of the State of Oregon Science Inquiry Scoring guide at this point in the study. Inquiry task averages were analyzed with a one-way ANOVA in order to test for the homogeneity of the classes in the study. It was anticipated that there would not be any statistical difference between any of the classes.

Phase II – Introduction of Treatment Conditions

The second phase of the study began nineteen weeks into the school year. All of the students were given the Chemistry Self Efficacy Survey (CSES) regardless of treatment condition. Following the survey, all of the classes in this study spent one day reviewing the state of Oregon science inquiry scoring guide.

To review the scoring guide, the students were given two sample science inquiry anchor papers that the state of Oregon had posted on its website (copies of the anchor papers are included in the appendix). Before the anchor papers were scored, students were asked to read through the scoring guide in order to determine what differentiated a score of “3” from a score of “4” in each of the scored dimensions. Students shared their ideas with those around them and then the researcher guided a whole class discussion.

After each class reached consensus on the meaning of the scoring guide language, the students then scored one of the work samples. All of the students wrote their marks on a whiteboard in the front of the classroom. This allowed students to compare their scores with those of their peers. A class discussion followed. The students were then shown the scores assigned by the state. This process was repeated for the second anchor paper. Student marks generally agreed with those assigned by the state.

The purpose for this scoring activity was two-fold. Primarily, the task was to ensure that students understood the criteria of the scoring guide. The secondary purpose was to provide concrete examples of what a passing paper looked like in comparison to a paper that was only close to passing.

In terms of content, students had been introduced to the concept of the mole, ionic and covalent compounds, and stoichiometry in the preceding eighteen weeks. The inquiry project in this phase of the study was conducted at the conclusion of a unit on gas laws. The students had seen demonstrations about the power of air pressure and had worked with various gas law equations. In the lab, they had worked with computer based pressure sensors in order to understand the relationship between temperature, volume, and pressure of a gas.

Prior to the inquiry task, the students were given a demonstration on how to determine the molar mass of a gas inside a cylinder of canned air commonly used to blow the dust off of electronics. Based on the demonstration, the students were asked to design their own experiment to determine the molar mass of the gas inside a cigarette lighter. Many of the students claimed that they already knew that the lighter contained butane, so they were asked to determine whether or not their data verified this assumption.

Similar to the initial inquiry task, for this second task, the students were given a question to explore. They were to use whatever resources they could find in order to design an experiment that would yield data to answer their question. Students only received help in developing their explanations and evaluating their results if they were in on of the treatment conditions. Again, this inquiry task was initially very teacher directed and then became more student directed once the investigation began.

The treatment groups were distributed in the following ways:

Class A (Control):

This class was randomly chosen from among the four general chemistry classes that were a part of this study. There were a total of twenty-eight students in this class, but only ten of the students in this class agreed to participate in the study. The researcher did not know which students had chosen to participate in the study until the final course grades were calculated at the conclusion of the school year. While all of the students in this class completed the same inquiry tasks as the other chemistry classes, no changes or modifications in teaching were made with this group.

This group acted as the control and allowed the researcher to monitor the extent to which changes in the treatment groups were due to differences in treatment or differences in the inquiry prompts themselves. It was anticipated that the science inquiry scores for the students in this group would not change significantly over the course of this study. On the CSES, it was anticipated that this group would see a general increase in their sense of self-efficacy for chemistry content while their sense of self-efficacy for science inquiry tasks would remain flat or even decrease. Considering the cyclical nature of self-efficacy, it was predicted that as student performance remained below state benchmark standards, students would begin to believe that they simply were not very good at completing inquiry tasks.

Class B (Proximal / Process Goal; Writing Stems):

This class was randomly chosen to receive partial application of the treatment. Of the 31 students in the class, 16 students volunteered to participate in the study. The researcher did not know which students in the class had chosen to participate until the final course grades were calculated at the conclusion of the school year. Prior to the second inquiry task, time was spent explicitly going through the guided-inquiry cycle (Magnusson & Palincsar, 2005). This took approximately half of a 90 minute class period. In this instructional time, students were asked to copy down the guided inquiry model used in this project and also asked to reflect on what would take place in each stage. Student responses were shared with the rest of the class in the format of a class discussion. Since these students were not prompted to reflect, they were presented with the guided-inquiry model that did not contain the self-regulatory reflection piece.

Students in this treatment were also given instruction on the nature and efficacy of goal setting in relation to student success. This process took approximately twenty minutes and concluded with students writing both a proximal and a distal goal. Students then compared their goals with a peer. Student volunteers were asked to read their responses and the class judged whether they thought that the goal was proximal or distal. Students were then asked to adopt a proximal / process goal for the upcoming inquiry task. Even though students were asked to adopt a proximal / process goal for the study, they were not asked to write this goal down.

These students were also introduced the knowledge-telling versus knowledge-transforming (Bereiter & Scardamalia, 1987) models of writing and the sentence stems that they were asked to use as they wrote their inquiry reports. Students were asked to write down the differences between the two writing models (see table 8). These characteristics were then compared back to the two benchmark papers from the state so that students could have a concrete example of these two writing styles.

Table 8: Comparison of Knowledge-Telling versus Knowledge-Transforming writing

Traits of Knowledge-Telling Writing	Traits of Knowledge-Transforming Writing
• No clear organization (ideas seem to be in random order) • Reads like a "stream of consciousness" • Lacks evidence of editing • Pieces tend to be short or highly repetitive • Includes information that is not necessarily relevant to the topic	• Clear focus • Ideas are clearly connected to each other • Structure of the writing (placement of topics, use of paragraphs, etc.) supports the overall purpose of the piece • Evidence of editing is present

The sentence stems for the inquiry write-up were given to the students approximately four weeks prior to the inquiry task in this second phase. Students were asked to glue the half-sheet of these sentence stems into their science notebooks on the back of the front cover so that these would be easily accessible. Every class period began with a question or problem that either related to the content from the prior class or to the content that would be learned that day. Students were given about five minutes to respond to this opening question / problem which was then followed by a class discussion. After the class discussion, students were asked to summarize what they had learned. This practice of summarization had been used consistently since the beginning of the school year, so the students were well versed with the process. For the purposes of this study, the students were asked to refer to the sentence stems that they had been given and find appropriate stems that could be used for the summarization activity. Each day, between four and eight students were asked to share their summaries with the rest of the class. This allowed the students to become familiar with the sentence stems that were available and also to hear how the sentence stems were used by other students.

As was already stated, students in this treatment condition were not prompted to reflect during their inquiry tasks. The purpose for this was to see the effect of reflection on student success. It was anticipated that for students in this group, the results would differ based on prior student achievement. The assumption was that high-achieving students already possessed well developed metacognitive skills (Davis, 2003; White & Frederiksen, 1998) and that they were already naturally reflective about their learning. It was anticipated that the sentence stems would benefit the high achieving group because these stems were assumed to expand on the already well developed language skills that these high achieving students already possessed. In contrast, it was assumed that low-achieving students tended to operate with metacognitive strategies that were not as well developed for thinking and writing. It was predicted that the introduction of sentence stems without the scaffolds provided by reflection would result in limited application of these sentence stems. The result was predicted to be one of little change in the science inquiry scores of these low achieving students over time.

In relation to the CSES, as with the control group, self-efficacy related to content knowledge was expected to increase over the time and whereas self-efficacy related to science inquiry was expected to remain fairly flat or even decrease. Low-achieving students were predicted to not benefit from this treatment; so their efficacy related to science inquiry was expected remain the same or potentially decrease based on feedback from the initial inquiry task. In contrast, the high achieving students were expected to benefit from an introduction to sentence stems. Even though high-achieving students were expected to benefit from the treatment, they were assumed to already be fairly self-efficacious related to science inquiry and so their self-efficacy scores would remain fairly flat.

Class C / Class D (Proximal / Process Goal; Writing Stems; Directed Reflection Prompts or Generic Reflection Prompts):

Classes C & D received the same training in goal setting as Class B and also were given the same instruction related to the guided inquiry model, knowledge-telling versus knowledge-transforming writing, and the use of sentence stems. These two classes were grouped together for the purposes of explaining the procedure treated as the only difference between the two was the nature of the prompts given to promote reflection. This study sought to understand the generalizability of Davis' (2003) findings that generic reflection prompts led to more effective reflection than directed reflection prompts. There was significant agreement in the literature regarding the relationship between effective reflection and achievement (Bransford, Brown, & Cocking, 2000; Davis, 2003; Flick & Tomlinson, 2006; White & Frederiksen, 1998). To investigate the effects of prompt types, the students in Class C were given directed reflection prompts with specific items to respond to as they reflected. Out of the thirty-two students in this class period, fifteen students elected to participate in the study. Again, the researcher was not aware of which students in the class were part of the study until final course grades were handed in at the end of the school year. During the Investigation phase of the inquiry task, the students were asked to reflect on the following three prompts:

1. My goal for the investigation phase of this project is…
2. As you begin planning your experiment, brainstorm a list of ideas, topics, and procedures that are relevant to this investigation.
3. At some point while you are doing the experiment, your teacher will ask you to respond to this question. "Do the data that you have collected so far make sense? Do you feel that any adjustments should be made to your procedure? Explain"

During the Reporting phase of the inquiry task, the students were asked to respond to the following three prompts:

1. My goal for the reporting phase of this project is…
2. As you begin planning your report, brainstorm a list of ideas, topics, formulas, and vocabulary words that you think you might use as you write. Put a star next to the items that you think are the most important.
3. At some point while you are doing the work on your report, your teacher will ask you to respond to this question. "Does your report have a clear focus? What ideas do you still plan to include? Explain"

These prompts directed the students to focus on specific aspects of their performance related to their inquiry task work at different phases within the cycle.

In contrast, the students in Class D were given generic reflection prompts where, similar to the Davis (2003) study, students responded to the reflection prompt, "Right now I am thinking about…" Out of thirty students that were in the class, twenty of the students elected to participate in the study. The researcher was not aware of which students had elected to participate until after final grades for the course had been handed in at the conclusion of the school year. Similar to Class C, there were three prompts given in the Investigation phase and three prompts given in the Reporting phase. In the Investigation phase, the students responded to the following prompts:

1. My goal for the investigation phase of this project is…
2. As you begin planning your experiment, please finish this sentence, "Right now I am thinking about…"
3. At some point while you are doing the experiment, your teacher will ask you to finish this sentence, "Right now I am thinking about…"

The prompts given during the Reporting phase were as follows:

1. My goal for the reporting phase of this project is…
2. As you begin planning your report, please finish this sentence, "Right now I am thinking about…"
3. At some point while you are doing the work on your report, your teacher will ask you to finish this sentence, "Right now I am thinking about…"

These generic prompts allowed for a much broader range of responses.

Since both classes experienced the same conditions related to setting goals, there was no difference between the goal setting prompts given to Class C and Class D. Students in both conditions were encouraged to focus their reporting goals on the use of sentence stems. Examples of the two prompting type sheets that were handed to the students are given in the appendix. With both classes after they were prompted to reflect, students were asked to share with a partner what they had written. In addition, time was given in both conditions so that students could share their responses with the rest of the class. Only students who volunteered were chosen to share. This allowed all students to hear a variety of reflections.

The class period after the data had been collected, each team of students was given between 3 and 5 minutes to talk about the data that they had collected and their initial findings. All of the teams wrote copies of their data sets on a board that was visible to the whole room so that the students could see the range of data other groups had collected. After each team had presented, the class was allowed to ask questions of each team. This was done to model the Reporting phase of the guided inquiry cycle, and also provide students the opportunity to see how other teams analyzed and interpreted their data.

For both of these classes, prior to turning in their inquiry tasks, the students were asked to self-assess their projects and predict what scores they anticipated receiving on their work. These predictions were collected by a colleague and held until the end of the study and then released to the researcher. This was to ensure that these student predictions did not affect the assessment of the inquiry tasks.

It was anticipated that both Classes C and D would exhibit an increase in their science inquiry scores. While it was anticipated that there would be little change in either Class with the high-achieving students, the low achieving students were expected to benefit from the specific instruction in inquiry and writing when paired with reflection prompts. More effective reflection would lead to a greater success in completing science inquiry. Since the research seemed to be divided, no predictions were made regarding the effectiveness of either reflection prompting type.

As science scores increased, it was anticipated that there would also be a corresponding increase in student self-efficacy as measured by the CSES. As with the other Classes, student self-efficacy for content was expected to increase throughout the year. Student self-efficacy for science inquiry tasks was expected to increase as well. Similar to Class B, it was predicted that there would be little self-efficacy change for the high achieving students because it was assumed that they already felt efficacious about their abilities in science inquiry. Low-achieving students, on the other hand, were expected to find this treatment to be beneficial. As inquiry scores increased and as students reflected on this, it was predicted that self-efficacy to successfully complete science inquiry tasks would increase as well.

Phase III – Continuation of Treatment Conditions

Phase III took place approximately 26 weeks into the school year. The phase began with the students taking the third iteration of the Chemistry Self-Efficacy Survey. The treatment conditions by class were identical to the treatment conditions that took place in Phase II. The goal was to replicate the experimental conditions in order to give students another experience with the guided inquiry cycle. In this phase of the study, the focus was on the chemistry of acids and bases.

The unit for this phase of the study began with an investigation using purple cabbage juice as an indicator to develop a definition of what types of things would be considered acidic and then by contrast, what types of things would be considered basic. While no time was spent reviewing the State of Oregon scoring guide, students in the treatment conditions were again asked to used the sentence stems from this study to write reflective learning summaries at various points in time during the unit to ensure that the students were still familiar with these stems. The students then completed a lab where they standardized a 0.1M sodium hydroxide solution ($NaOH$) with potassium hydrogen phthalate (KHP). The final lab of the unit asked students to titrate an 81 milligram aspirin tablet and a 325 milligram aspirin tablet and compare their findings with the stated values. Using their knowledge of acids and bases, the students were asked come up with a question or problem that could be addressed using titration and then carry out the investigation. Some of the investigation topics included things such as: a comparison of the acidity of various carbonated beverages (ie. diet versus regular), vitamin C content of various orange juice brands, and titrations of aspirin to compare milligrams of aspirin in generic and name-brand samples. When students struggled with developing a testable question or problem, they were encouraged to talk with other groups in order to help direct their thinking. The instructional treatments that took place in this phase were identical to the instructional treatments that took place in phase II.

This third inquiry task was much more student directed than the previous two tasks of this study. While the students were directed to focus on a question that could be addressed using and acid / base titration, they were free to choose what they wanted within those parameters. Once they chose their question, the students then had to develop their procedure to collect data such as what concentrations of their samples to use in order to complete their titrations in a timely manner. Similar to the second inquiry task, student received very little assistance in interpreting and presenting their data beyond the help they received in the treatment conditions.

Phase IV – Final Assessment

At the conclusion of the final inquiry task, all of the students were asked to respond to a final iteration of the Chemistry Self-Efficacy Survey (CSES). This survey was given approximately 34 weeks into the school year. At this point in the school year, all of the new content had been introduced, so the students would have had the opportunity to become acquainted with all of the content on the CSES. In addition, the final inquiry task of the year had already been returned, so the students had been able to look at their scores and process these as well.

Data Analysis - Addressing the Research Questions

Question 1

What is the relationship between self-efficacy for science inquiry / chemistry content and performance on inquiry tasks?

Feedback from the CSES was used to address this question. Specifically related to student self-efficacy for inquiry, it was anticipated that there would be a homogenous distribution of student self-efficacy scores across all of the classes in this study for the initial administration of the CSES. This assumption was tested using a one-way ANOVA. Students in Classes A and B were not expected to see any changes in their self-efficacy related to science inquiry over the course of the study. In contrast, it was anticipated that students in Classes C and D would benefit from the treatment and a corresponding rise in self-efficacy scores related to science inquiry was expected. In order to compare self-efficacy scores by experimental group over time, a mixed between-within ANOVA was conducted.

In addition, it was anticipated that there would be a strong correlation between a student's sense of efficacy and their score on the science inquiry task. This correlation was assessed using Pearson's product-moment correlation and was assessed independently for each separate inquiry task in this study.

The questions related to Chemistry Content on the CSES were also used to assess changes in student self-efficacy related to content. At each phase of this study, student scores were compared both to other classes and also to previous administrations of the survey. It was anticipated that there would not be any differences between the classes in this study because all of the classes were receiving the same instruction. In addition, it was expected that student self-efficacy scores for Chemical Content would increase over time as the students learned new material. To compare scores by experimental group over time, a mixed between-within ANOVA was used.

Question 2

Do metacognitive prompts given prior to, throughout, and following the science inquiry processes improve student performance on inquiry tasks?

This was assessed by looking at student scores on the three inquiry tasks outlined in this study. It was anticipated that all of the classes would have similar mean inquiry scores for the initial inquiry task. A one-way ANOVA was used to test this assumption. It was also predicted that Classes A and B would not see any increase in their science inquiry scores over phases II & III. In contrast, Classes C and D were expected to see an increase in their science inquiry scores as they moved through phases II & III. To compare the effects of metacognitive prompts by experimental group over time, a mixed between-within ANOVA was conducted.

A similar analysis was also conducted to assess the impact of these metacognitive prompts on the inquiry project scores of high- and low-achieving students. It was expected that high achieving students would not benefit much from the metacognitive prompts because it was predicted that high achieving students were high achieving because they already have developed effective metacognitive reflection skills. It was predicted that high achieving students would have high science inquiry scores and high self-efficacy scores for science inquiry. It was predicted that neither of these scores for the high achieving students would change much over time. In contrast, it was expected that low achieving students would benefit from these metacognitive prompts and sentence stems and resulting in an increase in their science inquiry task scores and their self-efficacy for science inquiry scores. A mixed between-within ANOVA was used to look at change with high and low achievers by experimental group over time. In addition, a mixed between-within ANOVA was used to compare the effects of metacognitive prompting types (generic versus directed) over time.

Question 3

What impact does the use of sentence stems have on a student's ability to successfully communicate their ideas related to a science inquiry tasks?

It was anticipated that students who used sentence stems would find that their science inquiry scores would be higher than students who did not use the sentence stems. Student results were compared within a phase using a two-tailed t-test. To compare results over time, student science inquiry scores were compared with a mixed between-within ANOVA.

Similar to the metacognitive prompts, an analysis was conducted to determine the impact of the sentence stems on the inquiry project scores of the high- and low-achieving students. It was anticipated that the sentence stems would provide a greater benefit to the low achieving students and have little effect on the high achieving students. Again, the reason for this was that it was anticipated that high-achieving students already possessed schema for effective writing and so would not need the prompting provided by the sentence stems. In contrast, the low-achieving students did not possess this same schema for elaborating on ideas and so the use of the sentence stems should result in an increase in the science inquiry task scores for the low achieving students. Science inquiry tasks scores were compared by experimental group using a mixed between-within ANOVA.

Table 9 provides an overview of the statistics utilized in this study. Of specific interest to this study were the trends and patterns that emerged from the data. Word count emerged as another variable that seemed to correlate with student scores. The relationship between word counts and science inquiry tasks scores was pursued in addition to the questions outlined above.

Table 9: Research Questions and Statistical Procedures used.

	Tests within a phase	Tests between phases
Question 1: Science Inquiry Self-efficacy scores		Mixed between-within subjects ANOVA
Question 1: Science Inquiry Self-Efficacy Scores versus actual scores from Science Inquiry tasks	Pearson product-moment correlation	
Question 1: Chemistry Content Self-Efficacy scores		Mixed between-within subjects ANOVA
Question 2: Effects of metacognitive prompting on Science Inquiry tasks scores	One-way ANOVA	Mixed between-within subjects ANOVA
Question 3: Impact of Sentence Stems on Science Inquiry task scores.	Two-tailed t-test	Mixed between-within subjects ANOVA
Question 3: Benefits of sentence stems on Science Inquiry task scores by achievement		Mixed between-within subjects ANOVA

Chapter 4: Data & Analysis

Introduction

This chapter examines the extent to which the two experimental conditions, metacognitive prompts and sentence stems, made a difference in improving students' abilities to conduct an inquiry task and then communicate their findings in a written format. To do this, the data analysis is in this chapter sought to address four main questions. The first three sections in this chapter correspond with the three research questions in this study:

1. What is the relationship between self-efficacy for science inquiry / chemistry content and performance on inquiry tasks?
2. Do metacognitive prompts given prior to, throughout, and following the science inquiry processes improve student performance on inquiry tasks?
3. What impact does the use of sentence stems have on a student's ability to successfully communicate their ideas related to science inquiry tasks?

The fourth question investigated in this chapter was not included as one of the original research questions as it emerged as a topic of interest during the course of the study. This fourth section sought to determine if there was a connection between the number of words that a student used and the resulting score they received. The chapter concludes with a summary of the significant study results.

Self-Efficacy and Inquiry Task Performance

Chemistry Self-Efficacy Survey (CSES) Psychometrics

The purpose of the survey was to track changes in student self-efficacy over time as they were involved in the study. Students in the study were asked to complete the Chemistry Self-Efficacy Survey (CSES) at four different points in time. There was a gap of at least twelve weeks between each administration.

The CSES was composed of two scales. One scale was composed of 10 items and assessed student self-efficacy related to chemistry content (Chemistry Content Self-Efficacy Scale) while the second scale was composed of 11 items and assessed student self-efficacy related to writing a science inquiry task (Science Inquiry Self-Efficacy Scale). Both scales in the instrument had good internal consistency with Cronbach alpha coefficients exceeding the recommended value of 0.7 (DeVellis 2003). For the Chemistry Content Self-Efficacy Scale, Cronbach alpha coefficients ranged from 0.77 to 0.92 with the internal consistency increasing with each administration. In a similar manner, the Cronbach alpha coefficients for the Science Inquiry Self-Efficacy Scale ranged from 0.85 to 0.95 and also increased with each administration. For each administration of the survey, all 21 survey items were subjected to a principal component analysis (PCA). Prior to performing each principal component analysis, the data from the survey was analyzed to assess its suitability.

For the initial administration of the survey, the Kaiser-Meyer-Oklin value for the CSES was 0.52 which was slightly lower than the recommended value of 0.6 (Kaiser, 1970 & 1974). Bartlett's Test of Sphericity, though, (Bartlett, 1954) did reach

Table 10: Pattern and structure matrix for PCA with oblmin rotation of a two factor solution of CSES items from the initial administration.

Item	Pattern Coefficients		Structure Coefficients		Communalities
	Component 1	Component 2	Component 1	Component 2	
Question 1	**.465**	.224	**.529**	.356	0.326
Question 2	-.168	**.620**	.009	**.573**	0.354
Question 3	**.373**	.166	**.421**	.272	0.202
Question 4	**.317**	.294	**.401**	.384	0.240
Question 5	**.644**	-.309	**.556**	-.125	0.397
Question 6	.270	**.434**	.393	**.511**	0.328
Question 7	**.837**	-.147	**.795**	.091	0.653
Question 8	**.606**	.040	**.618**	.212	0.383
Question 9	.029	**.661**	.217	**.669**	0.449
Question 10	**.557**	.115	**.590**	.274	0.360
Question 11	**.595**	.147	**.637**	.316	0.425
Question 12	**.486**	.165	**.533**	.304	0.310
Question 13	.370	**.422**	.490	**.527**	0.404
Question 14	-.015	**.784**	.208	**.779**	0.608
Question 15	.260	**.265**	.335	**.339**	0.177
Question 16	.020	**.645**	.203	**.651**	0.424
Question 17	**.789**	-.177	**.738**	.048	0.574
Question 18	**.735**	-.098	**.707**	.111	0.509
Question 19	.008	**.689**	.204	**.691**	0.478
Question 20	**.835**	-.115	**.802**	.122	0.655
Question 21	**.640**	.198	**.696**	.380	0.521

Note: Major loadings for each item are in bold.
The Chemical Content Scale contained questions: 1, 2, 6, 9, 10, 12, 14, 15, 16, 19
The Science Inquiry Scale contained questions: 3, 4, 5, 7, 8, 11, 13, 17, 18, 20, 21

statistical significance ($p < 0.0005$). An initial scree plot indicated that a two or three component solution would explain the greatest amount of the variance. Since the survey was designed to contain two factors, a two component solution was chosen. To aid in the interpretation of these two components, an oblimin rotation was performed.

A two component solution with an oblimin rotation explained 41.8% of the variance with the Science Inquiry Self-Efficacy Scale contributing 29.7% and the Chemistry Content Self-Efficacy Scale contributing 12.1%. The factors were only weakly correlated ($r = 0.29$). Ten of the eleven Science Inquiry items loaded on the first factor. Question 13 was the only item that loaded higher on the second factor. Both pattern coefficients and structure coefficients showed, though, that question 13 actually loaded well on either factor. Only seven of the ten Chemistry Content items loaded on the second factor. In looking at the Chemistry Content items that did not load as well on the second factor, three questions (questions 1, 10, & 12) covered content that would have been covered in a freshman science course and these three questions loaded well on the Science Inquiry scale. In this initial survey, students generally felt more confident about their abilities in inquiry than in chemistry content and so these content items loading on the inquiry scale was not surprising. Because these three questions were specifically related to content, they were left in the Chemistry Content scale.

For the second administration of the survey, again all the items on the survey were subjected to a principal component analysis (PCA). The Kaiser-Meyer-Oklin (KMO) value was 0.82 which was higher than the initial administration (KMO = 0.52) and exceeded the recommended value of 0.6 (Kaiser, 1970 & 1974). Bartlett's Test of Sphericity (Bartlett, 1954) reached significance ($p < 0.0005$) and an initial analysis of the scree plot indicated that almost all of the survey variance could be explained by a two component solution. To aid in the interpretation of these two components, an oblimin rotation was performed.

Table 11: Pattern and structure matrix for PCA with oblmin rotation of a two factor solution of CSES items from the second administration

Item	Pattern Coefficients		Structure Coefficients		Communalities
	Component 1	Component 2	Component 1	Component 2	
Question 1	**.732**	.101	**.768**	.368	.599
Question 2	-.229	**.603**	-.010	**.519**	.315
Question 3	**.831**	-.178	**.766**	.125	.614
Question 4	**.452**	.299	**.561**	.464	.392
Question 5	**.824**	-.169	**.763**	.132	.606
Question 6	**.610**	.181	**.676**	.403	.486
Question 7	**.837**	-.173	**.774**	.132	.624
Question 8	**.575**	.216	**.654**	.426	.468
Question 9	**.446**	.409	**.595**	.572	.500
Question 10	.223	**.617**	.448	**.698**	.531
Question 11	.307	**.430**	.464	**.542**	.375
Question 12	.135	**.652**	.373	**.702**	.508
Question 13	**.627**	.152	**.682**	.380	.486
Question 14	-.015	**.645**	.220	**.640**	.409
Question 15	**.566**	.289	**.671**	.495	.522
Question 16	.133	**.472**	.305	**.521**	.286
Question 17	**.515**	.340	**.639**	.528	.508
Question 18	**.831**	-.135	**.782**	.168	.627
Question 19	**.519**	.247	**.609**	.436	.424
Question 20	**.854**	-.167	**.794**	.145	.654
Question 21	**.562**	.230	**.646**	.434	.463

Note: Major loadings for each item are in bold.
The Chemical Content Scale contained questions: 1, 2, 6, 9, 10, 12, 14, 15, 16, 19
The Science Inquiry Scale contained questions: 3, 4, 5, 7, 8, 11, 13, 17, 18, 20, 21

A two-component solution explained 49.5% of the variance with the Science Inquiry Self-Efficacy scale contributing 40.4% and the Chemical Content Self-Efficacy scale contributing 9.1%. These two factors were only weakly correlated ($r = 0.36$). With this second administration, the two factor solution seemed to do a better job of explaining the variance (the initial administration two factor solution explained 41.8 % of the variance), although, the factor separation was not as clean. Though many of the items loaded well on both factors, ten of the eleven items on the Science Inquiry Self-Efficacy scale had their highest loadings on component 1. The only exception was item 11 which asked students about their beliefs in their ability to interpret data. On this administration of the CSES, Science Inquiry Self-Efficacy scores were generally higher than those for Chemical Content Self-Efficacy. That item 11 loaded better on the Chemical Content scale implied that students did not feel as efficacious about their abilities to interpret data when compared to other parts of the inquiry task.

On the Chemical Content Self-Efficacy Scale, only six out of the ten items loaded best on component 2. Three of the items in question (questions 6, 15, & 19) asked students about content that had just been covered. Because of the proximity between when this content was taught and the administration of this survey, it was not surprising that the scores on these items were slightly higher and therefore loaded better on the Science Inquiry scale. In looking at both the pattern and the structure coefficients, these items actually loaded well on both scales. The other item that did not load as well on the Chemistry Content scale was a question (question 1) that asked students about content that many had covered repeatedly since middle school (balancing chemical equations). That this item loaded better on the Science Inquiry scale was also not surprising. Again, in looking at the pattern and structure coefficients, even though this question loaded much better on the Science Inquiry factor, it also loaded well on the Chemical Content factor. Based on this analysis, no changes were made to the instrument.

A similar analysis was completed with the survey data from the third administration of the survey. An inspection of the correlation matrix revealed that many coefficients were 0.3 and above. The Kaiser-Meyer-Oklin value was 0.80 which exceeded the recommended value of 0.6 (Kaiser, 1970 & 1974). Bartlett's Test of Sphericity (Bartlett, 1954) reached statistical significance ($p < 0.0005$) which provided support for the existence of factors in the correlation matrix. An analysis of the screeplot indicated that a two factor solution would explain a majority of the variance. To aid in the interpretation of these two components, an oblimin rotation was performed.

A two-component solution explained a total of 57.4% of the variance. The items identified with the Science Inquiry Self-Efficacy scale explained 49.1% of the variance, and the items identified with the Chemical Content Self-Efficacy scale explained an addition 8.3 % of the variance. While this iteration of the CSES was able to explain a greater percentage of the variance, in this third iteration, the two scales were also more highly correlated ($r = 0.61$) than previous iterations (1st administration: $r = 0.29$, 2nd administration: $r = 0.36$).

Table 12: Pattern and structure matrix for PCA with oblmin rotation of a two factor solution of CSES items from the third administration

Item	Pattern Coefficients		Structure Coefficients		Communalities
	Component 1	Component 2	Component 1	Component 2	
Question 1	.004	**.882**	.541	**.885**	.782
Question 2	-.158	**.918**	.400	**.821**	.690
Question 3	**.761**	.038	**.784**	.502	.616
Question 4	.322	**.358**	.540	**.554**	.372
Question 5	**.662**	.128	**.740**	.531	.558
Question 6	.068	**.616**	.442	**.657**	.434
Question 7	**.800**	-.070	**.757**	.417	.576
Question 8	.318	**.594**	.679	**.788**	.684
Question 9	**.490**	.000	**.490**	.298	.240
Question 10	**.502**	.289	**.678**	.595	.512
Question 11	**.537**	.275	**.704**	.602	.544
Question 12	.241	**.636**	.629	**.783**	.650
Question 13	**.549**	.258	**.706**	.593	.541
Question 14	-.065	**.847**	.451	**.807**	.654
Question 15	**.475**	.230	**.615**	.519	.411
Question 16	.016	**.577**	.367	**.587**	.344
Question 17	**.813**	.017	**.823**	.512	.678
Question 18	**.844**	-.169	**.741**	.345	.567
Question 19	.240	**.715**	.675	**.861**	.777
Question 20	**.815**	.019	**.827**	.516	.684
Question 21	**.888**	-.053	**.856**	.488	.734

Note: Major loadings for each item are in bold.
The Chemical Content Scale contained questions: 1, 2, 6, 9, 10, 12, 14, 15, 16, 19
The Science Inquiry Scale contained questions: 3, 4, 5, 7, 8, 11, 13, 17, 18, 20, 21

Item loadings were analyzed to ensure that items from the survey loaded on the correct scale. The Science Inquiry Self-Efficacy scale was composed of eleven items. Nine of these items loaded best on this scale. The two remaining items in this scale, questions 4 & 8, actually loaded well on both scales. Question 4 asked students how they felt about their ability to represent data in graphical form and question 8 asked students how they felt about developing a science investigation question. Student self-efficacy for Chemical Content (M = 4.8, SD = 1.2) was slightly lower than student self-efficacy for Science Inquiry (M = 5.0, SD = 1.2). The implication was that students generally felt less efficacious in relation to these two Science Inquiry skills.

On the Chemical Content scale, seven of the ten items had their best loading on this factor. The three items that did not load as well were questions 9, 10, and 15. These three questions addressed content that was covered at various points during the year. Question 9 dealt with electrons and bonding, question 10 dealt with patterns in the periodic table, and question 15 covered stoichiometry. It was difficult to say why these questions may have loaded better on the Science Inquiry scale. Since these questions clearly cover chemistry content, the choice was made to keep these items in the scale.

For the fourth administration of the survey, as with the prior three administrations, an analysis was conducted to assess the psychometrics of the instrument. An inspection of the correlation matrix revealed that many coefficients were 0.3 and above, and the Kaiser-Meyer-Oklin value was 0.84 which exceeded the recommended value of 0.6 (Kaiser, 1970 & 1974). Bartlett's Test of Sphericity (Bartlett, 1954) reached statistical significance ($p < 0.0005$) which provided support for the existence of factors in the correlation matrix. An analysis of the screeplot indicated that a two factor solution would explain a majority of the variance. To aid in the interpretation of these two components, an oblimin rotation was performed.

A two-component solution explained 64.7% of the variance. The items identified with the Science Inquiry Self-Efficacy scale explained 55.5% of the variance. In contrast, items identified with the Chemical Content Self-Efficacy scale only explained 9.2% of the variance. Each successive iteration of the CSES was able to explain a greater percentage of the variance (% variance explained by 1st administration: 41.8%, % variance explained by 2nd administration: 49.5%, % variance explained by the 3rd administration: 54.7%). There was a strong positive correlation between the two scales ($r = 0.58$). Although stronger than the first two administrations (1st administration: $r = 0.29$, 2nd administration: $r = 0.36$), this correlation was similar to the 3rd administration of the survey ($r = 0.61$).

Item loadings were then assessed to ensure that survey items loaded on the correct scale (see table 13). For the Science Inquiry Scale, ten of the eleven items loaded best on this factor. The only exception was question 8. This question asked students how competent they felt about their ability to develop a scientific question. Student self-efficacy for Chemical Content ($M = 5.0$, $SD = 1.2$) was slightly lower than student self-efficacy for Science Inquiry ($M = 5.4$, $SD = 1.2$). The implication was that students did not feel as efficacious about their abilities to construct a scientific question when compared to the other skills involved in Scientific Inquiry. This was similar to prior administration of this survey, so the decision was made to keep this question in the scale. On the Chemical content scale, all ten items loaded as expected, so no changes were made.

Table 13: Pattern and structure matrix for PCA with oblmin rotation of a two factor solution of CSES items from the fourth administration

Item	Pattern Coefficients		Structure Coefficients		Communalities
	Component 1	Component 2	Component 1	Component 2	
Question 1	.144	**.641**	.515	**.724**	.538
Question 2	.333	**.435**	.585	**.628**	.468
Question 3	**.734**	.134	**.811**	.558	.670
Question 4	**.507**	.209	**.628**	.502	.423
Question 5	**.932**	-.222	**.803**	.317	.678
Question 6	-.013	**.774**	.434	**.766**	.587
Question 7	**.801**	.039	**.824**	.503	.680
Question 8	.296	**.548**	.613	**.719**	.576
Question 9	-.299	**.897**	.220	**.724**	.584
Question 10	.289	**.582**	.626	**.750**	.618
Question 11	**.609**	.281	**.771**	.633	.647
Question 12	.141	**.711**	.552	**.792**	.641
Question 13	**.636**	.352	**.839**	.720	.787
Question 14	.259	**.612**	.613	**.762**	.625
Question 15	.089	**.723**	.507	**.775**	.605
Question 16	.090	**.762**	.530	**.813**	.667
Question 17	**.887**	.086	**.937**	.599	.883
Question 18	**.814**	.052	**.844**	.523	.714
Question 19	.129	**.748**	.562	**.823**	.688
Question 20	**.841**	.017	**.851**	.503	.724
Question 21	**.855**	.052	**.885**	.547	.786

Note: Major loadings for each item are in bold.
The Chemical Content Scale contained questions: 1, 2, 6, 9, 10, 12, 14, 15, 16, 19
The Science Inquiry Scale contained questions: 3, 4, 5, 7, 8, 11, 13, 17, 18, 20, 21

Taken together, the data from the four administrations of the survey indicated that the CSES was psychometrically sound. The survey was designed to contain two factors and these two factors were consistently derived from the analysis of student responses. In addition, the two factors had high internal consistency with Cronbach's

α values ranging from 0.77 to 0.95. These data provided confidence that the CSES did indeed measure what it was designed to measure.

Chemistry Content Self-Efficacy: Comparison Across Administrations

All three of the inquiry tasks in this study were designed to allow students to use their current chemistry knowledge in order to apply it to a specific question. To that end, the CSES was used to monitor how efficacious students felt about their understandings of chemistry content. This was to ensure that student scores on their inquiry tasks were reflective of their abilities to do inquiry rather than on their abilities to understand chemistry content. A mixed between-within subjects ANOVA was conducted to determine if there were any differences in self-efficacy for chemistry content between the experimental groups (control, experimental group without reflective prompts, experimental group with directed reflective prompts, experimental group with generic reflective prompts) over time. The main effect for group membership was not significant, ($F(3,43) = 2.49$, $p = 0.07$, partial eta squared = 0.15), so the student scores for self-efficacy in Chemistry Content across all four classes were analyzed as one large group.

The trend in the data indicated that students generally felt more efficacious about their abilities in chemistry as the year progressed (see figure 12). A one-way repeated measures ANOVA was conducted to compare the Chemistry Content CSES scores over time. The effect for time was significant (Wilks' Lambda = 0.51, $F(3,44) = 14.17$, $p < 0.0005$). Post-hoc comparisons using a Bonferroni adjustment indicated that the mean Chemistry Content Self-efficacy scores from the initial administration of the survey were significantly lower than the subsequent three administrations ($p < 0.0005$). The increase in mean self-efficacy scores for Chemistry Content over the final three administration of the survey was not significant.

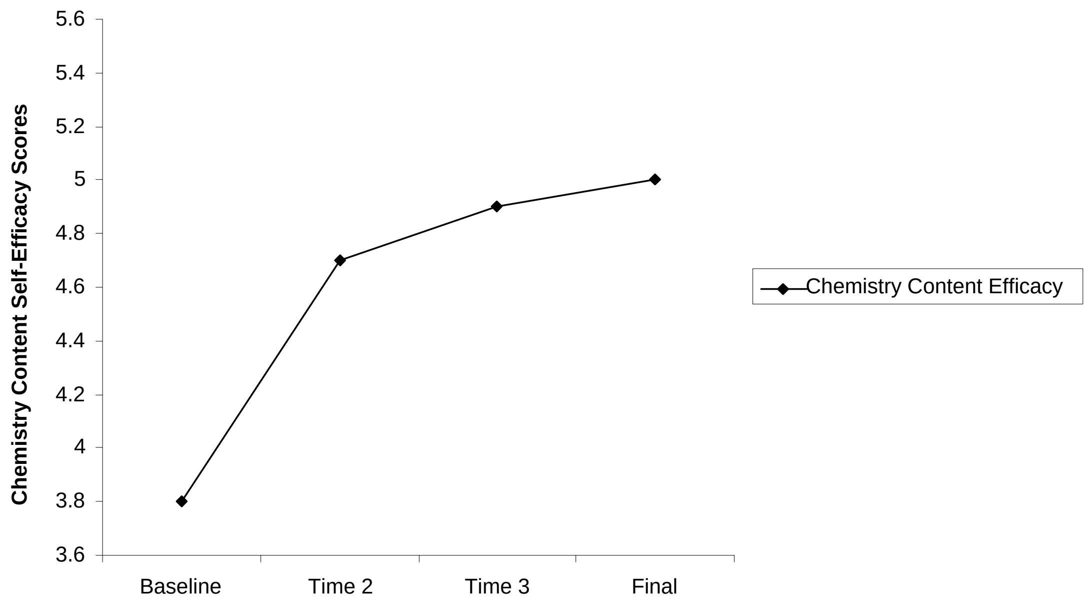


Figure 12: Mean Chemistry Content Self-Efficacy scores over time

Although student self-efficacy scores for content remained fairly stable over the final three administrations of the survey, a look at student responses to individual questions revealed interesting changes to student learning over time. Table 14 shows how student responses changed with each administration of the survey. The questions in table 14 are ordered by the sequence that they were taught over the course of the school year. For example, question #6 was taught at the beginning of the year whereas question #16 was taught towards the end of the school year. At the beginning of the year, the highest self-efficacy scores were to the left of the table. As the school year progressed, the items that received the highest score shifted towards the right of the table. The implication was that even though student self-efficacy for chemistry content remained stable over the latter part of the study, what the students felt efficacious about did change. Student efficacy tended to center around the content that was most recently taught.

Table 14: Comparison of Chemistry Self-Efficacy question means by order taught and administration.

	#6	#12	#10	#9	#1	#15	#19	#14	#2	#16
Initial Administration	5.3	5.5	5.4	2.4	4.4	3.6	2.6	2.9	2.9	3.1
2nd Administration	4.8	6.1	5.8	3.8	5.3	4.5	4.6	3.8	2.6	4.7
3rd Administration	4.5	5.8	5.3	3.5	5.2	4.1	4.9	5.7	5.1	4.1
4th Administration	4.3	5.7	5.6	4.2	5.5	4.6	5.0	5.4	4.9	5.0

A mixed between-within subjects ANOVA was conducted to see if male or female students had any advantage in relation to self-efficacy related to Chemistry Content. The effect for gender was not significant ($F(1,45) = 0.16$, $p = 0.70$). A similar test was conducted to compare high achieving and low achieving students. The main effect for time was significant (Wilks' Lambda = 0.53, $F(3,43) = 12.90$, $p < 0.0005$, partial eta squared = 0.47) which indicated that mean self-efficacy for Chemistry Content increased over time for both groups. The main effect for achievement was also significant ($F(1,45) = 7.92$, $p = 0.007$, partial eta squared =

0.15) which indicated that high achieving students tended to feel more efficacious about their abilities in chemistry than did their lower achieving counterparts. That high achieving students felt more efficacious about their abilities in chemistry was not surprising. It was expected that students who had demonstrated a greater understanding of the content would also have greater self-efficacy in relation to the content. These changing scores also indicated that student beliefs were slowly changing as they were involved in this study.

Science Inquiry Self-Efficacy: Comparison Across Administrations

In contrast to the Chemistry Content self-efficacy data, the results from the CSES indicated that student self-efficacy related to their abilities in science inquiry was quite stable across the time period of this study. Across all four administrations of the survey, there were no significant differences in mean self-efficacy scores for Science Inquiry by experimental group ($F(3,43) = 1.31$, $p = 0.28$, partial eta squared = 0.08). Because of this, student results for Science Inquiry Self-Efficacy were assessed as one group.

Over the course of the study, mean self-inquiry scores for science inquiry dropped and then rebounded at the end of study. A one-way repeated measures ANOVA indicated that even though mean scores for science inquiry self-efficacy changed over time, the main effect for time was not significant (Wilks' Lambda = 0.84, $F(3,44) = 2.78$, $p = 0.06$, partial eta squared = 0.16). The mean self-efficacy scores for science inquiry tended to be between 5.0 and 5.5 on a 7-point scale. The implication was that students generally felt efficacious about their abilities in science inquiry.

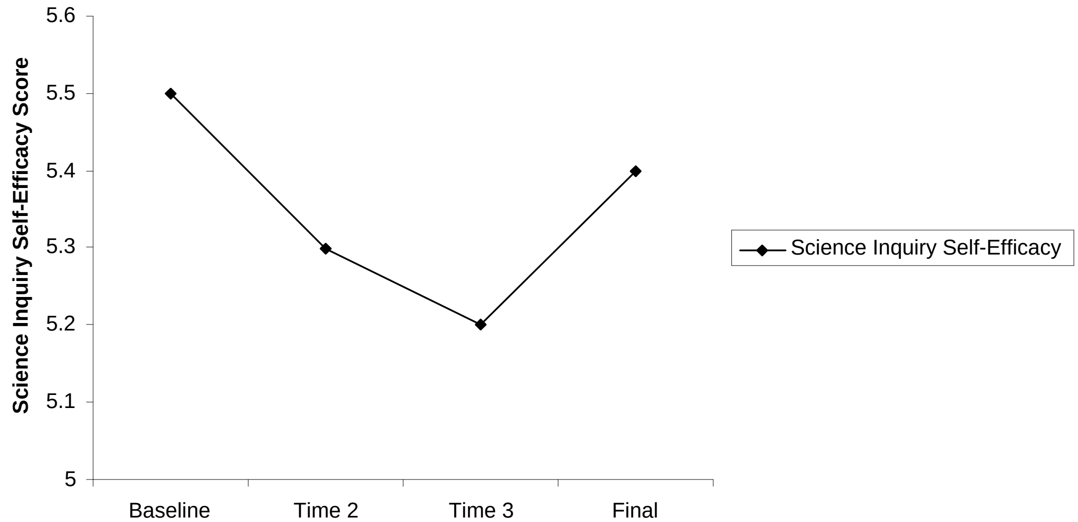


Figure 13: Mean Science Inquiry Self-Efficacy scores over time

Gender was evaluated to see if this played any role in science inquiry self-efficacy. There were no differences by gender over time (F(1,45) = 0.05, p = 0.83). This implied that both genders felt equally efficacious about their abilities in science inquiry.

A similar test was run to determine if there were differences in science inquiry self-efficacy by prior achievement. The main effect between achievement and time was not significant (Wilks' Lambda = 0.96, F(3,43) = 0.57, p = 0.64) whereas the main effect for time was significant (Wilks' Lambda = 0.84, F(3,43) = 2.84, p = 0.05, partial eta squared = 0.17) Interestingly, a follow-up comparison using a Bonferroni adjustment did not find any significant differences between the various survey administrations. With this in mind, the assumption was made that the differences that were seen over time were not significant and that mean science inquiry self-efficacy scores remain unchanged over the course of the study for the two achievement groups. The main effect for achievement was moderate and significant ($F(1,45) = 5.11$, $p = 0.03$, partial eta squared = 0.10). On average, high achievers mean Science Inquiry self-efficacy scores were 0.6 points higher than their lower achieving counterparts. That high achieving students had greater science inquiry self-efficacy was not surprising. Even though there was a significant difference in self-efficacy scores for high and low achieving students, this gap did not really change over the course of this study. These results indicated that for the students in this study, their self-efficacy for science inquiry was quite resistant to change.

When student Science Inquiry self-efficacy scores were broken down by inquiry category, an interesting pattern emerged. Across all four administrations of the survey, students consistently indicated that they felt less efficacious about their abilities in Framing and Analyzing and more efficacious about their abilities in Designing and Collecting (see table 15). In three out of four survey administrations, students felt significantly more efficacious about their abilities in Collecting data and felt significantly less efficacious about their abilities in Analyzing their data (see table 16). In a similar way, on two of the survey administrations, students felt significantly more efficacious about the abilities in Collecting data than they did about their abilities in Framing. Also, on all four administrations of the CSES, there was no significant difference in student self-efficacy for science inquiry between the sections of Framing and Analyzing (see table 16). These two sections required the greatest amount of expository writing and this demand may have contributed to students' lower self-efficacy in these areas. In contrast, the Collecting section was mostly composed of data tables, graphs, and sample calculations. The Collecting section required very little expository writing. This evidence provided support for the conjecture that the writing component of the science inquiry task posed a significant challenge for many students.

Table 15: Mean Science Inquiry Self-Efficacy scores by administration

	Administration 1	**Administration 2**	**Administration 3**	**Administration 4**
Framing	5.3	5.2	5.0	5.3
Designing	5.6	5.4	5.3	5.5
Collecting	6.0	5.5	5.3	5.7
Analyzing	5.3	5.1	5.0	5.4

Table 16: Comparison of mean Science Inquiry Self-Efficacy scores within each administration

	Framing	**Designing**	**Collecting**	**Analyzing**
Framing			1***, 4**	
Designing			1**	
Collecting				1***, 2**, 4**
Analyzing				

Number = mean difference between sections was significant during that particular administration
** significant at $p < 0.005$ / *** significant at $p < 0.0005$

Even though students felt less efficacious about the expository writing sections of the inquiry task, when chemistry content self-efficacy and science inquiry self-efficacy scores were compared, the results indicated that students generally felt more efficacious about their abilities in science inquiry as compared to their abilities in chemistry content. A series of paired-samples t-tests were conducted to compare chemistry content self-efficacy scores and Science Inquiry self-efficacy scores. For all administrations of the survey (1^{st} ($t(56) = 13.43$, $p < 0.0005$ (two-tailed), 2^{nd} ($t(56) = 5.58$, $p < 0.0005$ (two-tailed), 3^{rd} ($t(56) = 2.00$, $p = 0.05$ (two-tailed), and 4^{th} ($t(57) = 3.81$, $p < 0.0005$ (two-tailed)), student self-efficacy for science inquiry was significantly greater than student self-efficacy for chemistry content. In this study, students felt more confident about their abilities to successfully complete an inquiry task than they did about their abilities to successfully complete chemistry problems.

Science Inquiry Self-Efficacy: Correlation Analysis

The relationship between student self-efficacy and actual student scores on the three inquiry tasks was investigated using Pearson product-moment correlation coefficient. The correlation was positive and moderate to strong with significant r values ranging from 0.32 to 0.68 and associated p-values ranging from 0.02 to < 0.0005 (see table 17). Two conclusions emerged from this data. First, the correlation between science inquiry self-efficacy and actual inquiry project score provided more confirmation that the CSES was measuring what it was intended to measure. Second, the data provided support for the cyclical model of student self-efficacy which stated that there was a circular and reciprocal relationship between a student's sense of self-efficacy for science inquiry and their actual practice. These results will be discussed in more detail in the following chapter.

Table 17: Pearson product-moment correlations between self-efficacy for science inquiry and actual science inquiry scores

	Inquiry 1 – Score	Inquiry 2 – Score	Inquiry 3 - Score
CSES 1	0.35*	0.40**	0.24
CSES 2	0.32*	0.38*	0.21
CSES 3	0.53**	0.68**	0.36**
CSES 4	0.46**	0.51**	0.47**

* Correlation significant at the 0.05 level (2-tailed)
** Correlation significant at the 0.01 level (2-tailed)

The Effects of Metacognitive Prompting

Effects across Inquiry Tasks

Only two out of the four classes in this study were given reflective prompts as they worked both on the data collection and also on the post-inquiry write-up. Before the students began experimenting or reporting, they were asked to choose a goal for that particular part of the investigation. This was followed by two reflective prompts that the students were asked to complete as they were working.

To determine the effect of these reflective prompts, an analysis was done on the data from the first inquiry task in this study. All of the students, regardless of experimental condition, received the same instruction for this initial inquiry task. This was done to compare science inquiry task scores between classes to determine if there were any differences in range of abilities present in each experimental group. A one-way between-groups ANOVA was conducted to test for any differences. The analysis revealed that there was no significant difference in project scores between the groups ($F(3,54) = 0.42$, $p = 0.74$) for the initial inquiry task. To see if there was any effect over time by the treatment received in the experimental group, a mixed between-within subjects ANOVA was conducted. The interaction between time and reflection was not significant (Wilks' Lambda = 0.99, $F(2,52) = 0.23$, $p = 0.80$) whereas the main effect for time was significant (Wilks' Lambda = 0.74, $F(2,52) = 8.95$, $p < 0.0005$, partial eta squared = 0.26). In this study, mean student inquiry scores went from 12.5 to 13.9 out of a possible total of 24. As indicated by the significance of the main effect for time, this increase in student mean inquiry score was significant. There was no significant difference in mean inquiry score between the students who were given metacognitive prompts and those who were not ($F(1,53) = 0.00$, $p = 0.99$). For this study, it appeared that metacognitive prompts had no effect on students' abilities on to successfully complete a science inquiry task.

Metacognitive Prompts and Prior Achievement

An interesting trend emerged when the data was broken down by achievement. For this particular comparison, the students were separated based on whether or not they were given metacognitive prompts and also by achievement level. While mean science inquiry scores generally increased over time, it appeared that both prior achievement and metacognitive prompting had an effect (see figure 14). A mixed between-within subjects ANOVA was conducted to assess the impact of these metacognitivie reflective prompts on student inquiry scores when the results were separated by prior achievement. The interaction between group membership and time was not significant (Wilks' Lambda = 0.91, $F(6,100) = 0.81$, $p = 0.56$). In contrast, the main effect for time was significant (Wilks' Lambda = 0.75, $F(2,50) = 8.35$, $p = 0.001$, partial eta squared = 0.25) which indicated that the change observed in inquiry scores over time was indeed significant. The main effect for group membership was also significant ($F(3,51) = 7.15$, $p < 0.0005$, partial eta squared = 0.30). A series of post-hoc test were conducted using a Bonferroni adjustment. The only significant difference that was found was between the students with low prior achievement in the metacognitive reflection group and the students with high prior achievement in both the control group ($p = 0.007$) and the metacognitive reflection group ($p = 0.001$).

The metacognitive prompts that were used in this study did not assist the progress of the students with low prior achievement. This finding was opposite of what was expected. It was predicted that the metacognitive reflective prompts would help scaffold the thinking of these low prior achieving students so that their thinking would model that of their higher achieving peers. This help in modeling reflective thinking was predicted to result in higher science inquiry task scores. This point will be discussed further in the next chapter.

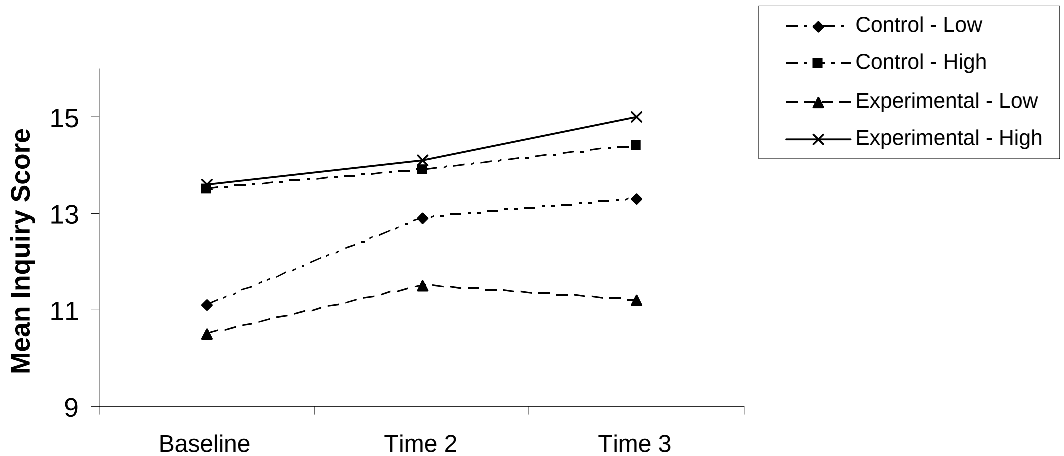


Figure 14: Comparison of mean student inquiry score by reflection type and also by prior student achievement

Generic versus Directed Metacognitive Reflective Prompts

Following the work of White & Frederiksen (1998) and Davis (2003), this study sought to determine if there was any difference between generic and directed prompting. Generic prompts are those that ask students to pause what they are doing and write down what they are thinking at that moment. In contrast, directed prompts ask students to reflect on specific things such as the development of a procedure or the process of data analysis. Both White & Frederiksen (1998) and Davis (2003) agreed that prompting was an effective way to help scaffold student thinking, but Davis proposed that generic prompts might be more effective in helping to promote student metacognitive thinking. Her reason was that the specificity of directed prompts might actually interfere with student thinking or lead them to think along the lines of "right answers" rather than true reflection. To investigate if there was a difference in student scores based on reflective prompt type, a mixed between-within subjects ANOVA was conducted. The interaction effect between reflection type and time was not significant (Wilks' Lambda = 0.91, $F(2,30) = 1.54$, $p = 0.23$). In contrast, the main effect for time was significant (Wilks' Lambda = 0.76, $F(2,30) = 4.65$, $p = 0.02$, partial eta squared = 0.24) agreeing with former observations that mean student inquiry scores tended to increase over the course of this study. The main effect for group membership was not significant ($F(1,31) = 0.05$, $p = 0.83$) which indicated that for this study, there was no difference by reflection type.

That there was no difference by reflection type was made even more apparent when the data was disaggregated by prior student achievement. Students were separated based on whether they were given generic or directed reflections and also by achievement, so there were a total of four groups. The data from these four groups were analyzed using a mixed between-within analysis of variance. The plot in Figure 15 shows mean student inquiry scores over time broken down by reflection type and prior achievement. The students with high prior achievement consistently outscored the students with low prior achievement regardless of whether they received generic or directed reflection prompts.

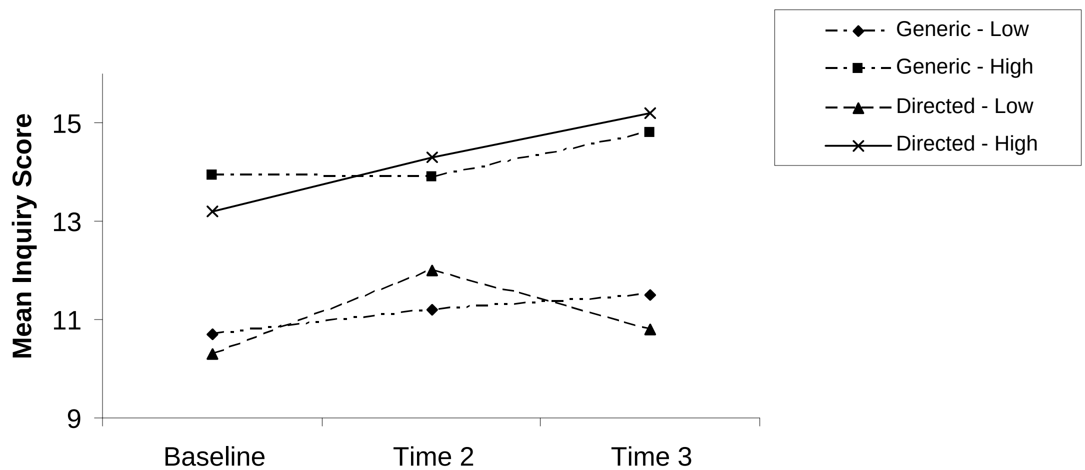


Figure 15: Comparison of mean inquiry score by both reflection type and prior student achievement

The results of the ANOVA indicated that there was no significant interaction between reflection type and achievement (Wilks' Lambda = 0.82, F(6,56) = 0.95, p = 0.47). The main effect for time (Wilks' Lambda = 0.80, F(2,28) = 3.51, p = 0.04) and for group membership (F(3,29) = 5.17, p = 0.006, partial eta squared = 0.35) were both significant. A follow-up post-hoc analysis using a Bonferroni adjustment indicated that the only significant differences were between the students with low prior achievement in the generic reflection group and the students with high prior achievement in both the generic reflection (p = 0.04) and the directed reflection (p = 0.04) groups. For the students in this study, the results indicated that there was no real difference between generic or directed reflection prompts.

The Effects of Goal Type (Proximal versus Distal)

At two places in the guided inquiry model (Magnusson & Palincsar, 2005), students were asked to take time to plan. Before investigating, students were asked to prepare to investigate by thinking through what they intended to investigate and what procedures they intended to use. In the same way, before reporting their results, students were asked to prepare to report their results. Students were asked to consider what they wanted to communicate and how they intended to communicate these findings. Zimmerman & Cleary (2006) referred to this as the Forethought phase. In this phase, students create goals for their upcoming task and plan steps needed to meet those goals. The literature on goal setting generally agreed that goal statements could be categorized as one of two types: proximal and distal (Bandura, 1997; Bandura & Schunk, 1981; Page-Voth & Graham, 1999; Zimmerman & Kitsantas, 1999). Proximal goals tended to focus on the immediate actions one could take to accomplish a task (eg. My goal is to use 3 sentences stems in my Framing and Analyzing & Concluding sections) whereas distal goals tended to focus on the long term outcomes (eg. My goal is to get a passing score on my Framing and Analyzing & Concluding sections). Specific to the writing process, proximal goals have tended to produce better outcomes (Bandura, 1997; Bandura & Schunk, 1981; Page-Voth & Graham, 1999; Zimmerman & Kitsantas, 1999).

In this study, students in all three of the experimental groups were asked to adopt proximal goals for both their investigation and reporting phases for their 2nd and

3rd inquiry tasks similar to the study by Bandura & Schunk (1981). Only two of these groups, though, were actually asked to write down their goals as a part of their reflective process. As a result, the data for this analysis was limited to the two classes that provided data related to their goals. The researcher examined student goal statements in order to determine if the goal statements were proximal or distal.

To assess whether a proximal or distal goal orientation made a difference, a series of independent-samples t-tests were conducted. These tests compared goal orientation to section score for the Framing and the Analysis section on the second and third inquiry tasks. The results indicated that there was no significant difference in mean section scores between the students that had adopted proximal goals and those that had adopted distal goals (2nd inquiry task Framing, $t(23) = 0.36$, $p = 0.72$ (two-tailed); 2nd inquiry task Analysis, $t(24) = 0.59$, $p = 0.56$ (two-tailed); 3rd inquiry task Framing, $t(21) = 1.14$, $p = 0.27$(two-tailed); 3rd inquiry task Analysis, $t(25) = 0.21$, $p = 0.83$ (two-tailed)). Even though the weight of evidence in the literature pointed towards the benefit of a proximal goal orientation, for the students in this study, goal orientation did not appear to make any difference.

The Effects of Sentence Stems

Comparison of Control and Treatment Groups

Three out of the four classes in this study were given training in the use of sentence stems to aid students as they wrote the reports for their inquiry tasks.

Students in the experimental group were given copies of the sentence stems to glue inside the front covers of their science notebooks. For about four weeks prior to the second inquiry task, students were asked to use these stems as they responded to short writing stems that were given in class. These sentence stems were used in an ongoing basis for the remainder of the school year.

An independent-samples t-test was conducted to compare student scores from the initial inquiry task to determine if there were any differences between student scores in the control and experimental group. The test indicated that there were no significant differences in mean inquiry task scores between the control group ($M = 11.9$, $SD = 3.2$) and the experimental group ($M = 12.4$, $SD = 2.6$), $t(56) = 0.55$, $p = 0.58$ (two-tailed) on the initial inquiry task. A mixed between-within subjects analysis of variance was then conducted to assess changes in mean inquiry score for the three inquiry tasks over time for both the control and experimental group. The interaction between time and group membership (Wilks' Lambda = 0.94, $F(2,52) = 1.71$, $p = 0.19$) was not significant. The same was true of the main effect for time (Wilks' Lambda = 0.96, $F(2,52) = 1.13$, $p = 0.33$) and group membership. These results indicated that there was no real change in scores over time and that instruction in the use of sentence stems did not make any difference.

Comparison by Actual Sentence Stem Use

A second analysis was conducted in order to determine how many students had actually used the sentence stems as they wrote their science inquiry reports. Instead of looking at the report as a whole, the focus turned to the Framing and Analysis sections. The focus shifted because these two sections required the greatest amount of expository writing. As it turned out, the number of students who actually used the sentence stems was quite low. On the second inquiry task, of the 51 students in the experimental group, only 12 of them (24%) used the stems in the Framing section and an even smaller number ($n = 8$, 16%) used the sentence stems to help write the Analysis. The same trend held true on the third inquiry task in this study. Out of the 46 students in the experimental group, only 7 students (15%) used the sentence stems in their Framing section while slightly more students ($n = 12$, 23%) used the sentence stems in their Analysis section.

A series of independent-samples t-tests were run to compare the section scores for students that had used the sentence stems with those who did not use the sentence stems. On the second inquiry tasks there was no difference between mean Framing section scores for the students who had used the sentence stems ($M = 3.6$, $SD = 0.4$) and those who had not used the sentence stems ($M = 3.2$, $SD = 0.8$), $t(53) = 1.83$, $p = 0.07$ (two-tailed). The same held true when scores from the Analysis section were evaluated. The difference in mean section score between the students who had used the sentence stems ($M = 3.4$, $SD = 0.6$) was not significantly different from the students who had not used the sentence stems ($M = 2.9$, $SD = 1.0$), $t(53) = 1.44$, $p = 0.16$ (two-tailed).

The same tests were run with the results from the third inquiry task. When scores from the Framing section were compared, students that used the sentence stems

(M = 4.1, SD = 0.5) had significantly higher scores than their counterparts who did not use the sentence stems (M = 3.2, SD = 0.6), t(53) = 3.46, p = 0.001 (two-tailed). For the Analysis section, mean student section scores for those who had used the stems (M = 3.4, SD = 0.6) were not significantly different from those who had not (M = 3.2, SD = 0.8), t(53) = 0.80, p = 0.43. In all cases, the students that had used the sentence stems had higher mean section scores than the students who did not choose to use the sentence stems. This difference, though, only reached significance in one of the four comparisons. It was predicted that if students used sentence stems, their science inquiry scores would increase, so while the results were encouraging, this treatment requires further study especially as it relates to implementation.

The low number of students in the experimental group to actually use the stems was initially surprising. That the students in this study were satisfied with inadequate strategies was evidenced by the high percentage of students who predicted that they would not receive passing marks on their papers, and yet handed them in anyway. On the 2nd inquiry task, 36 % percent of the students predicted a non-passing score on the Framing section and 59 % predicted a non-passing score on the Analysis section. On the 3rd inquiry task, 25 % percent of the students predicted a non-passing score on the Framing section and 43 % predicted a non-passing score on the Analysis section.

While it could have been argued that these students did not really understand the language of the scoring guide, a further analysis found that this was not the case. On the 2nd inquiry task, 14 of 22 students (64%) correctly predicted their score in the Framing section. This increased to 16 of 22 students (73%) correctly predicted their score for the Analysis section. Student section score predictions were similar on the 3rd inquiry task. For the Framing section, 22 of 44 students (50%) correctly predicted their section score and 28 of 44 students (64%) correctly predicted their section score for the Analysis section. This data demonstrated that students did indeed understand the requirements of the scoring guide.

Table 18: Comparison of student goals to actual use of sentence stems in the Framing section.

	2nd Inquiry Task (n = 26)	3rd Inquiry Task (n = 27)
Number of students who intended to use sentence stems in the Framing section	13	13
Number of students who used sentence stems in their Framing section	7	4

Table 19: Comparison of student goals to actual use of sentence stems in the Analysis section.

	2nd Inquiry Task (n = 26)	3rd Inquiry Task (n = 27)
Number of students who intended to use sentence stems in the Analysis section	14	13
Number of students who used sentence stems in their Analysis section	5	5

Student goal statements were also assessed to determine if there was a connection between these goal statements and actual student work. On both the second and third inquiry tasks, about half of the students who were in the experimental group indicated that they had adopted a proximal goal orientation of using sentence stems for the Framing and Analysis sections. Even though about half of the students indicated that they intended to use these sentence stems, very few students used these stems in their work (see Tables 18 & 19). There was little connection between student goal statements and their subsequent work.

The data from the second inquiry task was compared back to prior work to see the impact of the use of sentence stems. On the second inquiry task, students who used the sentence stems saw their Framing section scores go from a mean score of 3.7 (SD = 0.3) on their initial inquiry task to a mean score of 3.6 (SD = 0.4). Students who had not use the sentence stems saw their Framing section scores go from an initial mean score of 3.0 (SD = 0.7) to a mean score of 3.2 (SD = 0.8). The change in scores over time was assessed with a mixed between-within subjects ANOVA. The interaction term between group membership and time was not significant (Wilks' Lambda = 0.96, $F(1,52) = 2.46$, $p = 0.12$) and the same was true for the main effect of time (Wilks' Lambda = 0.99, $F(1,52) = 0.41$, $p = 0.53$) which implied that the scores were essentially flat over time. The main effect for group membership, though, was significant ($F(1,52) = 8.26$, $p = 0.006$, partial eta squared = 0.14). This implied that even though the students who had used the sentence stems on their second inquiry task had higher scores than those who did not use the sentence stems, the students who used the sentence stems were those that had high inquiry task scores on the initial inquiry task.

For the Analysis section, students who used the sentence stems saw their inquiry score increase from a mean of 3.0 (SD = 0.4) to a mean of 3.4 (SD = 0.6). The students who did not use the sentence stems saw their inquiry scores increase from a mean of 2.8 (SD = 0.8) to a mean of 2.9 (SD = 1.0). The was no significant interaction between time and group membership (Wilks' Lambda = 0.98, F(1,52) = 1.22, p = 0.27). The main effects for time (Wilks' Lambda = 0.96, F(1,52) = 1.94, p = 0.17) and for the use of sentence stems (F(1,52) = 1.58, p = 0.21) were both insignificant. While there were no real changes in student scores over time or by sentence prompt use for the Analysis section, the data from the Framing section implied that the students who chose to use the sentence stems were those that initially had higher science inquiry scores. The use of sentence stems did not increase student scores, rather it appeared that the use of sentence stems was associated with higher science inquiry task scores on the initial inquiry task.. It also appeared that when sentence stems were used, they had the greatest impact on scores in the Framing section.

This conclusion was supported with data from the third inquiry task. Scores were again compared using a mixed between-within subject ANOVA. For the Framing section, students who had used the sentence stems had mean section scores of 3.6 (SD = 0.5), 3.5 (SD = 0.5), and 4.1 (SD = 0.5) whereas the mean Framing section scores for students who did not use the sentence stems had mean section scores of 3.1 (SD = 0.5), 3.3 (SD = 0.6), and 3.2 (SD = 0.6). The interaction between the use of sentence stems and time was not significant (Wilks' Lambda = 0.89, F(2,49) = 2.96, p= 0.06) and neither was the main effect for time (Wilks' Lambda = 0.92, F(2,49) = 2.09, p = 0.13) which meant that the student scores over time were basically flat. In contrast, the students who had used sentence stems had significantly higher mean Framing section scores that those who had not used the sentence frames (F(1,50) =

8.08, $p = 0.006$, partial eta squared = 0.14). The results from the Analysis section were not as conclusive. Students who used the sentence stems saw their mean scores bounce around from 2.9 (SD = 0.5), to 3.5 (SD = 0.5), to 3.4 (SD = 0.6). The students who did not use the sentence stems saw their mean scores generally increase over time from 2.8 (SD = 0.8), to 2.8 (SD = 1.0), to 3.2 (SD = 0.8). Neither the interaction effect (Wilks' Lambda = 0.89, $F(2,49) = 3.00$, $p = 0.06$) or the main effect for sentence frame use ($F(2,49) = 1.99$, $p = 0.16$) were significant which indicated that scores were not greatly impacted by the use of sentence frames for the Analysis section. The main effect for time, though, was significant (Wilks' Lambda = 0.83, $F(2,49) = 4.93$, $p = 0.01$, partial eta squared = 0.17) which indicated that students scores did generally increase over time.

Taken together, these results implied that there was a correlation between high quality student work and the use of sentence stems. The data from this study, again, suggested that students who scored higher on the initial inquiry task already used language suggested by the sentence stems and so as these students adopted the language of these stems, there was no real change in their scores over time. In contrast, the students who initially did not score as well on their initial inquiry task did not use these sentence stems, so it was not possible to tell from this study whether or not the sentence stems provided a useful scaffold to the students they were designed to assist.

In addition, it was also intriguing to note that the use of the sentence stems appeared to have the greatest impact when used in the Framing section. There was very little difference in mean Analysis section scores for the students who used the sentence stems and those who did not. This might be due to the nature of the tasks in the two sections. In the Framing section, students reviewed prior work and concepts that were relevant to the study. The Framing section focused on summarizing current knowledge in order to provide a rationale for the study to follow. In contrast, the Analysis section asked students to review and interpret their data and then synthesize their findings with what they already knew. These findings will be discussed in more detail in the next chapter.

Word Counts and Inquiry Scores Study

Knowledge-Telling Model Analysis

The number of words that students used in the Framing and the Analysis sections emerged as a question of interest. As the inquiry papers were being assessed, it seemed that the Framing and Analysis sections were of similar word length. The reason that this was of interest was that in the knowledge-telling model of writing, students tend to focus on filling a pre-determined amount of space (eg. When a student asks “How many pages does this need to be?”) as opposed to communicating ideas (Bereiter & Scardamalia, 1987). If students used the same number of words in the Framing and the Analyzing sections, it would provide support for the idea that students tended to view the two sections as similar writing tasks. If the students were writing these sections from a knowledge-telling stance, they would tend to focus more on the length of the section and not as much on the content that they were communicating.

Table 20: Correlations for Framing and Analysis section word counts

	R value
Initial Inquiry	0.41*
Second Inquiry	0.60**
Third Inquiry	0.44**

* Correlation significant at the 0.01 level (2-tailed)
** Correlation significant at the 0.001 level (2-tailed)

There was a moderate to strong correlation between the number of words that a student used in the Framing section and the number of words that they used in the Analysis section. The strength of the correlation across all three of the inquiry tasks provided evidence that a majority of the students may have been utilizing knowledge-telling strategies as opposed to knowledge-transforming strategies. In light of the knowledge-telling model, these correlation values also indicated that students might view the process of Framing and Analyzing as similar types of writing tasks. While this is true in the sense that both sections are expository, as was discussed earlier in this chapter, the Framing section is focused predominantly on summarization whereas the Analysis section is more focused on knowledge synthesis. This will be discussed further in the next chapter.

Word Count and Score Correlation

Prior work by Bereiter & Scardamlia (1987) found that there was a positive correlation between the number of words that students wrote and the quality of the work that students produced. As the inquiry tasks in this study were assessed, it seemed that this correlation was also present. To determine if this was the case, a number of Pearson product-moment correlations were conducted between the number of words in a section and the score that the particular section received. Table 21 lists the results of this analysis. The correlation between the number of words in a section and score that this section received was significant, positive, and strong. The conclusion was that, in this study, when it came to producing quality science inquiry projects, students needed to be encouraged to write more.

But was it simply the process of writing more words that made a difference? The mean number of words that students used in the Framing section increased from 144 words ($SD = 51$) to 225 words ($SD = 113$). This increase was significant, $t(45) = 4.56$, $p < 0.0005$ (two-tailed). In contrast, mean Framing section scores began at 3.1 ($SD = 0.7$) and topped out at 3.3 ($SD = 0.7$). This increase was not significant ($t(55) = 1.66$, $p = 0.10$ (two-tailed)) which indicated that, in this study, Framing section scores over time did not change. For the Framing section, a significant increase in the number of words used did not necessarily result in an increase in the quality of the work produced.

Table 21: Pearson product-moment correlations between section word counts and scores

	Framing Section Score and Word Count	Analysis Section Score and Word Count
Initial Inquiry	0.70**	0.66**
Second Inquiry	0.57**	0.65**
Third Inquiry	0.74**	0.70**

** Correlation significant at the 0.0005 level (2-tailed)

This was not true, through, for the Analysis section. The mean number of words that students used in the Analysis section increased from 211 (SD = 121) to 266 (SD = 139). This increase in mean Analysis section word count was significant, t(45) = 3.10, p = 0.003 (two-tailed). This increase in mean word count was also accompanied by an increase in Analysis section scores. Mean Analysis section scores increased from 2.9 (SD = 0.7) to 3.2 (0.7) over the course of this study. This increase in mean Analysis section scores was significant, t(55) = 3.34, p = 0.002 (two-tailed). The data from the Framing section seemed to imply that when students wrote more words, they were simply filling space and were not adding any value to their work by writing more. In contrast, as students wrote more in the Analysis section, these additional words added important content to their work. Writing more words did not guarantee that work was of higher quality, but it was clear that writing too few words had a negative impact on student work.

Word Count by Achievement

To further investigate the relationship between word count and score, a mixed between-within subjects analysis of variance was conducted on the mean word counts for the three inquiry tasks over time. Students were divided into groups based on prior achievement. For mean word counts from the Framing section, the interaction term between time and group membership was not significant (Wilks' Lambda = 0.91, $F(2,40) = 2.08$, $p = 0.14$). In contrast, the main effect for time was significant (Wilks' Lambda = 0.72, $F(2,40) = 7.67$, $p = 0.002$, partial eta squared = 0.28) as was the main effect for prior achievement ($F(1,41) = 6.89$, $p = 0.01$, partial eta squared = 0.14) which indicated that mean word counts increased over time for both achievement groups and that students in the high prior achievement group had significantly more words in their Framing sections than did their low prior achievement counterparts.

For the mean word counts from the Analysis section, neither the interaction term (Wilks' Lambda = 0.99, $F(2,40) = 0.13$, $p = 0.88$) nor the main effect for time were significant (Wilks' Lambda = 0.88, $F(2,40) = 2.68$, $p = 0.08$) which indicated that mean word counts did not significantly change for either group over time. The main effect for group membership, though, was significant ($F(1,41) = 6.10$, $p = 0.02$, partial eta squared = 0.13) which indicated that students with high prior achievement consistently used more words in the Analysis section than did their low prior achievement counterparts. In both the Framing and Analysis sections, the higher achieving students wrote more words.

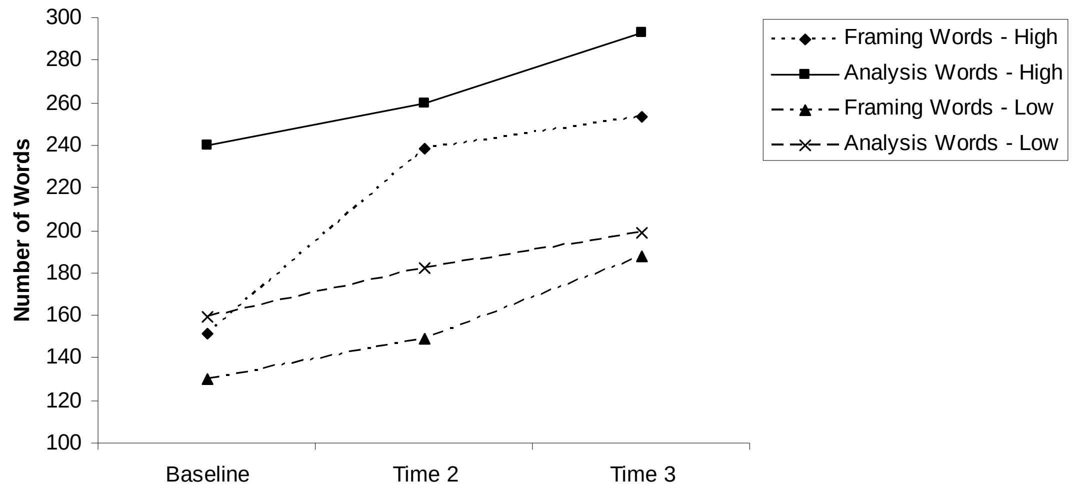


Figure 16: Comparison of mean number of words over time used in the Framing and Analysis sections by prior student achievement

Word Counts and Sentence Stems

Prior work in this chapter had indicated that there was a relationship between the use of sentence stems and high initial inquiry task score. Knowing that there was a correlation between word count and section score, the final analysis in this chapter explored the relationship between word count and the use of sentence stems. A series of independent-samples t-test were used to compare mean numbers of words in the Framing and Analysis sections for the second and third inquiry tasks. In the Framing section of the second inquiry task, students who used the sentence stems had a mean word count of 267 words (SD = 109) compared to a mean of only 201 words (SD = 110) for the students who did not use the sentence stems. In the Analysis section, students who used the sentence stems averaged 239 words (SD = 78) and those who did not use sentence stems only averaged 232 words (SD = 123). These differences in word counts were not significant (Framing section, t(52) = 1.84, p = 0.07 (two-tailed) and the Analysis section, t(52) = 0.15, p = 0.88 (two-tailed))

On the third inquiry task, students who used the sentence stems wrote 324 words (SD = 108) on average in the Framing section compared to only 209 words (SD = 106) in the Framing section for the students who did not use the sentence frames. For the Analysis section, the students who used the sentence stems actually wrote fewer words on average (M = 244, SD = 97) than the students who did not use the sentence stems (M = 253, SD = 145). The difference in mean word count for the Framing section was significant (t(53) = 2.68, p = 0.01 (two-tailed)) whereas the difference in mean word count for the Analysis section was not significant (t(53) = 0.21, p = 0.84 (two-tailed)).

In only one of the four instances did the group who used sentence stems write significantly more than the students who did not use the sentence stems. This finding was consistent with the prior conclusion that while more words does not necessarily guarantee a better project, writing more words might help produce a better product. Even more important, though, is the content of what is written. A host of words cannot make-up for a lack of content.

Conclusion

This study sought to deepen our understanding of the cognitive demands that students face as they engage in scientific inquiry tasks. This was done by attempting to integrate the frameworks of guided inquiry (Magnusson & Palincsar, 2005) and self-regulation (Zimmerman & Cleary) as a way of helping students understand how to approach various stages of an inquiry task. To aid in the communication demand specifically, this study relied heavily on the knowledge-transforming framework of Bereiter & Scardamalia (1987). Through this investigation, changes in student thinking were monitored through continuous evaluations of their self-efficacy (Bandura 1986, 1997).

In relation to self-efficacy, this study revealed three interesting results. First, student self-efficacy for science inquiry as measured by the CSES was very stable. Student beliefs about their abilities in science inquiry did not change over the course of this study even though inquiry scores generally increased over time. Second, there was a strong relationship between a student's sense of self-efficacy and the scores that they received on their science inquiry tasks. This self-efficacy measure should not be construed as related to students' actual abilities. However, the results of this study provide support for the conclusion that students' beliefs about their abilities were indeed related to the products they produced. Finally, the self-efficacy data provided strong evidence that students in this study did not view all parts of the inquiry task as equally challenging. Most students indicated that they felt less efficacious in their abilities in the Framing and Analyzing sections and more efficacious about their abilities in Collecting data. Between Framing and Analyzing, students consistently felt less efficacious about their abilities in Analyzing.

The results related to the application of self-regulation theory through the use of metacognitive prompts were mixed. Unlike prior studies (Davis, 2003; Coleman 1998; White & Frederikson, 1998), the students in this study who were given the metacognitive reflective prompts throughout their science inquiry tasks did not fare any better or worse than those who did not experience this condition. As was already stated, inquiry scores did tend to increase over time for the students in this study. This study also provided evidence that high achieving students are high achievers because they tend to be naturally metacognitive. Interestingly, in this study, the metacognitive prompts had no benefit for the lower achieving students that these prompts were designed to help.

Similar to the results for the metacognitive prompts, the results related to the effect of sentence stem use was mixed. The greatest difficulty was in actually getting students to adopt the use of these sentence stems as a strategy in their work. Students with high initially inquiry task scores seemed to have no problem adapting to the use of these sentence stems and their high scoring work remained high scoring. In contrast, students with low initial inquiry task scores tended to ignore the sentence stems so it was difficult to assess whether or not students with these low initial inquiry task scores might benefit from this type of scaffolding. The sentence stems used in this study seemed to have the greatest impact when used in the Framing section of the inquiry task.

In addition to sentence frames, section word counts also emerged as point of interest because the data collected in this study suggested that students were using a knowledge-telling strategy as opposed to a knowledge-transforming strategy. Higher achieving students tended to write more words than students who were lower achieving. While it became clear that students should be encouraged to write more if they wanted their projects to improve, it was not word count alone that mattered. The content of what students wrote was also important. The issue was one of both quantity and quality.

Chapter 5: Discussion

Introduction

This chapter takes the data from Chapter 4 in order to draw connections between the findings in this study and current research. The discussion generally follows the order presented by the three research questions guiding this study. Starting first with self-efficacy, this chapter examines the extent to which the data supports the model of student self-efficacy presented in Chapter 1. Student inquiry results are then examined in light of the guided inquiry model (Magnusson & Palincsar, 2005). Finally the feedback from the evaluation of student inquiry projects is used to evaluate the effectiveness of sentence stems to aid student thinking. The chapter concludes with recommendations for future research and a discussion of the limitations inherent to this study.

Self-Efficacy

In this study, the construct of self-efficacy was used to track changes in student cognition in order to understand how student thinking changed as a result of participation. Again, self-efficacy is an individual's belief in their ability to accomplish a specific task. This study tracked student self-efficacy for chemistry content knowledge and student self-efficacy for carrying out scientific inquiry.

One of the assumptions of this study was that student difficulties with science inquiry were based on challenges with the task of communication and not necessarily a reflection of difficulty with content. Student self-efficacy for chemistry content was tracked in order to determine if this was indeed the case with the students in this study. During the time that students were engaged in the experimental portion of the study, student self-efficacy scores for content hovered between 4.6 and 5.0 on average out of a 7 point scale. These numbers indicated that students felt slightly efficacious about their chemistry knowledge and that they had a general sense of confidence in their understanding of the course content.

It was interesting to note that average student self-efficacy for chemistry content jumped significantly between the first and second administrations of the survey but did not change much over the three final administrations of the survey. While average student self-efficacy for content was fairly static over the latter administrations, student responses to the various survey items were quite dynamic. Students tended to report that they felt most efficacious about the content that they had either just finished covering or were in the process of learning. The initial leap in student self-efficacy scores for content coupled with the dynamic nature of student responses to individual survey items provided evidence that student beliefs about their abilities in chemistry exhibited a positive change over the course of this study.

Even though student responses indicated that they felt generally efficacious about their abilities to correctly solve chemistry problems, the students in this study indicated that they felt more efficacious about their abilities in science inquiry than they did about their abilities in chemistry content. Unlike the self-efficacy results for content which did show growth over time, student self-efficacy for science inquiry did not significantly change over the course of this study. At the beginning of the year, students indicated that they felt fairly efficacious about their abilities in science inquiry and this changed little over the course of the school year.

This finding was interesting for a number of reasons. First, very few of the students in this study were able to produce an inquiry write-up that met the passing requirements in the four scored dimensions. Students believed that they were competent in their abilities even though their work did not meet the standards. The reflective pieces that the students completed indicated that they had a clear understanding of the requirements of the scoring guide, so confusion about the requirements of scoring guide was not an issue. A better explanation may be that students were satisfied with their current inquiry strategies and were not willing to trade these strategies for more effective ones. Or perhaps students did not believe that they possessed the necessary skills needed to utilize new strategies. Either way, it was clear that students did not adopt new strategies presented in this study and, instead, continued to use their current, marginally successfully strategies. This student resistance to change is not a new finding and is in line with other studies that have noted similar resistance (Brown et al, 1983; Zimmerman, 1995; Zimmerman & Cleary, 2006).

The second finding of interest was that student inquiry scores generally increased over time for the students in this study. Student inquiry scores were significantly higher at the end of the study. Even though a review of student work indicated that many of the students did not utilize the tools from the various treatment conditions, students generally saw their inquiry scores increase over the course of this study. This would have provided evidence to students that there was no real need to adopt new inquiry strategies. By sticking with and refining familiar strategies, scores increased. Albeit, only a few of the papers met the standards of the scoring guide, the fact that student scores generally increased would have reinforced the idea that there was no need to adopt new writing strategies because current strategies were having a positive effect.

An increase in inquiry scores may have also been due to students becoming more familiar with the requirements of the scoring guide. All of the students in this study were given instruction in the scoring guide. Feedback from students in two of the treatment conditions indicated that they were able to correctly apply the requirements of the scoring guide to their work. The increase in student scores may also have been the result of the two opportunities to practice doing and writing science inquiry tasks prior to the final task in this study. The initial two tasks may have served as opportunities for students to practice their communication skills.

The science inquiry self-efficacy data provided evidence about how students felt about their abilities to complete various parts of the inquiry task. In all but one of the administrations of the survey, students felt significantly more efficacious about their abilities to collect and present data than they did about their abilities to analyze their data and explain their findings. In a similar way, in two of the four survey administrations, students indicated that they felt more efficacious about their abilities to collect and present data than they did about their abilities to frame the investigation.

In contrast, in three of the four surveys, there was no real difference in student self-efficacy for designing an investigation or for collecting and presenting data. In all four administrations of the survey, there was no significant difference between student efficacy for framing an investigation and for analyzing the results. It was hypothesized in the opening chapter of this study that students' struggles with science inquiry were due in part to inabilities to communicate effectively. Students in this study had the least confidence on the parts of the inquiry task that required the greatest amount of expository writing. In contrast, students felt much more confident about their abilities on the parts of the inquiry task that required the least amount of writing.

This finding was in line with current research on the role of language in science. Yore & Treagust (2006) argued that the academic language requirements of science are unique to the discipline and are distinctly different from the language that students use in their everyday interactions with peers or even the academic language requirements of other content areas. They contended that the academic language of science must be explicitly taught to students if they are to find success in science classrooms. This is because the role and nature of communication in a science classroom has changed over time. Early views of language in science saw language as a fairly inflexible tool and a means to impart information (Klein, 2006). The key concern was ensuring that language was precise and accurate. In contrast, the present role of language in the science classroom is viewed as being much more dynamic and as a vehicle for constructing new knowledge (Klein, 2006). Students are active participants in the activity of knowledge construction in a language mediated process.

The students in this study felt most competent when the language demands of the inquiry task were merely factual recall (ie. steps in a procedure, labels on a graph, etc.). In contrast, when the language demands of the inquiry task asked that students synthesize their prior learning or construct new knowledge based on their findings, the students indicated that they felt much less capable. More time needs to be spent developing tools to help students with the writing demands of science inquiry tasks.

In addition to helping analyze the science inquiry task, the self-efficacy data also helped assess the model of student self-efficacy development that was presented in the first chapter. For the first two administrations of the CSES, student self-efficacy for inquiry was moderately and significantly correlated only to the scores from the first two inquiry tasks but was not significantly correlated to the scores on the third inquiry task. The first and second administrations of the CSES came directly before the first and second inquiry tasks respectively. In contrast, student self-efficacy for science inquiry as measured by the third and fourth administrations of the survey were moderately and significantly correlated to all of the science inquiry tasks in this study. Since this analysis was only correlational, it was not possible to determine causation. What was clear, though, was that there is a relationship between a student's sense of self-efficacy for science inquiry and their performance on science inquiry tasks.

This finding was in line with self-efficacy model proposed in the first chapter of this study and with other research in self-efficacy (Bandura, 1986, 1997; Pajares & Urdan, 2006; Tschannen-Moran, Woolfolk Hoy & Hoy, 1998; Woolfolk Hoy & Davis, 2006). In the self-efficacy model used in this study, an individual's performance acts as a new source of efficacy information. This information then goes through cognitive processing which leads to an adjustment in one's sense of self-efficacy. This updated sense of efficacy then influences future task performance. The results of this performance become a new source of self-efficacy information and the cycle then repeats itself. The data in this study provide further support for this model.

Even though the data provided support for the self-efficacy model in one sense, the support for the model from this study was limited. Student self-efficacy for science inquiry remained unchanged over the course of the study. This was even in spite of science inquiry task scores increasing over the course of this study. It would have been expected that as student inquiry scores increased, that student self-efficacy for science inquiry would increase as well. It may have been that students began the study with an artificially high sense of self-efficacy for science inquiry and that over the course of this investigation, student science inquiry self-efficacy scores remained unchanged because the scores on the inquiry tasks were simply catching up to what the students believed they were capable of doing. Another possibility is that students did not have enough time to internalize the results of their work in such a way that allowed them to incorporate this new information into their self-efficacy beliefs. Whatever the cause, it was clear that the treatments in this study did not have the intended effect of increased student self-efficacy for science inquiry.

There were two other results of interest. First, all through the study, high achieving students had significantly greater self-efficacy for science inquiry than the lower achieving students in the study. This finding was expected and provided evidence that the CSES measured what it was designed to measure. Second, there were no differences in science inquiry self-efficacy scores for male and female students. Both students felt equally efficacious about their abilities in science inquiry. There was a question based on a review of the literature as to whether inquiry tasks might favor one gender over the other (Jacobs et al, 2002; Pajares, 2003; Zimmerman & Martinez-Pons, 1990). In this study, both genders faired equally well.

Study Model

This study sought to apply the guided inquiry model (Magnusson & Palincsar, 2005) to a high school setting. The guided inquiry model was merged with the model of self-regulation (Zimmerman & Cleary, 2006) in order to highlight specific metacognitive strategies that students should employ as they worked their way through the inquiry task. This study looked at the effect of prompt types and goal types. Generally speaking there was little difference between the students in the control group and the experimental groups. The students in the experimental condition who received various prompts to reflect on their work did not have significantly different inquiry scores than the students who were not given the prompts to reflect. This was unexpected as the literature reviewed for this study indicated that high quality student work was associated with high quality student reflection (Coleman, 1998; Davis, 2003; Flick & Tomlinson, 2006; White & Frederiksen, 1998).

There are a number of reasons as to why this may have been the case. First, over the course of the inquiry task, the students in the reflection group were only asked to reflect a total of six times. In contrast, in the Davis (2003) study, students were asked to reflect a total of 11 times over the course of working through one project. White & Frederiksen (1998) did not specify the number of times that students were asked to reflect, but it was clear that the number was fairly high because when students were asked to give feedback on the on the study, a common complaint was that there was too much reflection. It may have been that in this study, six prompts to reflect were not enough to significantly impact or change student thinking.

Another reason why student reflection did not seem to make any difference might lie in the quality of reflection that the students produced. The work of both Coleman (1998) and Davis (2003) indicated that high quality student work was directly related to high quality student reflection. The second possibility, then, is that student reflections in this study were not of high enough quality to have had a beneficial impact on student scores. Students were not given any instruction in how to write quality reflections and this may have resulted in weak reflections. Student reflections were not evaluated for quality in this study, and so this conclusion is tentative.

A similar result was found related to student project goals. A number of studies have looked at the relationship between goal orientation and the quality of work produced. There was strong support for the benefit of proximal goals over distal goals (Bandura & Schunk, 1981; Page-Voth & Graham, 1999; Zimmerman & Kitsantas, 1999). Proximal goals are more immediate in nature and focus on an action that one plans to take to accomplish a task (ie. I will study for 15 minutes every night this week). In contrast, distal goals are those which focus on a future outcome, but do not specify the actions necessary to accomplish the goal (ie. I want to get an A on my test next week). In this study, students in two of the experimental groups were given instruction in proximal goals and then asked to write down a goal for the writing portion of the study. These goal statements were assessed and classified as either proximal or distal. In contrast to the research reviewed for this study, goal orientation did not make any difference on student scores. Student scores were the same regardless of whether students had adopted a proximal or a distal goal to the written portion of the inquiry task.

It would have been interesting to know why goal orientation did not have any effect. One alternative is that there was a discrepancy between the goals that the students wrote down on their reflection sheets and the goals that they actually adopted as they wrote their projects. Students had been given instruction on the nature of goal setting and the benefit of adopting a proximal goal attitude and were then asked to write a proximal goal for the written portion of their inquiry. There was often a break of a couple of days between instruction in goal setting and the subsequent writing. Even though about half of the students in the experimental condition stated that their writing goal was to use a specific number of sentence stems, the actual number of these students that actually used the sentence stems was quite small. This provided evidence that there was a disconnect between the goal that the students had stated and the goal that they actually used as they wrote. It seemed that even though students wrote down proximal goals, they did not internalize what the proximal goal actually asked them to do. This may be another piece of evidence that students were resistant to change. Students may have written down proximal goals simply because they had been asked to do so, but did not see the connection between their proximal goals and the nature of the writing task.

Another possibility is that the nature of the proximal goal was not adequate to the scope of work required for the science inquiry write-up. For the Zimmerman & Kitsantas (1999) study, the students had adopted proximal goals, but were only working on reconstructing sentences. In the Page-Voth & Graham (1999) study, while the students were asked to write short essays, they were asked to focus on increasing the number of arguments and counterarguments in their work. In this study, students were asked to adopt the proximal goal of sentence stem usage in order to keep the goal simple and attainable. It may have been that this proximal goal was too simplistic and did not adequately address student challenges related to reporting the results of an inquiry task which resulted in abandonment of the proximal goal.

Even though the results from this study indicated that the student inquiry scores in the experimental group with reflection were not significantly different from students in the groups that did not include reflection, the results from this study indicated that higher achieving students were higher achieving because they utilized reflective metacognitive strategies. There was no difference between the inquiry scores of high achieving students in the experimental condition with reflection and the inquiry scores of the high achieving students in the control group. This finding supports prior work (Coleman, 1998; Davis, 2003, White & Frederiksen, 1998) that concluded higher achieving students were higher achieving because they naturally utilized reflective metacognitive strategies.

Although the conclusions related to high achieving students and metacogntivie reflection was expected, the results related to low achieving students and metacognitive reflection was surprising. For the high achieving students, the act of asking students to reflect had no real effect on their work when compared to high achieving students in the control group. In contrast, low achieving student in the experimental group that included reflection actually did significantly worse than their high achieving peers. The data seemed to indicate that the act of asking these low achieving students to reflect on their work actually hindered their ability to successfully write a quality inquiry task. This was quite unexpected as prior work by White & Frederiksen (1998) and Coleman (1998) both found that asking low achieving students to reflect on their work resulted in a significant increase in the quality of work that these students were able to produce.

There are a number of potential reasons for this finding. First, there may have been too few reflective questions to make a real difference in thinking of these low achieving students. Another possibility is that the act of reflecting made incorrect thinking more concrete and more resistant to change. A third possibility is that these low achieving students found that the act of reflecting produced a significant amount of cognitive load which resulted less cognitive space to focus on the task of writing (Mayer, 2002). While there are a number of possibilities that explain this particular result, future study is needed to determine what it was about these prompts that seemed to hinder the success of these low achieving students.

Implications for Classroom Practice

While guided inquiry (Magnusson & Palincsar, 2005) provided the overarching framework for this study, the knowledge-transforming framework (Bereiter & Scardamalia, 1987) was adopted in order to help students with the communication demands of the inquiry task. Inquiry tasks, by their design, are intended to be knowledge-transforming experiences. Inquiry tasks ask students to apply their current knowledge to a problem in order to extend their understanding. Writing from a knowledge-transforming perspective requires that students develop a general plan about what to write, continually monitor meaning as they write, and then revise as necessary once the writing is complete. In contrast, text written from a knowledge-telling perspective lacks evidence of advance planning, appears to be the product of linear thinking, and shows little evidence of revision. Novice writers often tend to use a knowledge-telling orientation because the knowledge-telling orientation tends to mirror the familiar patterns of everyday speech. Bereiter & Scardamalia (1987) noticed that novice writers tended to struggle with producing content because they were in essence trying to carry two sides of a conversation without any outside assistance.

Sentence stems were used in this study to help students write from a more knowledge-transforming perspective. These stems provided a way for students to move away from writing as an extension of conversation to seeing writing as a separate and distinct mode of communication. These stems such as “An example of this is…” or “Another reason I think so…” provided opportunities for students expand on what they had written. The goal was for these stems to prompt further thought similar to the way that another person might ask for further clarification in the course of a conversation. In using these stems, students would be prompted to elaborate on an idea or introduce a new thought that had not yet been stated. The use of these sentence frames was also intended to help students create more content. Again, it was noted that one of the greatest struggles novice writers face was that of generating content (Scardamalia, Bereiter, & Steinbach, 1984). Scardamalia, Bereiter, & Steinbach (1984) had found that when students were given sentence stems, they tended to write more. To this end, sentence stems were used in this study in order to help students generate content.

The effect for the use of sentence stems was mixed. Only a small number of students actually used the sentence stems. In addition, the use of sentence stems did not significantly impact inquiry scores except in the Framing section on the 3rd inquiry task. On the 3rd inquiry task, students who had used the sentence stems had significantly higher Framing section scores than students who had not used the sentence stems. Further analysis revealed that the students who had higher inquiry task scores on the initial inquiry task were the ones who tended to use the sentence stems. The students with low initial inquiry task scores did not experience any positive benefits from the sentence stems because they did not use them.

While highly tentative, these data implied two things. First, the sentence stems in this study appeared to be best suited for the Framing section of the inquiry project. These stems were originally developed to helped students become better writers in the context of an English course (Bereiter & Scardamalia, 1987; Scardamalia et al, 1984). The Framing section was predominantly expository as the major focus of the Framing section was to summarize current understanding and set the stage for the purpose of the study. In contrast, the Analysis section required language constructions that were more germane to the literacy of argumentation and explanation used in science (Klein, 2006; Yore & Treagust, 2006). Sentence stems need to be developed that speak more specifically to the science literacy demands of the Analysis section in order to help students write coherently about their findings and conclusions.

Second, for the students in this study, it was assumed that the students with higher initial inquiry task scores chose to use the sentence stems because these stems mirrored sentence constructions that they already used. The question, then, is how to convince other students to use these sentence stems. Prior work with these stems had shown that these stems have had a positive impact on lower achieving students (Coleman, 1998; Scardamalia et al, 1984). This work, though, was done with elementary students in much smaller settings. More work is needed in order to help students understand how to use the sentence stems and gain more confidence with their application.

In addition to sentence stems, the number of words that students used in the Framing and Analysis sections emerged as a topic of interest. Similar to sentence stems and section score, the use of sentence stems did not produce any difference in the number of words used except in the Framing section of the 3rd inquiry task. In the Framing section of the 3rd inquiry task, the students who used sentence stems wrote significantly more words than the students who did not use sentence frames. This was also the only section where the students who used sentence stems scored significantly higher than those who did not use the sentence stems. In addition, it was clear that higher achieving students tended to use more words on their inquiry write-ups than their lower achieving counter parts and higher word counts were positively and significantly correlated to higher section scores. This is in line with the findings of Bereiter & Scardamalia (1987) who noted that novice writes tended to struggle with creating content whereas this was not necessarily the case with expert writers. Also, it should be noted, that more words did not necessarily mean higher scores. Simply writing more words was not a guarantee that a student’s score would increase. Content was also an important factor. In this study, though, it appeared that the low achieving students tended to write too few words, which then impacted their ability to be successful.

The word count analysis also provided some insight into the type of writing strategies that students used as they completed the written section of the inquiry task. In the knowledge-telling model (Bereiter & Scardamalia, 1987), novice writers viewed the goal of writing as that of filling up a pre-determined amount of space. Once enough content had been generated, novice writers tended to stop writing. If a novice writer found that they were not able to generate enough text to fill this pre-determined amount of space, they tended to go back and restate their ideas until the space was filled. In contrast, writers writing from a knowledge-transforming orientation were less concerned about the length of their writing but were more concerned about the content of their work.

The results from this study indicated that students may have been writing from a knowledge-telling perspective. That word counts from the Framing and Analysis sections were positively and significantly correlated could be interpreted as an indication that students were more focused on filling a pre-determined amount of space and not as focused on the generation of content. Student work was not assessed beyond word counts and student thinking was not assessed, so this conclusion is highly tentative.

While the data support the conclusion that many students may have been writing from a knowledge-telling perspective, student scores on their inquiry projects in this study also generally increased over time. This score increase was significant. The data indicated that experimental condition made no real difference in terms of student scores. Two conditions though, did appear to make a difference. First, before the second inquiry task in this study, all students were given instruction on the scoring guide that was used in this study. Students had the chance to read through the grading descriptors and then apply these descriptors to two inquiry tasks that were provided by the state. The value of this was two-fold. First, because the students had to use the scoring guide to assess two different student work samples, it meant that these students had to internalize what the words of the scoring guide meant in order to provide an accurate assessment. Secondly, one of the sample inquiry tasks met the benchmark standard on all four of the scored dimensions and the other sample inquiry task did not. This gave students a chance to see a model of what a passing sample looked like and also what types of characteristics were true of a study that did not meet the passing standards.

In addition to familiarity with the scoring guide, all of the students in this study participated in three inquiry tasks over the course of the school year. Students had the opportunity to learn from their previous work and then make adjustments as needed. This repeated exposure to science inquiry tasks appeared to have had a positive effect on student science inquiry scores.

Implications for Future Research

The goal of this work was to expand our understanding of student cognition as it related specifically to the task of science inquiry. Specifically, the focus of this study was on the language and communication demands of science inquiry tasks. There were a number of conclusions reached in this study that warrant further investigation.

Related to student cognition, student self-efficacy for science inquiry was positive and remained relatively stable over the course of this study. This was true even though few student samples completely met benchmark standards. Are these results typical for high school students? What types of activities and experiences cause the greatest changes in student self-efficacy for science inquiry at the high school setting? Why do students feel generally efficacious about their abilities in scientific inquiry even though their work does not meet the benchmark standards? More work is needed in order to address these questions.

More work is also needed to verify the accuracy of the efficacy instrument used in this study. Even though the instrument used in this study held up well to psychometric analysis, the validity of this instrument needs to be evaluated with a much larger sample size. An analysis of the instrument with a broader sample is needed if this instrument is to be used in other settings.

The findings related to the impact of the use of sentence stems to prompt student thinking also warrants further study. While there were indications that these sentence stems were beneficial, these results were by no means conclusive. Do sentence stems truly help reduce the cognitive demand that students face as they communicate the results of their investigations? Also, there seemed to be a general resistance to using the sentence stems. While student resistance to new strategies has been well documented (Brown et al, 1983), more work is necessary to understand how to help students overcome this resistance and release their reliance on ineffective strategies in order to adopt new, more effective ones.

In a similar way, the results related to the effectiveness of reflective prompts were surprising. It was expected that the students who were encouraged to reflect over the course of the investigation and writing portions of the inquiry task would do better than those that were not prompted to reflect. This was not the case. In this study, students in the experimental condition which included reflection did not do any better or worse than the students who were not prompted to reflect. There is strong support in the literature related to the benefit of asking students to be reflective about their work, so this result was unexpected. Was there something about the implementation of the reflective prompts that hindered their effectiveness? Also, these reflective prompts actually seemed to hinder the low achieving students in this study. What was it about these prompts that hindered these low achieving students from being successful and were these results unique to the student population in this study?

Finally, while much work has been done around science inquiry at the elementary and middle school levels, this study focused on the challenges that student face in a high school setting. The content that high school students are expected to master is more demanding as are the types of questions that these students are asked to investigate. In this study, it was evident that students found the Framing and Analysis sections to be the most difficult to complete. Students struggled with the academic language demands of communicating in scientific ways. Work in this area needs to continue in order to identify ways to help students become more proficient in the language of science.

Limitations of the Study

There are a number of limitations that were inherent to the design of this study. While these limitations by no means invalidate the conclusions of the study, these limitations do limit the broad generalizability of this work. This study was designed to be exploratory in nature and so future work evaluating the conclusions of this study with a broader group of students is both desired and necessary.

One of the limitations of this study was the size of the sample. While there were about 120 students in the four classes in this study, only 62 of these students agreed to participate in the study. This kept the sample size in each of the four classes in this study to about 15 students per experimental condition. For each analysis, the data was assessed to ensure that it did not violate any of the necessary assumptions. The small sample size, though, limits the ability of this sample to be viewed as representative and thus limits the generalizability of the findings in this study.

In addition to the small size of the sample, another limitation of the study was that only about half of the students in these chemistry courses chose to let their data be used in this study. Half of the data from the chemistry courses was not available for analysis. Anecdotal evidence from student feedback suggested that some of the students who elected to take part were the ones that felt favorably about the teacher and were highly motivated by the though of being involved in a research study. This is in line with the literature review by Rosenthal & Rosnow (1975) that characterized volunteers as having many positive qualities in comparison to nonvolunteers. And while the researcher did not know which students had chosen to take part in the study until the conclusion of the course, there were a number of students who approached the researcher at various times to communicate that they did not want to take part in the study because they felt insecure about their abilities in science. These were the students that this study was designed to assist. Students were randomly assigned to treatment conditions by class period in order to help limit this selection effect. Again, though, the nature of the sample and how it was obtained severely limits the generalizibity of this study beyond the student who chose to participate.

In addition to the selection effect, it is difficult to know the extent to which the Hawthorne effect also affected the results. Generally speaking, science inquiry task scores increased significantly over time for the all students in this study regardless of treatment condition. Was this due to the repeated exposure to inquiry tasks? A more thorough understanding of the scoring guide? Or did science inquiry task scores increase for the students in this study because they were motivated by the fact that they were involved in a research study? An attempt was made to limit this effect by imbedding the inquiry tasks into the chemistry curriculum. Also, these tasks were similar to ones that students had experienced in prior years in science courses, so the idea of a science inquiry task was not novel. Future studies should focus on obtaining a larger sample size and finding ways to make the study and treatments less obtrusive.

The sample in this study was a sample of convenience. While the classes were randomly assigned to the various experimental conditions in this study, the students were not. There were factors such as the placement of advanced music and math electives in the master schedule that affected what periods the students in this study were free to take chemistry. To generalize the results of this study, the students would need to have been randomly assigned to various experimental conditions.

Another factor that may have played a role in this study was the potential for the exchange of information between the control and experimental groups. While the conditions in the various classes were distinct, students in the various conditions were not prohibited from talking to each other as this would have been unenforceable. Student work from the control group was examined to see if any of these students had used the sentence stems. There was no evidence from student work, though, that this had been the case. There was no difference, though, in student inquiry scores between the control and the other experimental conditions, so it was possible that information was shared between experimental conditions.

Finally, since this study was exploratory in nature, it was designed to have high internal validity. The researcher conducted the various levels of treatment and so was able to monitor the fidelity of treatment and control conditions. The design included steps to check the reliability of each measure and established strong internal validity. This strong internal validity came at the expense of generalizability. There were indications that the treatments in this study, such as the use of sentence stems, could provide help with the communication demands of a science inquiry task. The conclusions related to the effectiveness of reflective prompting, though, were inconclusive. These findings merit further study. Future work should examine the extent to which the findings in this study can be replicated with other students and be extended to a broader population.

Bibliography

Ablard, K.E., & Lipschultz, R.E. (1998). Self-regulated learning in high-achieving students: relations to advanced reasoning, achievement goals, and gender. *Journal of Educational Psychology,* 90(1), 94-101.

Abrams, L.M. & Madaus, G.F. (2003). The lessons of high stakes testing. *Educational Leadership*, 61(3), 31-36.

American Association for the Advancement of Science (AAAS). (1990). *Project 2061: Science for all Americans*. New York: Oxford University Press.

American Association for the Advancement of Science (AAAS). (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

Baker, D.R. (2004). Focus on science literacy: The role of writing and speaking in the construction of scientific knowledge. *Eurasian Journal of Educational Research*, 16, 1-7.

Baker, W.P., Barstack, R., Clark, D., Hull, E., Goodman, B., Kook, J., Kraft, K., Ramakrishna, P., Roberts, E., Shaw, J., Weaver, D., & Lang, M. (2008). Writing-to-learn in the inquiry-science classroom: Effective strategies from middle school science and writing teachers. *The Clearing House,* 81(3), 105-109.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (Ed.). (1995). *Self-Efficacy in Changing Societies*. Cambridge: Cambridge University Press.

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

Bandura, A. (2006a). Adolescent development from an agentic perspective. In Pajares, F. & Urdan, T. (Eds.) *Self-Efficacy Beliefs of Adolsecents* (pp 1-43). Greenwich: Information Age Publishing.

Bandura, A. (2006b). Guide for creating self-efficacy scales. In Pajares, F. & Urdan, T. (Eds.) *Self-Efficacy Beliefs of Adolsecents* (pp 307-338). Greenwich: Information Age Publishing.

Bandura, A., & Schunk, D.H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology,* 41(3), 586-589.

Bartlett, M.S. (1954). A note on the multiplying factors for various chi square approximations. *Journal of the Royal Statistics Society*, 16 (series B), 296 – 298.

Bell, J.F. (2001). Investigating gender differences in the science performance of 16-year-old pupils in the UK. *International Journal of Science Education,* 23(5), 469-486.

Bereiter, C. & Scardamalia, M. (1987). *The Psychology of Written Composition.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Bransford, J.D., Brown, A.L., & Cocking, R.R. (2000). *How People Learn: Brain, Mind, Experience, and School*. Washington, DC: National Academy Press.

Brown, A. L., Bransford, J. D., Ferrara, R. A. & Campione, J. C. (1983). Learning, remembering, and understanding. In Flavell, J. H. & E. M. Markman. *Cognitive Development, Vol. 3* of P. H. Mussen (Ed.). *Handbook of child psychology, 4th Edition*. New York, NY: John Wiley & Sons.

Brown, A.L. & Campione, J.C. (1998). Designing a community of young learners: theoretical and practical lessons. In Lambert, N.M., & McCombs, B.L. (Eds.) *How Students Learn: Reforming Schools Through Learner-Centered Education* (pp 153-186). Washington, DC: American Psychological Association.

Center for Educational Policy Research. (2003). *Understanding University Success*. Eugene, OR: Center for Educational Policy Research. Full text available online at: http://www.s4s.org

Champagne, A. B., Kouba, V. L., & Hurley, M. (2000). Assessing inquiry. In J. Minstrell & E. H. van Zee (Eds.). *Inquiring into inquiry learning and teaching in science*. Washington, D.C.: Association for the Advancement of Science.

Coleman, E.B. (1998). Using explanatory knowledge during collaborative problem solving in science. *The Journal of the Learning Sciences*, 7(3&4), 387-427.

Davis, E.A. (2003). Prompting middle school science students for productive reflection: generic and directed prompts. *The Journal of the Learning Sciences,* 12(1), 91-142.

Dawson, C. (2000). Upper primary boys' and girls' interests in science: have they changed since 1980? *International Journal of Science Education,* 22(6), 557-570.

De La Paz, S. & Graham, S. (2002). Explicitly teaching strategies, skills, and knowledge: Writing instruction in middle school classrooms. *Journal of Educational Psychology*, 94(4), 687-689.

DeVellis, R.F. (2003). *Scale development: Theory and applications* (2^{nd} edn). Thousand Oaks, California: Sage.

Flick, L. (2000). Cognitive Scaffolding that Fosters Inquiry in Middle Level Science. *Journal of Science Teacher Education*, 11, 109-153.

Flick, L.B., & Tomlinson, M. (2006). Helping students understand the minds-on side of learning science. In McMahon, M., Simmons, P., Sommers, R., DeBaets, D. & Crawly, F. (Eds.) *Assessment in Science: Practical Experiences in Educational Research* (pp 183-196): Arlington, VA: NSTA Press.

Gaskill, P.J., & Woolfolk Hoy, A. (2002). Self-efficacy and self-regulated learning: the dynamic duo in school performance. In Aronson, J. (Ed.). *Improving Academic Achievement: Impact of Psychological Factors on Education* (pp 186-212). New York, NY: Academic Press.

Greenfield, T.A. (1997). Gender- and grade-level differences in science interest and participation. *Science Education*, 81(3), 259-276.

Hill A.G. (2000). Excellence in education versus high-stakes standardized testing. *Journal of Teacher Education*, 51(4), 293-304.

Hohenshell, L.M. & Hand, B. (2006). Writing-to-learn strategies in secondary school cell biology: A mixed method study. *International Journal of Science Education*, 28(2-3), 261-289.

Jacobs, J.E., Lanza, S., Osgood, D.W., Eccles, J.S., & Wigfield, A. (2002). Changed in children's self-competence and values: gender and domain differences across grades one through twelve. *Child Development*, 73(2), 509-527.

Kaiser, H. (1970). A second generation Little Jiffy. *Psychometrika*, 35, 401-415.

Kaiser, H. (1974). An index of factorial simplicity. *Psychometrika*, 39, 31-36.

Klein, P.D. (2006). The challenges of scientific literacy: From the viewpoint of second-generation cognitive science. *International Journal of Science Education*, 28(2-3), 143-178.

Kohn, A. (2000). Burn at the high stakes. *Journal of Teacher Education*, 51(4), 315-327.

Magnusson, S. J. (1996). Complexities of learning with computer-based tools: A case of inquiry about sound and music in elementary school. *Journal of Science Education and Technology*, 5(4), 297-309.

Magnusson, S.J., & Palincsar, A.S. (1995). The learning environment as a site of science education reform. *Theory Into Practice*, 34(1), 43-50.

Magnusson, S.J., & Palincsar, A.S. (2005). Teaching to promote the development of scientific knowledge and reasoning about light at the elementary school level. In Donovan, M.S, & Bransford, J.D. (Eds.). *How Students Learn: History, Mathematics, and Science in the Classroom*. Washington, DC: The National Academies Press.

Mayer, R.E. (2002). *The Promise of Educational Psychology Volume II: Teaching for Meaningful Learning*. Upper Saddle River, NJ: Pearson Education Inc.

McNeill, K. (2009). Teachers' use of curriculum to support students in writing scientific arguments to explain phenomena. *Science Education*, 93(2), 233-268.

Meece, J.L., & Painter, J. (2008). Gender, self-regulation, and motivation. In Schunk, D.H., & Zimmerman, B.J. (Eds.). *Motivation and Self-Regulated Learning: Theory, Research, and Applications* (pp 339-368). New York, NY: Lawrence Erlbaum Associates.

National Assessment of Educational Progress (2005). Retrieved April 9, 2008, from: http://nationsreportcard.gov/science_2005/s0101.asp?printver .

National Research Council. (1996). *National Science Education Standards*. National Committee on Science Education Standards and Assessment. Washington, DC: National Academy Press.

National Research Council (2000). *Inquiry and the National Science Education Standards*. National Committee on Science Education Standards and Assessment. Washington, DC: National Academy Press.

National Research Council. (2005). *How Students Learn: History, Mathematics, and Science in the Classroom.* Committee on *How People Learn*, A Targeted Report for Teachers, M.S. Donovan and J.D. Bransford, Editors. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Oregon Department of Education. (2007). Diploma requirements – get ready Oregon! Retrieved, April 11, 2008, from: http://www.getreadyoregon.org/newrequirements.aspx .

Oregon Department of Education. (2008). Scientific Inquiry Scoring Guides. Retrieved, July 22, 2008, from: http://www.ode.state.or.us/search/page/?id=1414 .

Page-Voth, V., & Graham, S. (1999). Effects of goal setting and strategy use on the writing performance and self-efficacy of students with writing and learning problems. *Journal of Educational Psychology,*91(2), 230-240.

Palincsar, A.S., & Brown, A.L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 1(2), 117-175.

Palincsar, A. S., Collins, K. M., Marano, N. L., & Magnusson, S. J. (2000). Investigating the engagement and learning of students with learning disabilities in guided inquiry science teaching. *Language, Speech, and Hearing Services in the Schools*, 31, 240-251.

Palincsar, A. S., Magnusson, S. J., Collins, K. M., & Cutter, J. (2001). Making science accessible to all: results of a design experiment in inclusive classrooms. *Learning Disabilities Quarterly*, 24(1), 15-32.

Pajares, F. (2002). Current directions in self-efficacy research. In Maehr, M. & Pintrich, P.R. (Eds.) *Advances in Motivation and Achievements*. Vol 10. Greenwich, CT: JAI Press.

Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: a review of the literature. *Reading and Writing Quarterly*, 19, 139-158.

Pajares, F. (2006). Self-efficacy during childhood and adolescence: implications for teachers and parents. In Pajares, F. & Urdan, T. (Eds.) *Self-Efficacy Beliefs of Adolsecents* (pp 339-367). Greenwich: Information Age Publishing.

Pajares (2007). *Overview of social cognitive theory and of self-efficacy*. Retreived December 5, 2007, from http://www.emory.edu/EDUCATION/mfp/eff.html.

Pajares, F. (2008). Motivational role of self-efficacy beliefs in self-regulated learning. In Schunk, D.H., & Zimmerman, B.J. (Eds.). *Motivation and Self-Regulated Learning: Theory, Research, and Applications* (pp 111-140). New York, NY: Lawrence Erlbaum Associates.

Pajares, F. & Urdan, T. (Eds.). (2006). *Self-Efficacy Beliefs of Adolescents*. Greenwich: Information Age Publishing.

Rosenthal, R. & Rosnow, R.L. (1975). *The volunteer subject*. New York, NY: John Wiley.

Rutherford, F.J. & Ahlgren, A. (1989). *Science For All Americans*. New York, NY: Oxford University Press.

Scardmalia, M. & Bereiter, C. (1985). Development of dialectical process in composition. In Olson, D.R., Torrance, N., & Hilyard, A. (Eds.). *Literacy, Language, and Learning: The Nature and Consequences of Reading and Writing* (pp 307-329). Cambridge, England: Cambridge University Press

Scardmalia, M., Bereiter, C. & Steinbach (1984). Teachability of the reflective process in written composition. *Cognitive Science*, 8, 173-190.

Schunk, D.H. (1996). Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal*, 33(2), 359-382.

Schunk, D.H. (2001). Social cognitive theory and self-regulated learning. In Zimmerman, B.J., & Schunk, D.H. (Eds.). *Self-Regulated Learning and Academic Achievement: Theoretical Perspectivest* (pp 125-151). New York, NY: Lawrence Erlbaum Associates.

Schunk, D.H., & Ertmer, P.A. (2000). Self-regulation and academic learning: self-efficacy enhancing interventions. In Boekaerts, M., Pintrich, P.R., & Zeidner, M. (Eds.) *Handbook of Self-Regulation* (pp 631-650). New York, NY: Academic Press.

Schunk, D.H., & Swartz, C.W. (1993). Goals and progress feedback: effects on self-efficacy and writing achievement. *Contemporary Educational Psychology*, 18, 337-354.

Schunk, D.H. & Zimmerman, B.J. (2007). Influencing children's self-efficacy and self-regulation of reading and writing through modeling. *Reading and Writing Quarterly*, 23, 7-25.

Shaw, K. (2004). *The Development of a Physics Self-Efficacy Instrument for Use in the Introductory Classroom*. AIP Conf. Proc. 720, 137, DOI:10.1063/1.1807273

Sloane, F.C. & Kelley, A.E. (2003). Issues in high-stakes testing programs. *Theory Into Practice*, 42(1), 12-18.

Smist, J.M. (1993, August). *General Chemistry and Self-Efficacy*. Paper presented at the National Meeting of the American Chemical Society, Chicago, IL.

Stewart, J., Cartier, J.L., & Passmore, C.M. (2005). Developing understanding through model-based inquiry. In Donovan, M.S, & Bransford, J.D. (Eds.). *How Students Learn: History, Mathematics, and Science in the Classroom*. Washington, DC: The National Academies Press.

The Ohio Core Curriculum: What's Required for Science? (2007). Retrieved April 11, 2008, from: http://www.ohiosci.org/TheOhioCoreCurriculumWhat'sRequiredforScience.pdf .

Tschannen-Moran, M., Woolfolk Hoy, A., & Hoy, W.K. (1998). Teacher efficacy: its meaning and measure. *Review of Educational Research*, 68(2), 202-248.

Warrick, P., Stephenson, P., & Webster, J. (2003). Developing pupils' written expression of procedural understanding through the use of writing frames in science: findings from a case study approach. *International Journal of Science Education*, 25(2), 173-192.

White, B.Y. & Frederiksen, J.R. (1998). Inquiry, modeling, and metacognition: making science accessible to all students. *Cognition and Instruction*, 16(1), 3-118.

Woolfolk Hoy, A. & Davis, H.A. (2006). Teacher self-efficacy and its influence on the achievement of adolescents. In Pajares, F. & Urdan, T. (Eds.) *Self-Efficacy Beliefs of Adolsecents* (pp 117-137). Greenwich: Information Age Publishing.

Yore, L.D., & Treagust, D.F. (2006). Current realities and future possibilities: Language and science literacy-empowering research and informing instruction. *International Journal of Science Education*, 28(2-3), 219-314.

Zimmerman, B.J. (1995). Self-efficacy and educational development. In Bandura, A. (Ed.) *Self-Efficacy in Changing Societies* (pp 202-231). Cambridge: Cambridge University Press.

Zimmerman, B.J. (2000). Attaining self-regulation: a social-cognitive perspective. In Boekaerts, M., Pintrich, P.R., and Zeidner, M. (Eds.), *Handbook of Self-Regulation* (pp 13-39). San Diego, CA: Adacemic Press.

Zimmerman, B.J., & Bandura, A. (1994). Impact of self-regulatory influences on writing course attainment. *American Educational Research Journal*, 31(4), 845-862.

Zimmerman, B.J., Bandura, A., & Martinez-Pons, M. (1992). Self-motivation for academic attainment: the role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal*, 29(3), 663-676.

Zimmerman, B.J., & Campillo, M. (2003). Motivating self-regulated problem solvers. In J.E. Davidson & R. Sternberg (Eds.), *The Psychology of Problem Solving*. New York: Cambridge University Press.

Zimmerman, B.J. & Cleary, T.J. (2006). Adolescents' development of personal agency. In Pajares, F. & Urdan, T. (Eds.) *Self-Efficacy Beliefs of Adolsecents* (pp 45 – 69). Greenwich: Information Age Publishing.

Zimmerman, B.J. & Kitsantas, A. (1997). Developmental phases in self-regulation: shifting from process goals to outcome goals. *Journal of Educational Psychology*, 89(1), 29-36.

Zimmerman, B.J. & Kitsantas, A. (1999). Acquiring writing revision skill: shifting from process to outcome self-regulatory goals. *Journal of Education Psychology*, 91(2), 241-250.

Zimmerman, B.J., & Martinez-Pons, M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal*, 23, 614-628.

Zimmerman, B.J., & Martinez-Pons, M. (1990). Student differences in self-regulated learning: relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology*, 82(1), 51-59.

Zimmerman, B.J., & Risemberg, R. (1997). Becoming a self-regulated writer: a social cognitive perspective. *Contemporary Educational Psychology*, 22, 73-101.

Zusho, A. & Pintrich, P.R. (2003). Skill and will: the role of motivation and cognition in the learning of college chemistry. *International Journal of Science Education*, 25(9), 1081-1094.

Appendix A: Study Instruments & Other Documents

Chemistry Self-Efficacy Survey

The purpose of this survey is to understand how you feel about your abilities in chemistry. There are no correct answers. Your answers will be kept confidential and you may choose to stop the survey at any time. Please respond to the prompts according to the following scale

A	B	C	D	E	F	G
Not confident		somewhat confident		quite confident		very confident

For each statement below, please rate how confident you feel that you could complete / do the following:

1. I can balance the following equations: $H_2 + O_2 \longrightarrow H_2O$ $KClO_3 \longrightarrow KCl + O_2$ $CuNO_3 + Mg \longrightarrow Mg(NO_3)_2 + Cu$	A B C D E F G
2. I can calculate the pH of a 0.500 M solution of H_2SO_4.	A B C D E F G
3. I can carry out an inquiry project that I have designed.	A B C D E F G
4. I can decide what type of graph to include that best represents my data.	A B C D E F G
5. I can design a data table to collect data that meets state standards (score at least a 4 according to the state scoring guide).	A B C D E F G
6. I can determine the number of moles and number of particles for the following: 40.0 g of KCl, 100.0 g $C_6H_{12}O_6$, 2.00 g Cu	A B C D E F G
7. I can develop a procedure for a scientific inquiry project such that it meets state standards (score at least a 4 according to the state scoring guide).	A B C D E F G
8. I can develop a question that could be investigated scientifically.	A B C D E F G
9. I can explain how the electrons act differently in the bonds that occur in the following molecules: LiBr, CaO, CH_4, CO_2.	A B C D E F G
10. I can explain two trends or patterns that are found in the periodic table as I move across a row or down a column.	A B C D E F G
11. I can find trends and patterns in data collected in a science inquiry project.	A B C D E F G
12. Given a periodic table, I can calculate the number of protons, electrons, and neutrons in a typical oxygen atom.	A B C D E F G
13. I can identify errors and their effect on data in a science inquiry procedure.	A B C D E F G
14. I can identify the following substances as acids or bases: HCl, HNO_3, $Mg(OH)_2$, KOH, H_2CO_3.	A B C D E F G
15. Given 20.0 g of H_2 and 20.0 g of O_2, I can calculate how many grams of water could I make and how much of the excess reactant would remain.	A B C D E F G
16. I can name the following compounds: Al_2O_3, KNO_3, P_2O_4, $HOCH_2CH_2CH_2CH_2OH$, $CH_3CH{=}CHCH_2CH_2CH_3$.	A B C D E F G

17. I can put enough background information in my inquiry project to meet state standards (score at least a 4 according to the state scoring guide).	A B C D E F G
18. I can successfully pass a science inquiry project with a score of at least 4 on each scored dimension (framing, design, collection, analysis) according to the state scoring guide.	A B C D E F G
19. I can use the ideal gas law to determine the volume taken up by 4 moles of H_2 gas at Standard Temperature and Pressure (STP)	A B C D E F G
20. I can write a conclusion to a science inquiry project that meets state standards (score at least a 4 according to the state scoring guide).	A B C D E F G
21. I can write a procedure for a scientific inquiry project such that it meets state standards (score at least a 4 according to the state scoring guide).	A B C D E F G

Inquiry Prompts

Phase I: For the initial phase of this study, the students were given the following information.

Determining the Percentage of Oxygen in a Chlorate

Background:

When a Chlorate is heated, it decomposes into a metallic chloride and oxygen gas. You can determine the percentage of oxygen in the compound by comparing the mass before and after heating. You will be using both potassium chlorate ($KClO_3$) and sodium chlorate ($NaClO_3$). Heating the chlorate in a Bunsen burner will drive off the oxygen and leave behind the resulting salt of either potassium chloride (KCl) or sodium chloride (NaCl). You will be given two compounds labeled “A” and “B”. Your task will be to design a procedure and then carry out an investigation to identify the unknown chlorates. Following the lab, you will need to type up a report of your investigation and findings (a guide to writing this project is on the back of this sheet). Here are the equations for the decomposition reactions that you’ll be working with in this lab.

$2\ KClO_3 + \text{heat} \rightarrow 2\ KCl + 3\ O_2$

$2\ NaClO_3 + \text{heat} \rightarrow 2\ NaCl + 3\ O_2$

The information below was on the back of the sheet that contained the prompt.

Writing the Inquiry Project
(to receive a 4)

Framing the Investigation:

- **Background knowledge**
  Talk about the reactions that took place in this lab (ie. what kind of reaction took place?, what is the balanced equation for this reaction?), what are some distinct properties of each compound, relevant information from the last unit of study.
- **Question or hypothesis**
  Pose a question or problem statement that this investigation will answer.
  Can the question or problem be answered?
  Does the background knowledge relate to the question / problem statement?

Designing the Investigation:

- **Logical and safe design**
  What did you use?
  How did you do it?
  Safety concerns?
- **Sufficient quality and quantity of data**
  How many trials did you perform?
  Is you data accurate/precise?
- **Detail**
  Could some one else perform your experiment based on your procedure?

Collecting and Presenting Data:

- **Data collection**
  Did you collect what your procedure stated that you would collect
- **Data transformation**
  Did you perform a calculation with your data? (ie. average, number of moles,…). If so, show example calculations.
- **Display (tables, charts, graphs…)**
  Are the neatly displayed?
  Are they correctly labeled (including correct units)

Analysis and Interpretation of Results:

- **Scientific terminology**
  Use science words (ie. mole, chemical formula, chemical reaction…)
  Identify trends or patterns that you see in your data
  Propose an explanation for the trends or patterns that you see
- **Review procedures**

What were some errors?
How would they have affected the experiment?

- **Support**
  Explicitly state how your results relate to your initial question / problem statement using some of your data.

Phases II & III: Students were not given any task prompts for either of these phases.

Writing Stems

Adapted from Bereiter & Scardamalia, 1987

New Idea

An even better idea is…
An important point I haven't considered yet is…
A better argument would be…
An important distinction is…
A consequence of (this is)…
The history of this is…
Something that is similar is…
Its features remind me of…
One thing that makes this different is…
A cause of this is…
A practical benefit of this is…

Improve

I could make my main point clearer by…
A criticism I should deal with in my report is…
I am getting off topic so…
This isn't very convincing because…
But many readers won't agree that…
I could describe this in more detail by adding…
To put it more simply…

Elaborate

An example of this is…
This is true, but it's not sufficient so…
The reason I think so…
Another reason that I think this is…
I could develop this idea by adding…
Another way to state this would be…
A good point on the other side of the argument is…
This results in…
My own experience with this is…

Goals

A goal for this investigation / report is…
My purpose is…

Putting it all together

My main point is…
I can tie this together by…

Student ID #__________________ Phase: II / III

Generic Reflection Prompts

There are no right or wrong answers to these prompts. I am interested in understanding what students think about as they work on inquiry projects. Feel free to use the back of this sheet or another sheet of paper if you need more space. Please hand this sheet in when you hand in your inquiry project.

1. My goal for the investigation phase of this project is…

2. As you begin planning your experiment, please finish this sentence, "Right now I am thinking about…"

3. At some point while you are doing the experiment, your teacher will ask you to finish this sentence, "Right now I am thinking about…"

---

4. My goal for the reporting phase of this project is…

5. As you begin planning your report, please finish this sentence, "Right now I am thinking about…"

6. At some point while you are doing the work on your report, your teacher will ask you to finish this sentence, "Right now I am thinking about…"

Student ID #___________________ Phase: II / III

Directed Reflection Prompts

There are no right or wrong answers to these prompts. I am interested in understanding what students think about as they work on inquiry projects. Feel free to use the back of this sheet or another sheet of paper if you need more space. Please hand this sheet in when you hand in your inquiry project.

1. My goal for the investigation phase of this project is…

2. As you begin planning your experiment, brainstorm a list of ideas, topics, and procedures that are relevant to this investigation.

3. At some point while you are doing the experiment, your teacher will ask you to respond to this question. "Does the data that you have collected so far make sense? Do you feel that any adjustments should be made to your procedure? Explain"

---

4. My goal for the reporting phase of this project is…

5. As you begin planning your report, brainstorm a list of ideas, topics, formulas, and vocabulary words that you think you might use as you write. Put a star next to the items that you think are the most important.

6. At some point while you are doing the work on your report, your teacher will ask you to respond to this question. "Does your report have a clear focus? What ideas do you still plan to include? Explain"

Student ID #___________________ Phase: II / III

Inquiry Project Self-Evaluation

This sheet will need to be filled out before you will have a chance to see your inquiry score. Using the Science Inquiry Scoring Guide, please assess your work on the four inquiry strands. Include a few sentences that explain why you gave yourself the score that you did.

	PREDICTED SCORE	ACTUAL SCORE
Forming a Question or Hypothesis:	_________	_________
Designing an Investigation:	_________	_________
Collecting and Presenting Data:	_________	_________
Analyzing and Interpreting Results:	_________	_________

Student ID #__________________ Phase: IV

Final Reflection

There are no right or wrong answers to these questions. These questions are a chance for you to let me know what sorts of things in this study were helpful to you and what sorts of things did not help you at all. As you answer, please be as specific as you can. To the extent that you can, use examples to help make your point. Feel free to use the back of this sheet or another sheet of paper if you need more space. Thank you for your help with this project!

1. Did your participation in this study cause you to think differently about scientific inquiry? Explain

2. What ideas / prompts / activities given in this study did you find to be useful in relation to scientific inquiry?

3. What ideas / prompts/ activities given in this study did you not find to be useful in relation to scientific inquiry?

4. Is there anything else that you would like to share about your involvement in this process?

Inquiry Scoring Guide

2002 - 2008 Official Scientific Inquiry Scoring Guide
CIM

	Forming a Question or Hypothesis *Based on observations and scientific concepts, ask questions or form hypotheses that can be answered or tested through scientific investigations.*	**Designing an Investigation** *Design a scientific investigation that provides sufficient data to answer a question or test a hypothesis.*	
6	A) Provides a focused rationale for the investigation by using the most relevant background science knowledge or preliminary observations. N) Forms a question or hypothesis that focuses and defines an investigation of scientific relationships (e.g., interaction, dependency, correlation, causation). C) Expresses question or hypothesis along with the application of background information clearly enough to suggest specific investigative procedures.	A) Applies knowledge of scientific research and procedures to create or adapt a design that is controlled, precise, safe, ethical, and consistent with accepted scientific practice. N) Presents a practical design that should provide reliable and valid data sufficient to answer the question or test the hypothesis and to explain the relationship(s). C) Communicates a unified (but flexible) design and logical, detailed procedures that can be fully replicated anticipating possible need for adjustment.	**6**
5	A) Provides background science knowledge or preliminary observations which are connected to the investigation. N) Forms a question or hypothesis that generally points toward an investigation of scientific relationships (e.g., interaction, dependency, correlation, causation). C) Expresses question or hypothesis along with the explanation of background information clearly enough to imply a particular investigative design.	A) Applies scientific knowledge to create or adapt a design with precise, safe, and ethical procedures. N) Presents a practical design that should provide data of sufficient quantity and quality to answer the question or test the hypothesis and investigate possible relationships. C) Communicates a unified design and logical, detailed procedures that can be reviewed, replicated, and revised.	**5**
4	A) Provides background science knowledge or preliminary observations that are relevant to the investigation. N) Forms a question or hypothesis that can be answered or tested using data gathered in a scientific investigation. C) Expresses question or hypothesis along with the explanation of background information clearly enough to imply an appropriate investigative approach.	A) Proposes scientifically logical, safe, and ethical procedures in a design with only minor scientific errors. N) Presents a practical design that should provide data of sufficient quantity and quality to answer the question or test the hypothesis (i.e., fair test). C) Communicates an organized design and procedures that have enough detail that they could be followed and revised.	**4**
3	A) Provides background science knowledge or preliminary observations that are either irrelevant or incomplete. N) Forms a question or hypothesis that can be investigated using data but not directly answered or tested. C) Expresses a question or hypothesis along with the explanation of background information that is understandable, but does not imply a direction for an investigation.	A) Proposes safe, ethical procedures in a design that contains some significant scientific errors. N) Presents a design that should provide relevant data but not sufficient to fully answer the question or test the hypothesis. C) Communicates a general plan and some procedures that can be followed.	**3**
2	A) Provides background science knowledge or preliminary observations that are inappropriate or substantially incorrect. N) Forms a question or hypothesis that cannot be investigated using data. C) Either question or hypothesis or background information is unclear.	A) Uses little scientific knowledge or does not consistently use reasonable, safe, or ethical procedures in a proposed design. N) Presents a design that should provide data somewhat applicable to the question or hypothesis. C) Communicates a summary of a plan that generally can be followed.	**2**
1	A) States a question or hypothesis without supporting background information. N) Forms a question or hypothesis that cannot be answered or tested. C) Background information is not included.	A) Uses minimal or incorrect scientific knowledge and unacceptable procedures in a proposed design. N) Presents a design that will not provide applicable data. C) Communicates a plan that is unclear or illogical.	**1**

[A= Application of Scientific Knowledge N= Nature of Scientific Inquiry C= Communication]

2002 - 2008 Official Scientific Inquiry Scoring Guide
CIM

	Collecting and Presenting Data *Collect, organize, and display sufficient data to facilitate scientific analysis and interpretation.*	**Analyzing and Interpreting Results** *Summarize and analyze data, evaluating sources of error or bias. Propose explanations that are supported by data and knowledge of scientific terminology.*	
6	A) Records accurate data consistent with complex procedures and deals with anomalous data, as needed. N) Transforms data into visually powerful displays/formats that clarify and highlight relationship(s) to be analyzed and explained. C) Creates precise and thorough displays (e.g., tables) for communicating observations or measurements, using appropriate units, in a logical and organized fashion.	A) Apply scientific terminology or notation correctly to analyze and explain relationship(s) investigated. N) Analyzes and critiques the design and procedures in light of the results and suggests insightful revisions or extensions. C) Explicitly analyzes the results of the investigation to support conclusions that address the question, hypothesis and relationship(s) investigated.	**6**
5	A) Records accurate data completely consistent with the planned procedure. N) Chooses data transformations that highlight information and patterns and support interpretation of relationships. C) Creates thorough displays (e.g., tables) for communicating observations or measurements, using appropriate units, that are logical and organized.	A) Uses scientific terminology or notation with minimal errors to report results, discuss relationships, and propose explanations. N) Provides evidence that the design, procedures, and results have been reviewed to identify important limitations and sources of error, suggesting design improvements when appropriate. C) Explicitly analyzes the results of the investigation to support conclusions that address the question or hypothesis and any relationships discovered.	**5**
4	A) Records reasonable data consistent with the planned procedure. N) Chooses data transformations that are valid and facilitate scientific analysis and interpretation. C) Creates displays (e.g., tables) for communicating observations or measurements, using appropriate units, that are logical and organized.	A) Uses scientific terminology with minimal errors to report results, identify patterns, and propose explanations. N) Provides evidence that the design, procedures, and results have been reviewed to identify some obvious limitations or sources of error. C) Explicitly uses the results of the investigation to support conclusions that address the question or hypothesis.	**4**
3	A) Records reasonable data consistent with the planned procedure with some obvious errors. N) Chooses data transformations that are valid and complete but do not facilitate scientific analysis and interpretation. C) Creates displays (e.g., tables) for communicating observations or measurements that are understandable, but somewhat incomplete or disorganized.	A) Uses scientific terminology, with some significant errors, to report results, identify patterns and propose explanations. N) Provides evidence that the design, procedures, and results have been reviewed but deals with errors and limitations in a trivial or illogical manner. C) Develops conclusions related to the question or hypothesis, but support from the investigation is either incomplete or not explicit.	**3**
2	A) Records data inconsistent with the planned procedure. N) Chooses data transformations that are sometimes invalid or incomplete. C) Creates displays (e.g., tables) for communicating observations or measurements that are substantially inaccurate, incomplete, or disorganized.	A) Uses scientific terminology incorrectly to report results, identify patterns or propose explanations. N) Provides minimal evidence that the design, procedures, and results have been reviewed. C) Presents interpretations or conclusions that are not clearly related to the question or hypothesis or supported by the results.	**2**
1	A) Records data unrelated to the planned procedure. N) Presents results in ways that are confusing or incorrect. C) Does not display data.	A) Does not clearly explain results or use scientific knowledge correctly. N) Does not provide evidence that the design or procedures have been reviewed. C) Does not present any interpretations.	**1**

[A= Application of Scientific Knowledge N= Nature of Scientific Inquiry C= Communication]

State of Oregon Anchor Paper #1

I. **Purpose**

The intent of this lab is to learn more about penny composition as it relates to the year the penny was manufactured. What are pennies made of, and have they always been made of the same substances? If not, in what year did they change? My hypothesis is that if I find the densities of the pennies, then I will know what year the materials were changed, because different substances have different densities.

II. **Background Information**

First of all, I know that matter is anything that has mass and volume. Mass is how much matter is in an object and volume is the amount of space an object occupies. Since scales depend on gravity to function, therefore measuring weight, one must use a balance to measure mass. Matter is classified by its properties. These properties can be extensive, meaning they depend on the quantity of matter present, or intensive, meaning they don't depend on the quantity of matter present. Extensive properties include mass, volume, and length while intensive are comprised of density, color, boiling point, and melting point, among others. Density is found by dividing mass by volume. By referencing the U.S. Mint online, I learned that before 1982, pennies were made of 95% copper and 5% zinc and had a mass of about 3.11 g. Since 1982, they have been made of about 99% zinc and around 1% copper and have a mass of about 2.5 g. The composition change occurred for the first time in 1982, but not all mints manufactured the newer pennies that year. The Compton's Encyclopedia says the melting point of zinc is 787° F (419° C) and calls it a bluish white metal. It gives zinc's density as 7.14 g/mL. Compton's says that copper melts at about 1083° C (about 1981° F) and has a density of 8.96 g/mL. The formula for calculating percent error is $(value_{accepted} - value_{measured}) / value_{accepted}$. When others melted pennies, they found a whitish, soft metal. Finally, I know that a Bunsen burner reaches about 1,000° C.

III. Procedure

A. Materials

1. Pennies (years 1979 – 1993)
2. Distilled water
3. (1) 50 mL graduated cylinder (smallest division – 1 mL)
4. (1) Electronic balance (smallest division - 0.01 g)

B. Technique

1. Found the mass of 10 pennies from 1979 using the electronic balance.
2. Found the volume of the same pennies by filling the graduated cylinder to 20 mL with distilled water, then sliding individual pennies into the GC on their edges (this is to minimize splashing). Read the volume, then subtracted the 20 mL of water from the total volume to find the volume of just the pennies.
3. Found the density of that set of pennies by dividing the mass by the volume.
4. Repeated steps 1-3 with a set of 10 pennies from 1980.
5. Found the mass, volume, and density of 10 pennies from 1981 using steps 1-3, then repeated those steps with two different sets of pennies from 1981.
6. Repeated step 5 for pennies from 1982 and 1983 (since 1982 is the year in question, I decided to obtain better accuracy in that year and the preceding and following year).
7. Found the mass, volume, and density of 1 penny each from 1984, '85, '86, '87, and '88 as a set using steps 1-3.
8. Found the mass, volume, and density of 1 penny each from 1989, '90, '91, '92, and '93 as a set using steps 1-3.
9. Compared the results with the densities of copper and zinc.
10. Calculated density by using the formula:
    Density = mass / volume
11. Calculated percentage of error for mass and density using background information and the formula:
    $(value_{accepted} - value_{measured}) / value_{accepted}$

IV. Data

Mass, Volume, and Density of Pennies

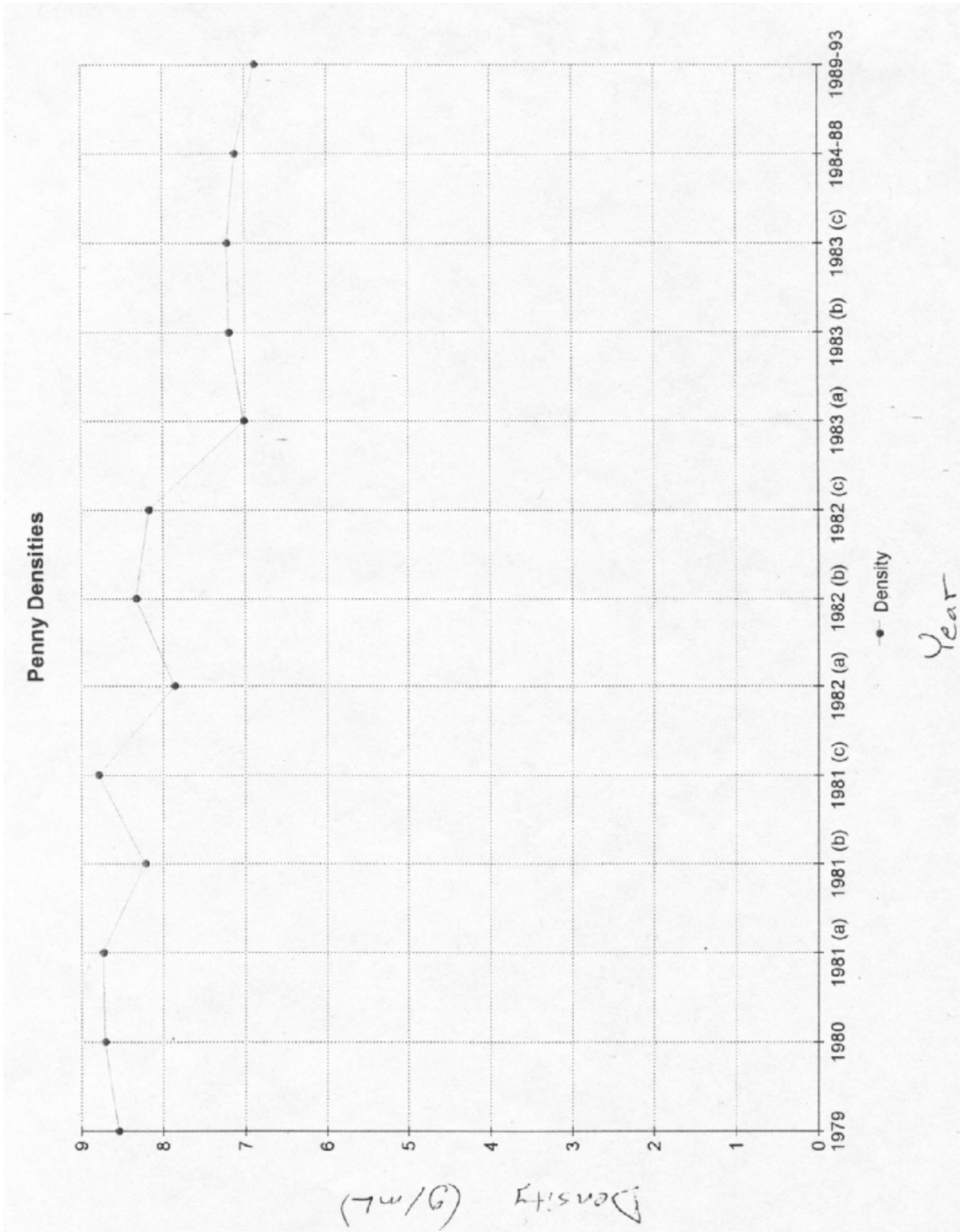
Year	Mass (g)	Volume (mL)	Density (g/mL)
1979	30.75	3.60	8.54
1980	30.74	3.53	8.71
1981	30.72 30.82 30.85	3.52 3.75 3.51	8.73 8.22 8.79
1982	29.49 29.14 28.59	3.75 3.50 3.50	7.86 8.33 8.17
1983	25.25 25.18 25.23	3.60 3.50 3.50	7.01 7.19 7.21
1984 – 1988	24.93	3.50	7.12
1989 – 1993	25.11	3.65	6.88

Melting Pennies
(Data provided by Period 1)

Year	Melted by Bunsen Burner?
1979	No
1980	No
1981	No
1982	No / Yes
1983	Yes
1984	Yes
1985	Yes
1986	Yes
1987	Yes
1988	Yes
1989	Yes
1990	Yes
1991	Yes
1992	Yes
1993	Yes

V. Analysis and Discussion

A. Describe



Leslie Phillips, Assessment Specialist
503-378-3600, Ext. 2317

Oregon Department of Education
Office of Assessment and Information Services

The "Mass, Volume, and Density" table shows that between 1979 and 1981, the densities were all around 8.5 g/mL. Then in 1982, the densities were more varied, ranging from 7.8 to 8.8 g/mL. Between 1983 and 1993, the densities were around 7 g/mL. The graph shows much the same thing – for 1979 – 1981 the densities were close to constant at ~8.5 g/mL. They then dip to around 8 g/mL in 1982, and finally level off at around 7 g/mL after 1983. The melting point table shows which years of pennies melted at around 1,000° C. Of these years, 1979-'81 didn't melt while some in '82 melted and some did not. After '82 all of the pennies melted.

B. Discuss

I was able to support my hypothesis because the data pointed to my conclusion: I will be able to find the year of change by using density. I learned that pennies have not always been made of the same substances, and that they changed in 1982. Finally, using intensive properties of matter, I was able to answer my question of what pennies are composed of.

C. Explain

I found that before 1982, pennies were made of almost pure copper. I know this from the densities found by myself, the inability of others to melt those pennies, and the color of the pennies. In 1982, some pennies were made of mostly copper like the years before, and some were made using other substances. I know this through research in the history of the U.S. Mint and from the fact that pennies from that year, when grouped, were in between they '79-'81 pennies and the '83-'93 pennies in terms of density. I found that after '82, all pennies were made of mostly zinc with a copper coating. This knowledge came from the density of zinc, the color of zinc, the melting point of zinc, the temperature of the Bunsen burner, and the research done on the Mint.

D. Errors

Percentage of Error

Year	Mass	Density
1979	1.125%	4.69%
1980	1.158%	2.79%
1981	1.222% 0.900% 0.804%	2.57% 8.26% 1.90%
1982	Incalculable (see analysis/discussion)	Incalculable
1983	1.000% 0.720% 0.920%	1.82% 0.70% 0.98%
1984-88	0.280%	0.28%
1989-93	0.440%	3.64%

One of the main errors occurs because in 1982 some mints manufactured the old pennies and some made the new pennies. Because of this, when I grouped 10 pennies together for mass and volume measurements, some where the old kind and some the new. This caused my density measurements to be in between the densities of the new and the old, and made calculations of percentage of error almost impossible. To find the numbers on the "Percent of Error" chart, I used the formula written in the background information. The table shows that the accuracy was pretty good for the penny masses and not bad for the densities. The densities were more off because the error from the mass was compounded with the error from the volume, making a larger error for density. One of the problems with measuring the volume is that the pennies caused the water in the graduated cylinder to splash and some of the water stuck to the sides of the cylinder. Another problem was that there weren't many graduations between measures on the cylinder, making it difficult to measure. Lastly, eyeballing the volume measurements is not much more than educated guesswork. To reduce error, I would use a graduated cylinder with more graduations, be more careful putting pennies in the cylinder, and find a set of pennies from 1982 that was entirely composed of the old pennies or entirely of the new.

E. Further Questions

I'm curious to know why the pennies were changed from mostly copper to mostly zinc in the first place. It would also be interesting to find out if pennies were made of other matters besides zinc and copper.

State of Oregon Anchor Paper #2

Penny Lab

In this lab we were asked to find which year the matter in pennies changed ,we were also asked to find the substance which it changed to if possible. This lab is used to see how accurate and precise we can be when conducting an experiment as well as seeing how we use the skills enabling us to properly identify a substance which we have acquired in the past two weeks.

We have access to several pennies from 1979 to 1993, these will be helpful in determining when the pennies began being made of a different material as well as helping us to find which material they were changed to. Since 1979 pennies have been made of at least two materials. Copper is one of the materials in pennies. Each element has its own properties which will help us determine the material. Melting point and density are very important in this lab, because they are intensive properties. Intensive properties are the properties which can if done right can properly and accurately identify a substance. Intensive properties show physical and chemical properties and they include density, melting point, and boiling point. Extensive properties which can also help to identify a material include mass, volume, and length these properties are useful but can not accurately identify a substance but ~~may be able to~~ verify the results of the intensive properties. We will use water distribution to determine the volume of the pennies and we will use the balances which go to .01 to find the mass. We will mainly use density to determine the substance. We have access to the busen burners which will be useful to determine the melting point. The bunsen burners temperature is around 1000 degrees C. We know that the measurements need to be very accurate and as precise as posible to come up with the correct answer. 7.14 is the density of zinc. 8.9 is the density of copper.

Procedure

Materials:

Electronic Balance
Water- distilled
Sets of pennies for years 1979-1993
Bunsen Burners and related equitment
50 mL graduated cylinder (GC)
Safety glasses

Steps:

1. Select 3 pennies from each year, keep in separate groups in order of weighing.
2. weigh the 3 pennies from each year on the balance (.01)

3. Record results and average the 3 pennies from each year.
4. Fill the GC with 10 mL of water
5. Add 6 pennies of the same year to the water.
6. Read from the <u>BOTTOM</u> of the meniscus where the line is at.
7. Do this for all 15 years and record how much volume was added to the water w/ the pennies.
8. take the amount of volume of the pennies and divide it by 6.
9. find the density of each year D= M/V M=the average of the 3 pennies mass, V= the amount of the volume of the pennies divided by 6
10. Look in a reference the see which elements come closest to your data to conclude which element the original element was changed to
11. look at your data to see which year you think the element changed
12. get on the internet and do a search to see when and to what the pennies were changed.

Year	Mass	Volume	Density
1979	3.07	0.33	9.3
1980	3.07	0.33	9.3
1981	3.1	0.16	8.8
1982	2.53	0.36	7.8
1983	2.49	0.33	7.7
1984	2.5	0.33	7.5
1885	2.5	0.33	7.6
1986	2.5	0.31	7.8
1987	2.52	0.31	7.6
1988	2.52	0.33	7.9
1989	2.52	0.35	7.2
1990	2.49	0.35	7.5
1991	2.51	0.16	7.2
1992	2.5	0.25	7.6
1993	2.52	0.33	7.6

Did the Pennies Melt? Y/N

Year	Yes	NO
1979		*
1980		*
1981		*
1982	*	*
1983	*	
1984	*	
1985	*	
1986	*	
1987	*	
1988	*	
1989	*	
1990	*	
1991	*	
1992	*	
1993	*	

Percent Error

Year	Percent
1979	4.5%
1980	4.5%
1981	1.1%
1982	9.2%
1983	7.8%
1984	5.0%
1985	6.4%
1986	9.2%
1987	6.4%
1988	10.6%
1989	.84%
1990	5.0%
1991	.84%
1992	6.4%
93	6.4%

In conclusion when I analyzed my data, my results showed that the pennies changed in 1982. I concluded that the material changed from Copper to copper plated Zinc. The densities I acquired were the most significant in concluding these educated guesses. The pennies from 1979-1982 came very close to the density of copper which is 8.92 g/cm3. In 1982the density was 7.8 right in the middle of copper and the other unknown substance. I then found that from 1983 – 1993 the densities were very close to each other when I researched my chemistry book to see what it was that the other material could be I found that it in my mind could have been Iron or Zinc; But when I found another group which had melted the pennies they said that all the pennies from 19983 on had melted, from my pervious knowledge I know that a Bunsen burner only got up to 1000-1200 degrees which is far to low for Iron to melt which made it clear to me that Zinc was the only logical solution.

The areas I need to improve on fro this lab to be truly successful is my accuracy with reading the bottom of the meniscus. I may have been able to do this if I had a graduated cylinder with more deviations on it this would have allowed me to read it closer to the actual milliliter. If I was ever to do this lab again I think I would use more samples in my volume because that is where I need the most samples to get an accurate reading/average. My percent errors weren't extremely high but I know with the previously mentioned things it could have been extremely lower.

All in all I think this lab was very useful. It in my mind was a total success it was put in place to have us use the tools we have recently acquired and put them to use in forming a puzzle which could be solved with the information we learned.

Appendix B: Data Analysis

Introduction

This study was conducted by looking at student work at three points in time over the course of a school year. The layout of this appendix follows the order of the surveys and tasks that the students experienced in this research project. The first section, *Initial Inquiry Task,* established a set of baseline data in order to compare the effects of the interventions that were assessed in this study. All of the students experienced similar conditions in this initial phase. The next two sections, *Second Inquiry Task* and *Third Inquiry Task*, looked at how student work differed in the separate experimental conditions. Since one group of students did not receive any interventions, this group provided a baseline in order to compare changes in the other groups. Each of the first three sections begins with an analysis of the *Chemistry Self-Efficacy Survey (CSES).* With each administration of the CSES, an assessment was run to evaluate the psychometrics of this scale. The final section of this appendix, *Final Assessment*, looked at the final administration of the CSES that the students took at the conclusion of the study.

Initial Inquiry Task – Baseline Study

Chemistry Self-Efficacy Survey (CSES)

1st Administration CSES - Validation

The study began with all of the students taking the Chemistry Self-Efficacy Survey (CSES). This survey took place approximately five weeks into the course and was given prior to any instruction on inquiry. Much of the first five weeks of the chemistry course was spent reviewing content from previous science courses. Time was also given to getting students familiar with the laboratory set-up of the chemistry room. All student responses on this initial administration of the CSES, then, primarily reflected beliefs and understandings gained from experiences in previous science classes. The survey was composed of two scales. One scale was composed of 10 items and assessed student self-efficacy related to chemistry content (Chemistry Content Self-Efficacy Scale) while the second scale was composed of 11 items and assessed student self-efficacy related to writing a science inquiry task (Science Inquiry Self-Efficacy Scale). An analysis of the Chemistry Content Self-Efficacy Scale yielded a Cronbach's $\alpha = 0.77$ with a mean inter-item correlation of 0.24 (range: -0.15 to 0.68). An analysis of the Science Inquiry Self-Efficacy Scale yielded a slightly higher Cronbach's $\alpha = 0.85$ with a mean inter-item correlation of 0.36 (range: 0.05 to 0.70).

All 21 items on this survey were subjected to principal component analysis (PCA). Prior to performing the principal component analysis, the data from the survey was analyzed to assess its suitability. The Kaiser-Meyer-Oklin value for the CSES was 0.523 which is slightly lower than the recommended value of 0.6 (Kaiser, 1970 & 1974). Bartlett's Test of Sphericity, though, (Bartlett, 1954) did reach statistical significance ($p < 0.0005$). An initial scree plot indicated that a two or three component solution would explain the greatest amount of the variance. Since the survey was designed to contain two factors, a two component solution was chosen. To aid in the interpretation of these two components, an oblimin rotation was performed.

A two component solution with an oblimin rotation explained 41.8% of the variance with the Science Inquiry Self-Efficacy Scale contributing 29.7% and the Chemistry Content Self-Efficacy Scale contributing 12.1%. The factors were only weakly correlated (r = 0.29). Ten of the eleven Science Inquiry items loaded on the first factor. Question 13 was the only item that loaded higher on the second factor. Both pattern coefficients and structure coefficients indicated, though, that question 13 actually loaded well on either factor. Only seven of the ten Chemistry Content items loaded on the second factor. In looking at the Chemistry Content items that did not load as well on the second factor, three questions (questions 1, 10, & 12) covered content that would have been covered in a freshman science course. These three questions loaded better on the Science Inquiry scale. In this initial survey, students generally felt more confident about their abilities in inquiry than in chemistry content. That these content items loaded better on the inquiry scale, then, is not surprising.

Table 22: Pattern and structure matrix for PCA with oblmin rotation of a two factor solution of CSES items from the initial administration.

Item	Pattern Coefficients		Structure Coefficients		Communalities
	Component 1	Component 2	Component 1	Component 2	
Question 1	**.465**	.224	**.529**	.356	0.326
Question 2	-.168	**.620**	.009	**.573**	0.354
Question 3	**.373**	.166	**.421**	.272	0.202
Question 4	**.317**	.294	**.401**	.384	0.240
Question 5	**.644**	-.309	**.556**	-.125	0.397
Question 6	.270	**.434**	.393	**.511**	0.328
Question 7	**.837**	-.147	**.795**	.091	0.653
Question 8	**.606**	.040	**.618**	.212	0.383
Question 9	.029	**.661**	.217	**.669**	0.449

Question 10	**.557**	.115	**.590**	.274	0.360
Question 11	**.595**	.147	**.637**	.316	0.425
Question 12	**.486**	.165	**.533**	.304	0.310
Question 13	.370	**.422**	.490	**.527**	0.404
Question 14	-.015	**.784**	.208	**.779**	0.608
Question 15	.260	**.265**	.335	**.339**	0.177
Question 16	.020	**.645**	.203	**.651**	0.424
Question 17	**.789**	-.177	**.738**	.048	0.574
Question 18	**.735**	-.098	**.707**	.111	0.509
Question 19	.008	**.689**	.204	**.691**	0.478
Question 20	**.835**	-.115	**.802**	.122	0.655
Question 21	**.640**	.198	**.696**	.380	0.521

Note: Major loadings for each item are in bold.
The Chemistry Content Scale contains questions: 1, 2, 6, 9, 10, 12, 14, 15, 16, 19
The Science Inquiry Scale contains questions: 3, 4, 5, 7, 8, 11, 13, 17, 18, 20, 21

1st Administration CSES – Student Score Analysis

The average scores from the first iteration of this survey were compared with a one-way ANOVA. The mean score for each class on the Science Inquiry Self-Efficacy scale ranged from a score of 5.4 to 5.7 on a 7-point Likert scale where a low value indicated low self-efficacy and a high value indicated high self-efficacy. In relation to Science Inquiry efficacy, there was no significant difference between the classes ($F(3,53) = 0.195$, $p = 0.90$). A similar analysis was conducted with the results from the Chemistry Content Self-Efficacy Scale. Mean scores ranged from 3.1 to 4.3 on a 7-point Likert scale where a low value indicated low self-efficacy and a high value indicated high self-efficacy. The initial one-way ANOVA yielded a Levene's Statistic = 3.69, $p = 0.017$. This implied that the variance between the classes was not the same. Because the variances were not the same, the more conservative Welch and Brown-Forsythe tests were used. Both tests indicated that there was a significant difference between the groups (Welch = 6.92, $p = 0.02$, Brown-Forsythe = 4.21, $p = 0.013$). Post-hoc comparisons using the Tukey HSD test indicated that the mean scores for period 4 ($M = 3.1$, $SD = 0.6$) were significantly lower than the mean scores for period 2 ($M = 4.2$, $SD = 1.2$) and period 5 ($M = 4.3$, $SD = 1.1$), $p = 0.03$ and $p = 0.01$ respectively.

Table 23: Summary results from the 1st administration of the Chemistry Self-Efficacy Survey

Period	n	Chemistry Content Self-Efficacy		Science Inquiry Self-Efficacy	
		Mean	SD	Mean	SD
2	9	4.2	1.2	5.6	0.9
4	15	3.1	0.6	5.5	0.8
5	14	4.3	1.1	5.7	1.0
6	19	3.9	1.0	5.4	

The mean scores for each efficacy question related to chemistry content were analyzed. The questions in the table are listed in the order that the topics identified in those questions were taught. For example, the content in survey question #6 on the survey covered content that was taught at the beginning of the school year whereas survey question #16 covered content that was taught at the end of the school year. The general trend from this first administration indicated that students generally felt more efficacious about the content that was covered earlier in the chemistry course and less efficacious about content that had not yet been covered. This result was not surprising.

Table 24: Summary results from the 1st administration of the Chemistry Self-Efficacy Survey of mean Chemistry Content Self-Efficacy by order taught and by class period

Period	Chemistry Self-Efficacy Survey – Chemistry Content Self Efficacy Questions									
	#6	#12	#10	#9	#1	#15	#19	#14	#2	#16
2	6.3	6.2	5.4	2.6	4.6	3.8	2.3	4.5	3.3	3.9
4	4.4	4.8	5.3	1.4	4.1	2.8	1.4	1.5	3.0	1.4
5	5.6	5.8	5.8	3.5	4.8	3.7	4.1	3.4	2.9	3.2
6	5.3	5.4	5.2	2.4	4.3	4.2	2.6	2.9	2.7	4.0

Student self-efficacy related to their abilities to complete specific aspects of the scientific inquiry task were also assessed. The State of Oregon Science Inquiry Scoring Guide broke down the write-up of the inquiry task into four phases. In the first phase (Framing), students wrote about what they intended to investigate and also included other background information that was relevant to their study. In the second phase (Designing), the students outlined their plan to collect data relevant to their investigation. In the third phase of writing (Collecting), the students presented their data and other graphic displays of their findings. Finally, in the last phase of the writing task (Analyzing & Concluding), the students presented their conclusions and assessed any limitations or errors that took place in the experiment. The self-efficacy survey contained between two and three questions for each of the four phases. Mean student scores for each phase were broken out by class in table 25. Four ANOVA's were run in order to compare the student mean self-efficacy scores within each phase.

These ANOVA's revealed that there were no significant differences by class period within each phase (Framing, $F(3,53) = 0.13$, $p = 0.95$; Designing, $F(3,53) = 0.04$, $p = 0.99$; Collecting, $F(3,53) = 0.34$, $p = 0.80$; Analyzing & Concluding, $F(3,53) = 1.37$, $p = 0.26$).

Table 25: Summary results from the 1st administration of the Chemistry Self-Efficacy Survey of mean Science Inquiry Self-Efficacy by inquiry category and period

Period	Framing Self-Efficacy		Designing Self-Efficacy		Collecting and Presenting Self-Efficacy		Analyzing & Concluding Self-Efficacy	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2	5.2	1.2	5.6	1.4	6.1	0.8	5.6	0.9
4	5.4	1.3	5.5	1.0	6.2	0.6	5.1	1.2
5	5.4	1.5	5.6	1.4	6.0	1.0	5.7	0.9
6	5.2	1.0	5.6	1.0	5.9	1.0	5.1	1.1

The four phases were then compared to each other using paired-samples t-tests. Student self-efficacy related to Framing ($M = 5.3$, $SD = 1.2$) was significantly lower than their self-efficacy for the Designing ($M = 5.6$, $SD = 1.1$) and Collecting ($M = 6.0$, $SD = 0.9$) phases ($t(56) = 2.29$, $p = 0.03$ (two-tailed) and $t(56) = 4.72$, $p < 0.0005$ (two-tailed) respectively). Student self-efficacy for Designing ($M = 5.6$, $SD = 1.1$) and Analyzing & Concluding ($M = 5.3$, $SD = 1.0$) were both significantly lower than student efficacy for the Collecting ($M = 6.0$, $SD = 0.9$) phase ($t(56) = 3.55$, $p = 0.001$ (two-tailed); $t(56) = 5.32$, $p < 0.0005$ (two-tailed) respectively). The differences between student self-efficacy for Framing ($M = 5.3$, $SD = 1.2$) and Analyzing & Concluding ($M = 5.3$, $SD = 1.0$) were not significant ($t(56) = 0.41$, $p = 0.68$ (two-tailed)) and the same was true for the differences in mean self-efficacy between the

Designing (M = 5.6, SD = 1.1) and Analyzing & Concluding (M = 5.3, SD = 1.0) phases (t(56) = 1.82, p = 0.08 (two-tailed). These results implied that students felt most confident in their abilities related to Designing and Collecting and less efficacious about their abilities related to Framing and Analyzing & Concluding.

Table 26: Summary results from the 1st administration of the Chemistry Self-Efficacy Survey of mean Science Inquiry Self-Efficacy by inquiry category.

Framing Self-Efficacy		Designing Self-Efficacy		Collecting and Presenting Self-Efficacy		Analyzing & Concluding Self-Efficacy	
Mean	SD	Mean	SD	Mean	SD	Mean	SD
5.3	1.2	5.6	1.1	6.0	0.9	5.3	1.0

1st Administration CSES – Effects of Gender

An independent samples t-test was run on the scores from the Chemistry Self-Efficacy Survey to see if there was a difference in efficacy by gender related either to Chemistry Content or Science Inquiry.

Table 27: Comparison of mean student scores in Chemistry Content Self-Efficacy and Science Inquiry Self-Efficacy by gender from the 1st administration of the CSES

	Chemistry Content Self-Efficacy		Science Inquiry Self-Efficacy	
	Mean	SD	Mean	SD
Male	3.9	0.9	5.6	0.8
Female	3.8	1.2	5.5	0.9

On Chemistry Content, males (M = 3.9, SD = 0.9) scored slightly higher than females (M = 3.8, SD = 1.2) but this difference was not significant; t(55) = -0.403, p = 0.69.

The same was true on the Science Inquiry scale where males (M = 5.6, SD = 0.8) scored slightly higher than the females (M = 5.5, SD = 0.9). Again, this difference was not significant; t(55) = -0.441, p = 0.66.

1st Administration CSES – Effects of Achievement

This study also sought to understand the relationship between student achievement and student beliefs about their abilities in science inquiry. For the purposes of this study, students were categorized as either high achieving or low achieving. Students were considered to be high achieving if their cumulative percent at the end of the first semester was 85% or higher whereas students who received an 84% or lower were categorized as low achieving. An independent samples t-test was run to compare the mean scores from the Chemistry Self-Efficacy Survey. Although in both cases, high achieving students (Chemistry Content M = 3.9, SD = 0.9; Science Inquiry M = 5.7, SD = 0.9) generally had higher mean self-efficacy scores than the lower achieving students (Chemistry Content M = 3.7, SD = 1.2; Science Inquiry M = 5.4, SD = 0.8), this difference was not significant for either factor (Chemistry Content t(55) = 0.947, p = 0.35 (two-tailed); Science Inquiry t(55) = 1.160, p = 0.25).

Table 28: Comparison of mean student scores in Chemistry Content Self-Efficacy and Science Inquiry Self-Efficacy by achievement from the 1st administration of the CSES

	Chemistry Content Self-Efficacy		Science Inquiry Self-Efficacy	
	Mean	SD	Mean	SD
High Achievers	3.9	0.9	5.7	0.9
Low Achievers	3.7	1.2	5.4	0.8

Initial Inquiry Task

Initial Inquiry Task Analysis – Comparison of Class Grades and State Test Scores

The initial base-line study took place about five weeks into the school year. Different class periods represented different comparison groups. No interventions were introduced during the first semester of the school year, so all students in this study experienced the same set of teaching conditions. In order to ensure that all of the class periods were statistically the same for later comparative purposes, a number of one-way ANOVA's were run in order to assess the extent to which this assumption was true. Because sample sizes were relatively small, with each ANOVA, a Levene's test for homogeneity of variance was run in order to insure that the variances within each group were similar.

A one way between-group analysis of variance was conducted in order to assess the distribution of student grades for each class period. This test was run using student percentages from the first semester only because the study itself did not take place until the beginning of the second semester. The mean percentage for each class ranged from 82.3% to 88.5%. For the chemistry course, an "A" was given to students who scored a 90% or higher whereas a "B" was for students whose cumulative point total was between an 80% and 90%. There was no statistically significant difference between the four classes in terms of student grades ($F(3,57) = 0.907$, $p = 0.44$). A similar comparison was conducted using scores from the Oregon Assessment of

Knowledge and Skills (OAKS) test in science. Class mean scores ranged from 243 to 248. Scores ranged from 217 and as high as 270. In order to meet State of Oregon benchmark standards, students must have earned a score of 240 or higher. There was no statistically significant difference in OAKS scores between the classes ($F(3,57) = 1.035$, $p = 0.38$).

Table 29: Summary results of mean semester 1 class percent and mean state test scores by period

Period	Percent		OAKS	
	Mean	SD	Mean	SD
2	84.4	11.0	244	8.5
4	82.4	9.1	243	9.6
5	88.5	8.0	248	7.1
6	82.3	16.5	246	9.5

Initial Inquiry Task Analysis – Comparison of Initial Inquiry Task Scores

Each of the classes were also compared on their initial abilities related to performing and then writing a science inquiry task. Again, the goal was to see if there were any significant differences between the classes. Scores on these inquiry tasks could range from 0 to 24 where a score of 16 was considered to meet the benchmark criteria. This initial task asked students to use their understanding of moles and percent composition to experimentally identify two different unknown samples. Overall, mean class scores on this task ranged from 11.9 to 12.8 and a one-way ANOVA indicated that there were no significant differences between the classes

(F(3,54) = 0.416, p = 0.742). Two more analyses were run that looked specifically at the written portions of the science inquiry task, namely the Framing section and the Analyzing & Concluding section. Each section was scored on a scale of 0 to 6 where a score of 4 or higher was required to meet benchmark standards. Mean class scores on the Framing section ranged from 2.9 to 3.1 and were not significantly different across the four classes (F(3,54) = 0.134, p = 0.94). The same was true when the mean scores for each class from the Analyzing & Concluding section were compared (F(3,54) = 0.391, p = 0.76). Mean class scores ranged from 2.6 to 2.9.

Table 30: Summary results of mean inquiry task scores and mean inquiry section task scores by period for the initial inquiry task

Period	Inquiry Task		Framing Section		Analyzing & Concluding Section	
	Mean Score	SD	Mean Score	SD	Mean Score	SD
2	11.9	3.2	2.9	1.2	2.6	1.3
4	11.9	2.8	3.1	0.7	2.8	0.6
5	12.4	2.2	3.1	0.5	2.8	0.6
6	12.8	2.8	3.1	0.9	2.9	1.0

Initial Inquiry Task Analysis – Score and Efficacy Comparison

The relationship between a student’s self-efficacy for science inquiry and their actual inquiry score was investigated using Pearson’s product-moment correlation coefficient. A preliminary analysis was performed in order to ensure that the assumptions of normality, linearity, and homoscedasticity were not violated. There was a moderate positive correlation between the two, r = 0.35, n = 54, p = 0.01 (two-tailed) which indicated that students with a greater sense of self-efficacy for science inquiry also tended to receive higher scores on their written inquiry tasks.

Initial Inquiry Task Analysis – Word Count Study

As student work was being assessed, it appeared that there might be a connection between the number of words in a section and a student's score in that section. An analysis was run on the data to see if there was a correlation between the number of words in a section and the score that a student received. This relationship was first investigated with the data from the Framing section of this initial inquiry using Pearson product-moment correlation coefficient. Preliminary analyses were performed in order to ensure that assumptions of normality, linearity, and homoscedasticity were not violated. There was a strong positive correlation between these two variables, $r = 0.70$, $n = 49$, $p < 0.0005$, where a high word count was associated with a high score.

The same analysis was run to see if there was a correlation between number of words in the Analyzing & Concluding section and a student's score for this initial inquiry task. Preliminary analyses were again run in order to ensure that Pearson product-moment correlation coefficient assumptions were not violated. This correlation also turned out to be positive and quite strong as well, $r = 0.66$, $n = 49$, $p < 0.0005$, where again, more words were associated with a higher score.

It also seemed, as the data was being collected, that the number of words a student used in the Framing section was often very close to the number of words used in their Analyzing & Concluding section. This correlation was also investigated using a Pearson product-moment correlation coefficient. A preliminary analysis of the data indicated that none of the assumptions were violated. In this case, the correlation between words in one section and words in the other was only moderate but significant, r = 0.41, n = 49, p = 0.003.

Table 31: Pearson product-moment correlations between numbers of words and section scores for the initial inquiry task

	1.	2.	3.	4.
1. Words in Framing Section		0.41**	0.70**	
2. Words in Analyzing & Concluding Section				0.66*
3. Framing Section Score				0.66**
4. Analyzing & Concluding Section Score				

* p < 0.005 (2-tailed)   ** p < 0.0005 (2-tailed)

Finally, this study looked at the correlation between scores in the Framing and Analyzing & Concluding sections. The correlation between section scores was large and significant, r = 0.66, n = 58, p < 0.0005. This indicated that students who tended to have high Framing section scores also had high Analyzing & Concluding scores.

Because there was such a strong correlation between word count and section scores, classes were compared on the number of words that were written in both the Framing and the Analyzing & Concluding section. In the Framing section, the mean class word count ranged from 120 words to 149 words. The differences were not statistically significant (F(3,45) = 0.831, p = 0.48). Class mean word counts from the Analyzing & Concluding section were slightly higher between 161 words and 223 words. Again, the mean differences between classes in word count for the Analyzing & Concluding section were not statistically significant ($F(3,45) = 0.409$, $p = 0.75$). These data support the claim that the four classes in this study were composed of a homogenous mix of students.

Table 32: Summary results of mean Framing word count and mean Analyzing & Concluding word count by period for the initial inquiry task

Period	Framing		Analyzing & Concluding	
	Mean # of Words	SD	Mean # of Words	SD
2	147	75	161	112
4	149	48	193	86
5	145	61	206	101
6	120	52	223	169

Initial Inquiry Task Analysis – Effects of Gender

An independent samples t-test was also conducted to compare the mean science inquiry scores for male and female students. Although males (M = 12.0, SD = 3.0) tended to score lower than females (M = 12.6, SD = 2.4), this difference was not significant; $t(56) = 0.755$, $p = 0.45$ (two-tailed). Independent samples t-tests were also run on student scores from the Framing and Analyzing & Concluding sections. On both sections, this trend continued where females (Framing M = 3.2, SD = 0.6; Analyzing & Concluding M = 2.9, SD = 0.8) tended to have higher scores than their male counterparts (Framing M = 3.0, SD = 1.0; Analyzing & Concluding M = 2.6 SD = 0.9). Again, neither of these differences were statistically significant (Framing:

t(56) = 0.918, P = 0.37 (two-tailed); Analyzing & Concluding: t(56) = 1.567, p = 0.12 (two-tailed)). Independent-samples t-tests were also used to compare word counts by section. In the Framing section, there were no significant differences between the mean number of words used by male (M = 135, SD = 58) and female (M = 140, SD = 57) students, t(47) = 0.25, p = 0.80 (two-tailed). There were also no significant differences in the mean number of words that were used by male (M = 165, SD = 86) and female (M = 225, SD = 140) students on the Analyzing & Concluding section, t(47) = 1.69, p = 0.10 (two-tailed).

Table 33: Comparison of mean student scores on Framing and Analyzing & Concluding sections on the initial inquiry task by gender

	Initial Inquiry Task		Framing				Analyzing & Concluding			
	Mean Score	SD	Mean Score	SD	Mean # of Words	SD	Mean Score	SD	Mean # of Words	SD
Male	12.0	3.0	3.0	1.0	135	58	2.6	0.9	165	86
Female	12.6	2.4	3.2	0.6	140	57	2.9	0.8	225	140

Initial Inquiry Task Analysis – Effects of Achievement

For this study, students were categorized as either high achieving or low achieving. Students were considered to be high achieving if their cumulative percent at the end of the first semester was 85% or higher whereas students who received an 84% or lower were categorized as low achieving. When high achieving students were compared to low achieving students using an independent samples t-test, the mean difference between high achievers (M = 13.4, SD = 2.0) and low achievers (M = 10.7, SD = 2.8) was significant; $t(56) = 4.23$, $p < 0.0005$. The magnitude of the differences in the means (mean difference = 2.68, 95% CI: 1.41 to 3.95) was quite large (eta squared = 0.24). High achieving students did significantly better on the Framing section (high achieving M = 3.4, SD = 0.4; low achieving M = 2.5, SD = 0.9), $t(56) = 4.93$, $p < 0.0005$ (two-tailed) and the Analyzing & Concluding section (high achieving M = 3.0, SD = 0.5; low achieving M = 2.4, SD = 1.0), $t(56) = 3.45$, $p = 0.001$ (two-tailed). Interestingly, the difference in mean number of words in the Framing section was not significant (high achieving M = 149, SD = 54; low achieving M = 120, SD = 59), $t(47) = 1.75$, $p = 0.09$ (two-tailed) whereas in the Analyzing & Concluding section, the difference in mean number of words was significant (high achieving M = 235, SD = 129; low achieving M = 144, SD = 93), $t(47) = 2.65$, $p = 0.01$ (two-tailed).

Table 34: Comparison of mean student scores on Framing and Analyzing & Concluding sections on the initial inquiry task by achievement

	Initial Inquiry		Framing				Analyzing & Concluding			
	Mean Score	SD	Mean Score	SD	Mean # of Words	SD	Mean Score	SD	Mean # of Words	SD
High Achievers	13.4	2.0	3.4	0.4	149	54	3.0	0.5	235	129
Low Achievers	10.7	2.8	2.5	0.9	120	59	2.4	1.0	144	93

Initial Inquiry Task Analysis – Summary

Taken together, the data from this baseline study provided strong evidence that the classes in this study were comparable in terms of student composition and ability levels. The only exception was related to the Chemistry Content Self-Efficacy of the students in period 4. It was evident though, that every class contained a similar range of ability levels. This provided further justification for using each individual classroom as the unit of analysis. This initial analysis also indicated that gender did not make any difference in relation to student scores. Male and female students felt equally efficacious in regards to their knowledge of chemistry content and their abilities in science inquiry. The same was true of males and females regarding their scores on the inquiry tasks.

Second Inquiry Task – Intervention Introduction

Chemistry Self-Efficacy Survey (CSES)

2nd Administration CSES - Validation

This phase took place approximately nineteen weeks into the school year. Before any interventions were introduced, the students were again asked to complete the Chemistry Self-Efficacy Survey (CSES). This was to see if student perceptions of their abilities related to chemistry content or scientific inquiry had changed since the initial administration of the survey. In the time between the initial and second administrations of the survey, the students had completed and received scores for one inquiry task and had also had been introduced to new chemistry content.

An analysis was conducted to again measure the psychometrics of the CSES. The Chemistry Content Self-Efficacy scale yielded a Cronbach's α coefficient of 0.80 with a mean inter-item correlation of 0.30 (range: -0.03 to 0.61). A similar analysis of the Science Inquiry Self-Efficacy scale yielded a Cronbach's α coefficient of 0.90

with a mean inter-item correlation of 0.45 (range: 0.13 to 0.70). Both of these Cronbach's α coefficients were slightly higher than the initial administration of this instrument (initial Chemistry Content Self-Efficacy Cronbach's α = 0.77; initial Science Inquiry Self-Efficacy Cronbach's α = 0.85).

The CSES was also analyzed to see if the initial determination of the factor structure was still valid. Again all the items on the survey were subjected to a principal component analysis (PCA). The Kaiser-Meyer-Oklin (KMO) value was 0.82 which was higher than the initial administration (KMO = 0.52) and exceeded the recommended value of 0.6 (Kaiser, 1970 & 1974).

Table 35: Pattern and structure matrix for PCA with oblmin rotation of a two factor solution of CSES items from the second administration

Item	Pattern Coefficients		Structure Coefficients		Communalities
	Component 1	Component 2	Component 1	Component 2	
Question 1	**.732**	.101	**.768**	.368	.599
Question 2	-.229	**.603**	-.010	**.519**	.315
Question 3	**.831**	-.178	**.766**	.125	.614
Question 4	**.452**	.299	**.561**	.464	.392
Question 5	**.824**	-.169	**.763**	.132	.606
Question 6	**.610**	.181	**.676**	.403	.486
Question 7	**.837**	-.173	**.774**	.132	.624
Question 8	**.575**	.216	**.654**	.426	.468
Question 9	**.446**	.409	**.595**	.572	.500
Question 10	.223	**.617**	.448	**.698**	.531
Question 11	.307	**.430**	.464	**.542**	.375
Question 12	.135	**.652**	.373	**.702**	.508
Question 13	**.627**	.152	**.682**	.380	.486
Question 14	-.015	**.645**	.220	**.640**	.409

Question 15	**.566**	.289	**.671**	.495	.522
Question 16	.133	**.472**	.305	**.521**	.286
Question 17	**.515**	.340	**.639**	.528	.508
Question 18	**.831**	-.135	**.782**	.168	.627
Question 19	**.519**	.247	**.609**	.436	.424
Question 20	**.854**	-.167	**.794**	.145	.654
Question 21	**.562**	.230	**.646**	.434	.463

Note: Major loadings for each item are in bold.
The Chemistry Content Scale contained questions: 1, 2, 6, 9, 10, 12, 14, 15, 16, 19
The Science Inquiry Scale contained questions: 3, 4, 5, 7, 8, 11, 13, 17, 18, 20, 21

Bartlett's Test of Sphericity (Bartlett, 1954) reached significance ($p < 0.0005$) and an initial analysis of the scree plot indicated that almost all of the survey variance could be explained by a two component solution. To aid in the interpretation of these two components, an oblimin rotation was performed.

A two-component solution explained 49.5% of the variance with the Science Inquiry Self-Efficacy scale contributing 40.4% and the Chemistry Content Self-Efficacy scale contributing 9.1%. These two factors were only weakly correlated ($r = 0.36$). With this second administration, the two factor solution seemed to do a better job of explaining the variance (the initial administration two factor solution explained 41.8 % of the variance), although, the factor separation was not as clean as the initial administration. Though many of the items loaded onto both factors, ten of the eleven items on the Science Inquiry Self-Efficacy scale had their highest loadings on component 1. The only exception was item 11 which asked students about their beliefs in their ability to interpret data. On this administration of the CSES, Science

Inquiry Self-Efficacy scores were generally higher than those for Chemistry Content Self-Efficacy. That item 11 loaded better on the Chemistry Content scale implied that students did not feel as efficacious about their abilities to interpret data when compared to other parts of the inquiry task.

On the Chemistry Content Self-Efficacy Scale, only six out of the ten items loaded best on component 2. Three of the items in question (questions 6, 15, & 19) asked students about content that we had either just covered or were in the process of covering. Because of the proximity between when this content was taught and the administration of this survey, it was not surprising that the scores on these items were slightly higher and therefore loaded better on the Science Inquiry scale. In looking at both the pattern and the structure coefficients, these items actually loaded well on both scales. The other item that did not load as well on the Chemistry Content scale was a question (question 1) that asked students about content that many had covered repeatedly since middle school. That this item loaded better on the Science Inquiry scale is also not surprising. Again, in looking at the pattern and structure coefficients, even though this question loaded much better on the Science Inquiry factor, it also loaded well on the Chemistry Content factor. It was anticipated that students' efficacy on both scales should increase over the course of the school year, so finding items that loaded well on both scales was not surprising.

2nd Administration CSES – Student Score Analysis

Class averages were compared with the results from the second administration of the CSES with one-way ANOVA's. Class means in Chemistry Content Self-Efficacy ranged from 4.3 to 5.1 on a 7-point Likert scale where a low value indicated low self-efficacy and a high value indicated high self-efficacy. There was no statistically significant difference between the classes ($F(3,53) = 1.956$, $p = 0.13$). A similar analysis was done with the results of the CSES with mean Scientific Inquiry Self-Efficacy scores. When a one-way ANOVA was run, the results were as follows. Class means ranged between 4.8 and 5.2 where a low value indicated low self-efficacy and a high value indicated high self-efficacy. Similar to the Chemistry Content Self-Efficacy, there were no significant differences between the classes ($F(3,53) = 1.263$, $p = 0.30$).

Table 36: Summary results from the 2nd administration of the Chemistry Self-Efficacy Survey of mean Chemistry Content Self-Efficacy and mean Science Inquiry Self-Efficacy by period

Period	Chemistry Content Self-Efficacy		Science Inquiry Self-Efficacy	
	Mean	SD	Mean	SD
2	4.9	0.8	4.8	1.2
4	4.3	0.8	5.0	1.0
5	5.1	1.0	5.6	0.8
6	4.3	1.3	5.2	1.3

Similar to the first administration of the CSES, student efficacy for Chemistry content was analyzed to see how a student's responses to individual questions related to the content that had already been covered (table 37). With the exception of question #9, student self-efficacy seemed to be the greatest in the first half of the survey and then declined towards the second half of the survey. Since the initial questions referred to content that had already been covered and the latter half referred to content that had not yet been covered, this was expected.

Table 37: Summary results from the 2nd administration of the Chemistry Self-Efficacy Survey of mean Chemistry Content Self-Efficacy by order taught and by class period

Period	Chemistry Self-Efficacy Survey – Chemistry Content Self Efficacy Questions									
	#6	#12	#10	#9	#1	#15	#19	#14	#2	#16
2	5.3	6.7	5.3	3.9	5.4	4.1	4.8	4.1	3.7	5.6
4	4.4	5.9	5.9	3.1	5.2	4.6	3.5	4.3	2.7	3.7
5	4.4	6.1	5.6	3.8	5.0	4.4	4.6	2.4	2.2	4.5
6	4.8	6.1	5.8	3.8	5.3	4.5	4.6	3.8	2.6	4.7

An analysis of student self-efficacy for the four inquiry phases of the written product was also conducted. Four ANOVA's were run in order to assess the extent to which student self-efficacy scores in each phase varied by class period. All of the ANOVA's indicated that there were no significant differences between the class periods related to self efficacy scores in each phase (Framing, $F(3,53) = 2.32$, $p = 0.09$; Designing, $F(3,53) = 0.87$, $p = 0.46$; Collecting, $F(3,53) = 0.17$, $p = 0.92$; Analyzing & Concluding, $F(3,53) = 1.84$, $p = 0.15$).

Table 38: Summary results from the 2nd administration of the Chemistry Self-Efficacy Survey of mean Science Inquiry Self-Efficacy by inquiry category and period.

Period	Framing Self-Efficacy		Designing Self-Efficacy		Collecting and Presenting Self-Efficacy		Analyzing & Concluding Self-Efficacy	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2	4.7	1.1	4.8	2.0	5.3	1.5	4.8	1.0

4	4.7	1.0	5.2	1.3	5.5	1.2	4.9	1.2
5	5.8	1.1	5.7	0.8	5.5	1.5	5.0	0.7
6	5.3	1.6	5.4	1.6	5.5	1.4	5.1	1.3

Since there were no significant differences between the classes within each phase, the scores across all the classes were combined by each phase in order to determine if there were differences between student efficacy for each of the phases. The phases were compared using paired-samples t-tests. The only comparison of the six that yielded a significant results was the comparison between student self-efficacy for Collecting data ($M = 5.5$, $SD = 1.4$) and student self-efficacy for Analyzing & Concluding ($M = 5.1$, $SD = 1.1$), $t(56) = 2.80$, $p = 0.007$ (two-tailed). Of the remaining comparisons, two comparisons approached significance (Framing ($M = 5.2$, $SD = 1.3$) and Collecting ($M = 5.5$, $SD = 1.4$), $t(56) = 1.82$, $p = 0.07$ (two-tailed); Designing ($M = 5.4$, $SD = 1.4$) and Analyzing & Concluding ($M = 5.1$, $SD = 1.1$), $t(56) = 1.90$, $p = 0.06$(two-tailed)) and three comparisons were essentially equivalent: Framing ($M = 5.2$, $SD = 1.3$) and Designing ($M = 5.4$, $SD = 1.4$), $t(56) = 1.04$, $p = 0.30$ (two-tailed); Framing ($M = 5.2$, $SD = 1.3$) and Analyzing & Concluding ($M = 5.1$, $SD = 1.1$), $t(56) = 0.74$, $p = 0.47$ (two-tailed); Designing ($M = 5.4$, $SD = 1.4$) and Collecting ($M = 5.5$, $SD = 1.4$), $t(56) = 1.07$, $p = 0.29$ (two-tailed)). These results again support the idea that students feel most confident in their abilities to collect and present data.

Table 39: Summary results from the 2nd administration of the Chemistry Self-Efficacy Survey of mean Science Inquiry Self-Efficacy by inquiry category

Framing Self-Efficacy		Designing Self-Efficacy		Collecting and Presenting Self-Efficacy		Analyzing & Concluding Self-Efficacy	
Mean	SD	Mean	SD	Mean	SD	Mean	SD
5.2	1.3	5.4	1.4	5.5	1.4	5.1	1.1

2nd Administration CSES – Effects of Gender

CSES survey results were then assessed with an independent-samples t-test in order to determine if there were any differences due to gender. On the Chemistry Content Self-Efficacy scale, males (M = 4.6, SD = 1.4) and females (M = 4.6, SD = 0.8) were virtually identical; $t(55) = -0.058$, $p = 0.95$ (two-tailed). In contrast, on the Science Inquiry Self-Efficacy Scale, female student scores (M = 5.3, SD = 0.9) were generally higher than their male counterparts' scores (M = 5.1, SD = 1.4). This difference, though, was not statistically significant; $t(55) = 0.495$, $p = 0.62$ (two-tailed).

Table 40: Comparison of mean student scores in Chemistry Content Self-Efficacy and Science Inquiry Self-Efficacy by gender from the 2nd administration of the CSES

	Chemistry Content Self-Efficacy		Science Inquiry Self-Efficacy	
	Mean	SD	Mean	SD
Male	4.6	1.4	5.3	0.9
Female	4.6	0.8	5.1	1.4

2nd Administration CSES – Effects of Achievement

Student scores on the CSES were also broken down by achievement. Again, students with a grade percentage of 85% or higher from their first semester in chemistry were counted as high achievers for the purposes of this study. Any student with a grade percentage of 84% or lower was considered to be a lower achiever. When mean scores from the Chemistry Content scale were compared, high achievers (M = 4.9, SD = 1.0) scored significantly higher than the lower achieving students (M = 4.2. SD = 1.1); t(55) = 2.53, p = 0.014 (two-tailed). The magnitude of the difference between these means (mean difference = 0.71, 95% CI: 0.146 to 1.271) was moderate (eta squared = 0.10). The mean scores for the Science Inquiry scale also indicated that high achievers (M = 5.3, SD = 1.2) tended to have higher scores than low achieving students (M = 5.0, SD = 1.0). The independent-samples t-test implied, though, that this difference was not significant; t(55) = 1.00, p = 0.32.

Table 41: Comparison of mean student scores in Chemistry Content Self-Efficacy and Science Inquiry Self-Efficacy by achievement from the 2nd administration of the CSES

	Chemistry Content Self-Efficacy		Science Inquiry Self-Efficacy	
	Mean	SD	Mean	SD
High Achievers	4.9	1.0	5.3	1.2
Low Achievers	4.2	1.1	5.0	1.0

2nd Administration CSES – Comparison to Baseline: Scale Scores

A one-way repeated measures ANOVA was also conducted on the data to compare Chemistry Content Self-Efficacy scores back to baseline data. The means and standard deviations are presented in table 42. There was a significant increase in mean Chemistry Content Self-Efficacy, Wilks' Lambda = 0.72, $F(1,52) = 20.53$, $p < 0.0005$. The multivariate partial eta squared = 0.28 which indicated a large effect size.

A one-way repeated measures ANOVA was also conducted in order to compare the Science Inquiry Self-Efficacy scores back to baseline data. While there was a slight drop in the mean Science Inquiry Self-Efficacy, this drop was not statistically significant (Wilks' Lambda = 0.944, $F(1,52) = 3.083$, $p = 0.085$). The partial eta squared = 0.06 indicated a moderate effect size. An effect size this large was somewhat surprising because the difference in means over time was not statistically significant, although the difference was approaching significance ($p = 0.085$).

Table 42: Comparison of mean student Chemistry Content Self-Efficacy scores from the 2nd administration back to baseline data

	N	Mean	SD
Baseline	53	3.9	1.1
2nd Administration	53	4.6	1.1

Table 43: Comparison of mean student Science Inquiry Self-Efficacy scores from the 2nd administration back to baseline data

	N	Mean	SD
Baseline	53	5.6	0.9
2nd Administration	53	5.3	1.1

2nd Administration CSES – Comparison to Baseline: Gender

A mixed between-within subjects ANOVA was run in order to compare male and female scores on the Chemistry Content scale back to the baseline data that was initially collected for this survey. There was no significant interaction between gender and time, Wilks' Lambda = 0.999, $F(1,51) = 0.037$, $p = 0.85$. As was shown in the previous section, there was a significant main effect for time, Wilks' Lambda = 0.726, $F(1,51) = 19.283$, $p < 0.0005$, partial eta squared = 0.27, with both males and females mean Chemistry Content Self-Efficacy scores showing improvement. The main effect comparing males and female scores was not significant, $F(1,51) = 0.037$, $p = 0.85$.

A similar comparison was run on the mean score data for the Science Inquiry scale. Again there was no significant interaction between gender and time, Wilks' lambda = 1.00, $F(1,51) = 0.022$, $p = 0.88$. And while the effect of time approached significance, (Wilks' Lambda = 0.944, $F(1,51) = 3.024$, $p = 0.088$, partial eta squared = 0.056) the main effect for gender was not significant, $F(1,51) = 0.051$, $p = 0.82$ which implied that there was no real difference between the mean Science Inquiry Scale scores of male and female students over time.

Table 44: Comparison of mean Chemistry Content Self-Efficacy scores from the 2nd administration to baseline data by gender

	Male – Chemistry Content			Female - Chemistry Content		
	n	Mean	SD	n	Mean	SD
Baseline	22	3.9	0.9	31	3.8	1.2
2nd Administration	22	4.6	1.4	31	4.6	0.8

Table 45: Comparison of mean Science Inquiry Self-Efficacy scores from the 2nd administration to baseline data by gender

	Male – Science Inquiry			Female – Science Inquiry		
	n	Mean	SD	n	Mean	SD
Baseline	22	5.6	0.8	31	5.5	0.9
2nd Administration	22	5.3	1.3	31	5.3	0.9

2nd Administration CSES – Comparison to Baseline: Achievement

A comparison was also run in order to look at how the scores of high achieving students compared to scores of lower achieving students. A mixed between-within subjects ANOVA was run in order to compare mean scores back to baseline data on the Chemistry Content scale. The interaction between time of survey administration and achievement was not significant, (Wilks' Lambda = 0.97, $F(1,51) = 1.63$, $p = 0.20$) whereas the main effect for time was significant (Wilks' Lambda = 0.74, $F(1,51) = 17.59$, $p < 0.0005$, partial eta squared = 0.26) and both groups showed an increase in mean Chemistry Content Self-Efficacy over time. The main effect that compared the mean scores of high and low achievers over time was also significant, $F(1,51) = 4.09$, $p = 0.05$, partial eta squared = 0.07, which indicated that the efficacy gains of the higher achievers was significantly greater than efficacy gains of the lower achieving students.

A similar comparison was run on the mean Science Inquiry data. Again, the interaction between time and achievement was not significant (Wilks' Lambda = 1.00, $F(1,51) = 0.00$, $p = 0.98$) nor was the main effect for time (Wilks' Lambda = 0.95, F(1,51) = 2.91, p = 0.09, partial eta squared = 0.05). The main effect for achievement approached significance, F(1,51) = 3.84, p = 0.06, partial eta squared = 0.07, but there was no real difference between the student mean scores on the Science Inquiry scale over time.

Table 46: Comparison of mean Chemistry Content Self-Efficacy Scale scores from the 2nd administration to baseline data by achievement

	High Achievers - Chemistry Content			Low Achievers - Chemistry Content		
	n	Mean	SD	n	Mean	SD
Baseline	32	4.0	0.9	21	3.7	1.3
2nd Administration	32	4.9	1.0	21	4.2	1.2

Table 47: Comparison of mean Science Inquiry Self-Efficacy Scale scores from the 2nd administration to baseline data by achievement

	High Achievers - Science Inquiry			Low Achievers - Science Inquiry		
	n	Mean	SD	n	Mean	SD
Baseline	32	5.7	0.9	21	5.3	0.8
2nd Administration	32	5.4	1.1	21	5.0	1.0

Second Inquiry Task

2nd Inquiry Task Analysis – Comparison of Second Inquiry Task Scores

The interventions in this phase began after the second administration of the CSES. In the three classes that were part of the experimental condition, about 3 weeks were spent introducing students to a model of the inquiry process and how this related to the process of writing. In addition, all of the classes in the experimental condition were given sentence stems to use in the Framing and Analyzing & Concluding sections. In this inquiry task, students were asked to use their knowledge of gas law chemistry to experimentally determine the molar mass of a gas that was inside of a cigarette lighter. Using this molar mass data, students were asked to compare their experimental values with molar mass values for known gasses. Students were then asked to make a conclusion about the identity of their gas and defend their choice using their data.

Table 48: Summary results of mean inquiry task scores and mean inquiry section task scores by period for the 2nd inquiry task

Period	Inquiry Task		Framing		Analyzing & Concluding	
	Mean Score	SD	Mean Score	SD	Mean Score	SD
2	13.4	1.3	3.3	0.5	3.1	0.4
4	13.6	3.0	3.4	0.7	3.1	1.1
5	13.7	1.8	3.3	0.6	3.0	1.0
6	13.0	2.8	3.1	1.0	2.7	1.2

Within the experimental condition, two out of the three classes were also asked to reflect at various points in time during the actual inquiry task and the subsequent writing. One group was asked to reflect using questions that had been written up in advance while the other group was asked to respond to a more generic prompt, “right now I am thinking...” Responding to a specific prompt was called “Directed” reflection whereas responding to the more generic prompt was called “General” reflection. Treatments differed by class and each class was assigned the following condition: period 2 – control group; period 4 – instruction in inquiry and writing but no reflection; period 5 – instruction in inquiry and writing with directed reflection; period 6 – instruction in inquiry and writing with general reflection.

A one-way ANOVA was used to compare class means on overall inquiry project scores in addition to the scores on the Framing and Analyzing & Concluding sections. Overall class mean scores on this second task ranged from 13.0 to 13.7 where scores could have ranged from 0 to 24 and a 16 was considered a passing mark. There was no statistical difference between any of the classes ($F(3,53) = 0.253$, $p = 0.86$). An independent-samples t-test also confirmed that there was no significant difference between the mean scores of students in the experimental group ($M = 13.4$, $SD = 2.6$) and those in the control group ($M = 13.4$, $SD = 1.3$); $t(55) = 0.035$, $p = 0.97$. A follow-up one way ANOVA compared mean scores on this second inquiry task between students in the control group (per2), students in the experimental group who were not asked to reflect (per 4) , and students who were asked to reflect (per 5 & 6). This test revealed that there were no significant differences between the mean student scores ($F(2,54) = 0.09$, $p = 0.91$).

When the class means for Framing scores were analyzed with a one-way ANOVA, the means ranged from 3.1 to 3.4 out of a scale of 6 possible points where a score of 4 was considered passing. These means were not significantly different ($F(3,53) = 0.314$), $p = 0.82$). Class means for the Analyzing & Concluding sections ranged from 2.7 to 3.1 out of a scale of 6 possible points where a score of 4 was considered passing. Again, these means were not significantly different ($F(3,53) = 0.512$, $p = 0.68$).

2nd Inquiry Task Analysis – Score and Efficacy Comparison

The relationship between a student's self-efficacy for science inquiry and their actual inquiry score was investigated using Pearson's product-moment correlation coefficient. A preliminary analysis was performed in order to ensure that the assumptions of normality, linearity, and homoscedasticity were not violated. There was a moderate positive correlation between the two, $r = 0.38$, $n = 53$, $p = 0.005$ (two-tailed) which indicated that students with a greater sense of self-efficacy for science inquiry also tended to receive higher scores on their written inquiry tasks. In the initial inquiry task, the correlation coefficient was 0.35, so there was a slight increase in the strength of the correlation between a student's sense of efficacy and their subsequent inquiry task score.

2nd Inquiry Task Analysis – Word Count Study

Word counts were analyzed with one-way ANOVA's to see if there were any differences in the number of words that students wrote in the Framing and Analyzing & Concluding sections. In the Framing section, class mean word counts ranged from 200 to 240 words. In the Analyzing & Concluding section, mean word counts were between 228 and 240 words. For both of these analyses, the mean differences were not significant ($F(3,50) = 0.389$, $p = 0.76$; $F(3,50) = 0.023$, $p = 0.99$ respectively) which meant that the general number of words that students used did not vary much from class to class.

The relationship between word count and score was investigated for both the Framing and the Analyzing & Concluding sections using a Pearson product-moment correlation coefficient. An initial scatter plot analysis ensured that no assumptions regarding normality, linearity, and homodescedasticity were violated. For both sections, there was a strong positive correlation (Framing: $r = 0.57$, $n = 54$, $p < 0.0005$; Analyzing & Concluding: $r = 0.65$, $n = 54$, $p < 0.0005$) between the number of words used and the resulting score. A follow-up analysis using a Pearson product-moment correlation coefficient indicated that there was a significant and strong relationship between the number of words a student used in the Framing section and the number of words that the student used in the Analyzing & Concluding section ($r = 0.60$, $n = 54$, $p < 0.0005$). The same was also true of the correlation between the section scores ($r = 0.81$, $n = 57$, $p < 0.0005$).

Table 49: Summary results of mean Framing word count and mean Analyzing & Concluding word count by period for the 2nd inquiry task

Period	Framing - Word Count		Analyzing & Concluding - Word Count	
	Mean # of Words	SD	Mean # of Words	SD
2	203	94	241	74
4	210	123	231	127
5	200	79	228	125
6	240	138	235	126

Table 50: Pearson product-moment correlations between numbers of words and section scores for the 2nd inquiry task

	1.	2.	3.	4.
1. Words in Framing Section		0.60**	0.57**	
2. Words in Analyzing & Concluding Section				0.65**
3. Framing Section Score				0.81**
4. Analyzing & Concluding Section Score				

** $p < 0.0005$ (2-tailed)

2nd Inquiry Task Analysis – Sentence Stem Use Study

When student work from the experimental groups was analyzed, it became clear that only a small number of students actually used the sentence stems in their writing. Out of the 51 students in the experimental group, only 12 of them (24%) actually used the stems in the Framing section and an even smaller number (n = 8, 16%) used the sentence stems to help write the Analysis and Conclusion section. Even though these numbers were small, the students who used these sentence stems were spread across all of the experimental conditions.

Table 51: Comparison of student sentence prompt use across experimental conditions on the 2nd inquiry task

Period	Number of Students using Sentence Stems in the Framing Section	Number of Students using Sentence Stems in the Analyzing & Concluding Section
4 (n = 16)	4	2
5 (n = 15)	3	1
6 (n = 20)	5	5

Table 52: Comparison of student sentence prompt use across experimental conditions on the 2nd inquiry task

	Framing Section Number of Students using Sentence Stems	Analyzing & Concluding Section Number of Students using Sentence Stems
Females (n = 15)	7	4
Males (n = 17)	5	4

Approximately equal numbers of males (n = 5) and females (n = 7) used the sentence stems in their Framing sections. The same was true of males (n = 4) and females (n = 4) who used sentence stems in their Analyzing & Concluding sections. The sample of students who used the sentence stems was too small to assess any differences with a Chi-Square test. Generally, though, it was the high achieving students that tended to use the sentence stems.

Table 53: Comparison of sentence stem use by achievement level on the 2nd inquiry task

Achievement Level	Number of Students using Sentence Stems in the Framing Section	Number of Students using Sentence Stems in the Analyzing & Concluding Section
High	11	6
Low	1	2

A follow up analysis was run in order to compare the results from students who used the sentence stems with those who did not use any sentence stems in their projects. An independent-samples t-test was run on the Framing scores from the 2nd inquiry task. The mean score from those who used the sentence stems (M = 3.6, SD = 0.4) was only slightly higher than the mean scores for students who did not use these sentence stems ($M = 3.2$, $SD = 0.8$). This difference (mean difference = -0.43, 95% CI: -0.91 to 0.04) approached significance ($t(53) = -1.83$, $p = 0.07$ (two-tailed)), but did not reach the $p < 0.05$ threshold. The results were similar when looking at the relationship between sentence prompt use and scores for the Analyzing & Concluding section. Again, the mean score was slightly higher for students who used the sentence stems ($M = 3.4$, $SD = 0.6$) than those that did not use these sentence stems ($M = 2.9$, $SD = 1.0$), but this difference was not significant ($t(53) = -1.44$, $p = 0.16$ (two-tailed)).

Knowing that there was a relationship between the number of words written and the section score, another independent-samples t-test was run to determine if students who used the sentence stems wrote more words in their sections than students that did not use these sentence stems. In the Framing section, the mean number of words that written by students who used the sentence stems ($M = 267$, $SD = 109$) was higher than the mean number of words written by students that did not use the sentence stems ($M = 201$, $SD = 110$). This difference in word count (mean difference = -66, 95% CI: -138 to 6.) approached significance ($t(52) = -1.84$, $p = 0.07$(two-tailed)), but did not quite reach the $p < 0.05$ threshold. In the Analyzing & Concluding section, the mean word counts were practically the same regardless of whether students used sentence stems ($M = 239$, $SD = 78$) or not ($M = 232$, $SD = 123$); $t(52) = -0.15$, $p = 0.88$ (two-tailed).

Table 54: Comparison of mean section scores by sentence stem usage on the 2nd inquiry task

	Framing		Analyzing & Concluding	
	Mean Score	SD	Mean Score	SD
Did use Sentence Stems	3.6	0.4	3.4	0.6
Did not use Sentence Stems	3.2	0.8	2.9	1.0

Table 55: Comparison of mean section word counts sentence stem usage on the 2nd inquiry task

	Framing		Analyzing & Concluding	
	Mean # of Words	SD	Mean # of Words	SD
Did use Sentence Stems	267	109	239	78
Did not use Sentence Stems	201	110	232	123

2nd Inquiry Task Analysis – Effects of Goal Type

Two of the classes in the experimental condition were asked to reflect on their work at various points during the experimentation and writing phases of this second inquiry task. Student reflections were collected when students turned in their written projects. Students were asked to write down their reflections in order to investigate student thinking that took place as these projects were completed. Related specifically to the writing process, one of the student reflection questions asked students to choose a goal related specifically to how they were going to write their inquiry task. Time was spent in each of the classes in the experimental condition talking about the difference between proximal and distal goals. It was emphasized that proximal goals often lead to better products because they focus on an immediate action whereas distal goals tend to focus on an outcome without specifying the actions needed to reach that goal. Even though all three of the classes in the experimental condition were given instruction related to goal setting, only the students in the reflection groups were asked to leave a written record.

Both of the classes in the reflection group had approximately equal numbers of students who turned in their written reflection sheets. A little more than half of the students in each class had written proximal goals while slightly less than half had written distal goals. Most of the students' goals pertained to both the Framing and Analyzing & Concluding sections while a few students had goals that focused only on one section.

Table 56: Comparison of the number of student goal orientations for the Framing section of the 2nd inquiry task

Period	Proximal Goal	Distal Goal	No Goal
5 (n = 13)	7	5	1
6 (n = 14)	9	4	1

Table 57: Comparison of the number of student goal orientations for the Analyzing & Concluding section of the 2nd inquiry task

Period	Proximal Goal	Distal Goal	No Goal
5 (n = 13)	8	5	0
6 (n = 14)	9	4	1

Student scores were then assessed to see if their goal type had any effect on scores for the second science inquiry task. An independent-samples t-test was run to see if goal type had and effect on student scores. The mean scores on the Framing section for students who chose proximal goals (M = 3.33, SD = 1.06) was slightly higher than those for the students who chose distal goals (M = 3.19, SD = 0.50), but this difference was not significant; t(23) = 0.36, p = 0.72 (two-tailed). The same test was run to compare the mean scores on the Analyzing & Concluding section. Again, mean scores for students who chose proximal goals (M = 3.07, SD = 1.25) were slightly higher than for those students who chose distal goals (M = 2.81, SD = 0.61). This difference, though, was not significant; t(24) = 0.585, p = 0.56 (two-tailed). An independent-samples t-test on word counts revealed the same thing. There were no significant differences in words counts for the Framing section (t(22) = 0.85, p = 0.40 (two-tailed)) or for the Analyzing & Concluding section (t(23) = 0.04, p = 0.97 (two-tailed)).

Table 58: Comparison of mean section scores by goal orientation on the 2nd inquiry task

Goal Type	Framing				Analyzing & Concluding			
	Mean Score	SD	Mean # of Words	SD	Mean Score	SD	Mean # of Words	SD
Proximal	3.3	1.1	238	105	3.1	1.2	259	116
Distal	3.2	0.5	201	94	2.8	0.6	256	142

As instruction was given in goal setting, students were urged to adopt the proximal goal of using a minimum of three sentence stems in both their Framing and Analyzing & Concluding sections. Student goals around sentence prompt use were then compared with actual student work to see if the goals matched the work produced. In the directed reflection experimental group (per 5), the number of students who had goals to use sentence stems was about the same as those that did not form a goal around the use of sentence stems. This was true for both the Framing and the Analyzing & Concluding section. Generally, though, the majority of the students in the direct reflection experimental group did not actually use the sentence stems in their writing. This is most pronounced in the results from the Analyzing & Concluding section where none of the students used any of the sentence stems in their writing even though approximately half of the students (6 out of 13) that they intended to use these sentence stems. The results from students in the general reflection experimental group (per 6) generally showed that if students had a goal to use the sentence stems then these stems showed up their reports. If these students did not have a goal around sentence prompt usage, then these stems generally did not show up in their repots. Again, this was true for both the Framing and the Analyzing & Concluding sections.

Table 59: Comparison of student goal to use sentence stems compared to actual use of sentence stems in the Framing section of the 2nd inquiry task

Period	Goal to use sentence stems and did use sentence stems	Goal to use sentence stems but did not use sentence stems	No goal to use sentence stems but did use sentence stems	No goal to use sentence stems and did not use sentence stems
5 (n = 13)	2	3	1	7
6 (n = 13)	5	3	0	5

Table 60: Comparison of student goal to use sentence stems compared to actual use of sentence stems in the Analyzing & Concluding section on the 2nd inquiry task

Period	Goal to use sentence stems and did use sentence stems	Goal to use sentence stems but did not use sentence stems	No goal to use sentence stems but did use sentence stems	No goal to use sentence stems and did not use sentence stems
5 (n = 13)	0	6	0	7
6 (n = 13)	5	3	0	5

2nd Inquiry Task Analysis – Post-Task Student Predictions

In addition to setting goals, students in both reflection experimental conditions were asked to score themselves on the four components of the science inquiry task (Framing, Design, Data Collection, and Analyzing & Concluding). The intent was to determine the extent to which students understood the science inquiry grading criteria and then apply those criteria to their own work. On both the Framing and the Analyzing & Concluding section, a majority of the students were able to correctly predict their scores. Between the two sections, though, it seemed that students had a better sense of the grading criteria for the Analyzing & Concluding section.

Table 61: Comparison of student predicted score to their actual score on the Framing and Analyzing & Concluding sections of the 2nd inquiry task

	Correctly Predicted Section Score		Incorrectly Predicted Section Score	
	n	%	n	%
Framing	14	63.6	8	36.4
Analyzing & Concluding	16	72.7	6	27.3

Since a majority of students understood the grading criteria, a follow up analysis looked at how motivated students were to achieve passing marks. Again, each section of the science inquiry task was graded on a scale of 0 to 6 where a score of 4 was considered to meet benchmark standards. Students were sorted into one of two groups based on a comparison of their predicted scores to the scores they actually received:

Group 1: Students predicted a passing score

Group 2: Students predicted a non-passing score

In the Framing section, a little over a third of the students (36.3%) predicted that they would not receive a passing score on that section and yet did not attempt to make any changes to their report in order to bring that section up to a passing mark. On the Analyzing & Concluding section, the percentage of students who predicted that they would not get a passing score and yet still turned in their report increased to 59.1%.

Table 62: Comparison of the percentage of students who predicted passing and non-passing scores on the Framing and Analyzing & Concluding sections of the 2nd inquiry task

	Predicted a Passing Score		Predicted a Non-Passing Score	
	n	%	n	%
Framing	14	63.6	8	36.6
Analyzing & Concluding	9	40.9	13	59.1

2nd Inquiry Task Analysis – Effects of Gender

This study also sought to understand whether differences in gender had an effect on student scores on this second inquiry task. A two-way between-groups ANOVA was conducted to explore if there was a difference in gender and experimental condition (the four class periods represented different experimental conditions) as measured by the Total Inquiry score from this 2nd inquiry task. The interaction between gender and class period was not significant, $F(3,49) = 1.93$, $p = 0.14$, partial eta squared = 0.01. Neither of the main effects for gender ($F(1,49) = 2.17$, $p = 0.15$, partial eta squared = 0.04) or class period ($F(3,429) = 0.20$, $p = 0.89$, partial eta squared = 0.01) reached statistical significance. This implied that experimental conditions did not produce significantly different scores for male or female students.

As a follow-up, a number of independent-sample t-tests were conducted. There was a significant difference in mean Total Inquiry scores for males ($M = 12.5$, $SD = 2.8$) and females ($M = 14.0$, $SD = 2.0$); $t(55) = 2.315$, $p = 0.024$ (two-tailed)) on this second inquiry task. The magnitude of this difference was moderate (eta squared = 0.09) with female scores being generally higher than their male counterparts (mean difference = 1.48, 95% CI: 0.20 to 2.76). A follow-up analysis revealed that females tended to score significantly higher than their male counterparts on both the Framing ($t(55) = 2.215$, $p = 0.031$ (two-tailed)) and the Analyzing & Concluding ($t(55) = 2.487$, $p = 0.016$ (two-tailed)) sections. Mean number of words in each section were also analyzed by gender using an independent-samples t-test. On the Framing section, there was no difference between the mean number of words that males used (M = 200, SD = 135) and the number of words that females used (M = 225, SD = 96), t(52) = 0.80, p = 0.43 (two-tailed). In contrast, the mean number of words that male (M = 176, SD = 84) and female (M = 270, SD = 121) used in the Analyzing & Concluding section were significantly different, t(52) = 3.10, p = 0.003 (two-tailed). The magnitude of the differences in the means (mean difference = 93, 95% CI: 33 to 154) was quite large (eta squared = 0.16).

Table 63: Comparison of scores on Framing and Analyzing & Concluding sections by gender on the 2nd inquiry task

	2nd Inquiry Task		Framing				Analyzing & Concluding			
	Mean Score	SD	Mean Score	SD	Mean # of Words	SD	Mean Score	SD	Mean # of Words	SD
Male	12.5	2.8	3.0	0.9	200	135	2.5	1.1	176	84
Female	14.0	2.0	3.5	0.6	225	96	3.2	0.9	270	121

2nd Inquiry Task Analysis – Effects of Achievement

A two-way between subjects ANOVA was also conducted to determine how achievement and experimental condition impacted student Total Inquiry scores on the second inquiry task. The interaction between achievement and class period was not statistically significant, F(3,49) = 0.56, p = 0.63, partial eta squared = 0.03. In a similar way, the main effect for class period was also not significant, F(3,49) = 0.61, p = 0.61, partial eta squared = 0.04. The main effect for achievement was significant, F(1,49) = 5.37, p = 0.025. The effect size was moderate (partial eta squared = 0.10) which indicated that student scores did vary based on prior achievement, but student scores did not vary based on experimental condition.

An independent-samples t-test on the mean scores between high and low achieving students was also conducted. As expected, there was a significant difference between the means scores for high achieving students (M = 14.0, SD = 1.9) and low achieving students (M = 12.4, SD = 2.9); t(55) = 2.32, p = 0.03 (two-tailed). The magnitude of this difference in the means (mean difference = 1.62, 95% CI: 0.20 to 3.04) was moderate (eta squared = 0.09). High achieving students scores were significantly higher on both the Framing (high achievers M = 3.5, SD = 0.5; low achievers M = 2.9, SD = 0.9), t(55) = 2.84, p = 0.006 (two-tailed) and the Analyzing & Concluding sections (high achievers M = 3.2, SD = 0.8; low achievers M = 2.5, SD = 1.2), ,t(55) = 2.65, p = 0.01 (two-tailed). Word counts for high achievers were significantly higher than the low achievers for the Framing (high achievers M = 248, SD = 114; low achievers M = 161, SD = 87), t(52) = 2.95, p = 0.005 (two-tailed) and the Analyzing & Concluding sections (high achievers M = 259, SD = 106; low achievers M = 189, SD = 123), t(52) = 2.19, p = 0.03 (two-tailed).

Table 64: Comparison of scores on Framing and Analyzing & Concluding sections by achievement on the 2nd inquiry task

	2nd Inquiry Task		Framing				Analyzing & Concluding			
	Mean Score	SD	Mean Score	SD	Mean # of Words	SD	Mean Score	SD	Mean # of Words	SD
High Achievers	14.0	1.9	3.5	0.5	248	114	3.2	0.8	259	106
Low Achievers	12.4	2.9	2.9	0.9	161	87	2.5	1.2	189	123

2nd Inquiry Task Analysis – Comparison to Baseline: Introduction

The purpose of these next sections is to compare student results on this 2nd Inquiry back to the Baseline task. The intent was to see what effects, if any, were produced by the various treatment conditions and to track the effect of these treatment conditions over time. Because sample sizes in this study tended to be small and also a large number of statistics were run, the choice was made to go with the more conservative multivariate analysis of variance statistic.

2nd Inquiry Task Analysis – Comparison to Baseline: Inquiry Task Scores

In order to determine if there were any effects due to the treatments introduced in the experimental conditions, group mean total inquiry scores were compared back to baseline data using a mixed between-within subjects ANOVA. The interaction between time and group was not significant, Wilks' Lambda = 0.89, $F(3, 51) = 2.07$, $p = 0.12$, partial eta squared = 0.11. The main effect for time turned out to be significant, Wilks' Lambda = 0.90, $F(1,51) = 5.95$, $p = 0.02$, partial eta squared = 0.10, which indicated that there was a general increase in total inquiry scores between the initial and second administration of the inquiry tasks. Total inquiry scores for the students in the experimental conditions generally increased while total inquiry scores for the students in the control group generally decreased. These changes, though, were not significant, $F(3,51) = 0.12$, $p = 0.95$.

Table 65: Comparison of mean inquiry scores by period from the 2nd inquiry back to baseline data

	Per 2 (n = 7)		Per 4 (n = 15)		Per 5 (n = 15)		Per 6 (n = 18)	
	Mean Score	SD	Mean Score	SD	Mean Score	SD	Mean Score	SD
Baseline	13.4	1.1	11.9	2.8	12.1	2.2	12.8	2.9
2nd Inquiry	13.3	1.4	13.5	3.1	13.7	1.8	13.0	2.9

Period 2 – Control Group, Period 4 - Sentence Stems Only, Period 5 - Sentence Stems + Directed Reflection, Period 6 - Sentence Stems + General Reflection

Because the main effect for time was significant, two paired-samples t-tests were run in order to determine if these increased scores were due to increased scores in the Framing or Analyzing & Concluding sections. Because there was not a significant effect for group membership, students were not divided into groups for this comparison. Mean scores in the Framing section increased from 3.2 (SD = 0.7) to 3.3 (SD = 0.8), but this difference was not significant, t(54) = 1.32, p = 0.19 (two-tailed). The same was true with the scores from the Analyzing & Concluding section. Mean scores increased from 2.9 (SD = 0.8) to 2.9 (SD = 1.0), but this increase was also not significant, t(54) = 0.54, p = 0.59 (two-tailed). This implies that either increases in student scores were in the Design and Collection sections or that incremental (insignificant) increases in each of the four sections added up to a significant increase overall.

Table 66: Comparison of mean inquiry scores by entire task and inquiry sub-sections between results from the 2nd inquiry task back to baseline data

	n	2nd Inquiry Task		Framing		Analysis	
		Mean Score	SD	Mean Score	SD	Mean Score	SD
Baseline	55	12.6	2.5	3.2	0.7	2.9	0.8
2nd Inquiry	55	13.4	2.5	3.3	0.8	2.9	1.0

A follow up analysis looked to see if there was an effect for students who were in any of the treatment groups. All of the students in the experimental condition were aggregated into one group for this analysis. A mixed between-within ANOVA was run in order to compare mean inquiry task scores from this second inquiry task back to baseline data. There was no significant interaction between time and group membership; Wilks' Lambda = 0.97, $F(1, 53) = 1.67$, $p = 0.20$, partial eta squared = 0.03. The main effect for time was also not significant, Wilks' Lambda = 0.98, $F(1, 53) = 0.92$, $p = 0.34$, partial eta squared = 0.02. The mean total inquiry scores generally decreased between the initial and second inquiry task for students in the control group while total inquiry scores generally increased for students in the experimental groups. The main effect comparing these two conditions was not significant, $F(1, 53) = 0.25$, $p = 0.62$.

Table 67: Comparison of mean total inquiry score for control and experimental groups with results from the 2nd inquiry back to baseline data

	Control Group			Experimental Group		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	7	13.4	1.1	48	12.4	2.6
2nd Inquiry	7	13.3	1.4	48	13.4	2.6

2nd Inquiry Task Analysis – Comparison to Baseline: Word Count

Knowing that scores had generally increased from the initial inquiry task to the second inquiry task, two paired-samples t-tests were run in order to compare word counts in both the Framing and the Analysis & Conclusion sections from baseline to the second inquiry task. The number of words used in the Framing section increased from a mean of 144 words (SD = 51 words) to a mean of 208 words (SD = 106 words). This increase was significant, $t(43) = 4.03$, $p < 0.0005$ (two-tailed). The mean increase was 64 words (95% CI 32 words to 96 words) with an eta squared statistic = 0.27 which indicated a large effect size. When a similar analysis was run on the number of words in the Analyzing & Concluding section, there was also an increase in the mean number of words used from 212 words (SD = 123 words) to 233 words (SD = 120 words) used. The mean increase of 21 words (95% CI -10 words to 52 words), though, was not significant, $t(43) = 1.38$, $p = 0.18$ (two-tailed).

Table 68: Comparison of Framing and Analyzing & Concluding section word counts comparing results from the 2nd inquiry to back to baseline data

	n	Framing		Analyzing & Concluding	
		Mean# of Words	SD	Mean # of Words	SD
Baseline	44	144	51	212	123
2nd Inquiry	44	208	106	233	120

A mixed between-within ANOVA was then used to see if group membership had any effect on the number of words that a student used. The initial analysis used data from the Framing section. The interaction between time and group membership was not significant, Wilks' Lambda = 0.92, F (3, 40) = 1.23, p = 0.31, partial eta squared = 0.09. There was a significant main effect for time, Wilks' Lambda = 0.82, F(1,40) = 9.10, p = 0.004, partial eta squared = 0.19, which agrees with the results of the prior paired-samples t-test. Although the number of words in the Framing section generally increased for students in the experimental groups, the number of words for the students in the control group tended to decrease (see table 69). These differences, though, were not significant, F(3,40) = 0.243, p = 0.87, partial eta squared = 0.02.

A similar test was conducted in order to compare the number of words in the Analyzing & Concluding section from the second inquiry back to baseline data. Again, the interaction between time and group membership was not significant, Wilks' Lambda = 0.91, F(3,40) = 1.37, p = 0.27, partial eta squared = 0.09. As was expected, the main effect for time was also not significant, Wilks' Lambda = 0.96, F(1,40) = 1.87, p = 0.18, partial eta squared = 0.05. All of the groups saw the mean number of words in the Analyzing & Concluding section increase except for the experimental group that also was also asked to response to general reflection prompts (see table 70). This difference in groups was not significant, F(4,30) = 0.04, p = 0.99.

Table 69: Comparison of Framing section mean word count by period from the 2nd inquiry task back to baseline data

	Per 2		Per 4		Per 5		Per 6	
	Mean # of Words	SD	Mean # of Words	SD	Mean # of Words	SD	Mean # of Words	SD
Baseline	185	39	153	48	145	61	120	34
2nd Inquiry	172	42	223	133	205	79	211	126

Period 2 – Control Group, Period 4 - Sentence Stems Only, Period 5 - Sentence Stems + Directed Reflection, Period 6 - Sentence Stems + General Reflection

Table 70: Comparison of Analyzing & Concluding section mean word count by period from the 2nd inquiry task back to baseline data

	Per 2		Per 4		Per 5		Per 6	
	Mean # of Words	SD	Mean # of Words	SD	Mean # of Words	SD	Mean # of Words	SD
Baseline	209	89	191	90	206	101	235	174
2nd Inquiry	224	42	260	125	225	129	222	132

Period 2 – Control Group, Period 4 - Sentence Stems Only, Period 5 - Sentence Stems + Directed Reflection, Period 6 - Sentence Stems + General Reflection

2nd Inquiry Task Analysis – Comparison to Baseline: Sentence Stem Use

In order to assess if student use of sentence stems had any effect on scores, a number of mixed between-within subjects ANOVA's were run on the data. The first analysis compared the Framing section scores for students who used the sentence stems to students who did not use the stems. The scores from the second inquiry task were compared back to scores from the baseline task. The interaction between time and stem use was not significant, Wilks' Lambda = 0.96, $F(1,52) = 2.46$, $p = 0.12$, partial eta squared = 0.05. The main effect for time was also not significant, Wilks' Lambda = 0.99, $F(1,52) = 0.41$, $p = 0.53$, partial eta squared = 0.01. Interestingly, the mean scores for students who used the stems in the Framing section generally fell slightly overtime whereas the mean scores for students who did not use the stems generally increased slightly. The mean scores for students who chose to use the stems on the second inquiry task generally had higher Framing section scores on the initial inquiry task and these scores remained higher than those who did not use stems. This difference in scores was significant, F(1,52) = 8.26, p = 0.006, partial eta squared = 0.14.

Table 71: Comparison of mean Framing section score by sentence stem use from the 2^{nd} inquiry task to back to baseline data

	Sentence Stem Use			No Sentence Stem Use		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	12	3.7	0.3	42	3.0	0.7
2^{nd} Inquiry	12	3.6	0.4	42	3.2	0.8

Table 72: Comparison of mean Analyzing & Concluding section by sentence stem use from the 2^{nd} inquiry task to back to baseline data

	Sentence Stem Use			No Sentence Stem Use		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	8	3.0	0.4	46	2.8	0.8
2^{nd} Inquiry	8	3.4	0.6	46	2.9	1.1

A similar analysis was done comparing sentence stem use to Analyzing & Concluding section scores over time. Again, there was no significant interaction between time and stem use, Wilks' Lambda = 0.98, F(1,52) = 1.22, p = 0.27, partial eta squared = 0.02. In addition, the main effect for time was also not significant, Wilks' Lambda = 0.96, F (1, 52) = 1.94, p = 0.17, partial eta squared = 0.04. Both groups' scores on the Analyzing & Concluding section increased over time, but the increase was greater for the students who had used the sentence stems in their writing. The main effect for sentence stem use was not significant, F (1, 52) = 1.58, p = 0.21, partial eta squared = 0.03.

In addition to section scores, the number of words used was also compared between students who used sentence stems to those who did not. A mixed between-within subjects ANOVA was conducted using the word counts in Framing sections from the initial inquiry task and the second inquiry task. There was a significant interaction between time and sentence stem use, Wilks' Lambda = 0.91, $F(1,42) = 4.38$, $p = 0.04$. This implied that the change of scores over time in one group was significantly different than the change in scores of the other group. In looking at the group means, the students who used the sentence stems used more words in the Framing section than the students who did not use the sentence stems (see table 73). For this interaction, the partial eta squared = 0.10 indicating a moderate to large effect size. Since the interaction effect was significant, caution is needed in interpreting the main effects. The main effect for time was significant, Wilks' Lambda = 0.66, $F(1, 42) = 21.83$, $p < 0.0005$, with the partial eta squared = 0.34 indicating a very large effect size. Both of the groups, whether they used the sentence stems or not, did generally increase the number of words that they wrote in the Framing section. Students who used the sentence stems tended to have the larger gain in word count, but this difference was not significant, $F(1,42) = 2.69$, $p = 0.11$, partial eta squared = 0.06.

Table 73: Comparison of mean Framing section number of words by sentence stem use from the 2nd inquiry task to back to baseline data

	Sentence Stem Use			No Sentence Stem Use		
	n	Mean # of Words	SD	n	Mean # of Words	SD
Baseline	11	143	40	33	144	54
2nd Inquiry	11	262	113	33	189	98

The same analysis was done, but this time with the word count from the Analyzing & Concluding section. This time, the interaction between time and sentence stem use was not significant, Wilks' Lambda = 1.00, $F(1,42) = 0.004$, $p = 0.95$. In addition, the main effect for time was also not significant, Wilks' Lambda = 0.98, $F(1, 42) = 0.91$, $p = 0.35$, partial eta squared = 0.02. Both groups showed an increase in word usage, but in contrast to the Framing section, the students who were in the group that did not use the stems tended to use more words in the baseline inquiry task and this trend was maintained through the second inquiry task (see table 74). This difference did not turn out to be significant, $F(1,42) = 0.06$, $p = 0.80$.

Table 74: Comparison of mean Analyzing & Concluding section number of words by sentence stem use from the 2nd inquiry task to back to baseline data

	Sentence Stem Use			No Sentence Stem Use		
	n	Mean # of Words	SD	n	Mean # of Words	SD
Baseline	7	203	72	37	213	132
2nd Inquiry	7	222	66	37	235	128

2nd Inquiry Task Analysis – Comparison to Baseline: Effects of Goal Type

This comparison focuses specifically on the two groups of students who were asked to reflect at various points in time. Proximal goals are those goals that tend to focus on actions and can be assessed in the immediate future. In contrast, distal goals are those goals that tend to focus on future outcomes and can only be assessed after some amount of time has passed.

Students in the two reflection groups were compared in order to assess if there was any difference in outcomes between those who had adopted proximal goals and those who had adopted distal goals. A mixed between-within subjects ANOVA was conducted using the Framing section scores collected from the initial inquiry task and the second inquiry task. There was no significant interaction between time and goal type, Wilks' Lambda = 0.97, $F(1, 23) = 0.81$, $p = 0.38$, partial eta squared = 0.03. The main effect for time approached significance, Wilks' Lambda = 0.88, $F(1, 23) = 3.27$, $p = 0.08$, partial eta squared = 0.12, but did not quite reach the $p < 0.05$ threshold. While both goal groups showed an increase in their Framing section score, those who had adopted proximal goals had the greater score increase (see table 75). The main effect comparing the two goal orientations was not significant, $F(1, 23) = 0.03$, $p = 0.87$.

The same analysis was run using data from the Analyzing & Concluding section. Again, the interaction between goal type and time was not significant, Wilks' Lambda = 0.99, $F(1, 24) = 0.32$, $p = 0.56$, partial eta squared = 0.01. The main effect for time not significant either, Wilks' Lambda = 0.99, $F(1, 24) = 0.20$, $P = 0.66$, partial eta squared = 0.01. Student who had adopted proximal goals tended to see their scores on the Analyzing & Concluding section increase while students who had adopted distal goals tended to see their scores remain the same (see table 76). This difference was not significant, F (1, 24) = 0.24, p = 0.63, partial eta squared = 0.01.

Table 75: Comparison of mean Framing section score by goal orientation from the 2nd inquiry task back to baseline data

	Proximal Goal			Distal Goal		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	16	3.1	1.0	9	3.1	0.3
2nd Inquiry	16	3.3	1.1	9	3.2	0.5

Table 76: Comparison of mean Analyzing & Concluding section score by goal orientation from the 2nd inquiry task back to baseline data

	Proximal Goal			Distal Goal		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	17	2.9	0.6	9	2.8	0.7
2nd Inquiry	17	3.1	1.2	9	2.8	0.6

In addition word counts were also compared by goal orientation. A mixed between-within subjects ANOVA was conducted in order to determine if goal orientation had any effect on the number of words that a student wrote. The interaction of time and goal type was not significant, Wilks' Lambda = 0.98, F (1, 19) = 0.36, p = 0.56, partial eta squared = 0.02. The main effect for time did turn out to be significant, Wilks' Lambda = 0.64, F (1, 19) = 10.57, p = 0.004, which implied that the number of words increased for both goal orientations. The partial eta squared statistic = 0.36 which indicated a very large effect size. While both goal orientations did increase their number of Framing section words over time, students who had adopted proximal goals tended to see the greater gain in word count. This main effect comparing these two goal types, though, was not significant, F(1,19) = 0.61, p = 0.45, partial eta squared = 0.03, which indicated that there was no real difference between the two.

Table 77: Comparison of mean Framing section number of words by goal orientation from the 2nd inquiry task back to baseline data

	Proximal Goal			Distal Goal		
	n	Mean # of Words	SD	n	Mean # of Words	SD
Baseline	12	147	53	9	140	41
2nd Inquiry	12	236	110	9	201	94

The same analysis was run with the word count data from the Analyzing & Concluding section. There was no significant interaction between time and goal orientation, Wilks' Lambda = 0.86, F (1, 20) = 3.28, p = 0.09, partial eta squared = 0.14. The main effect for time was also not significant, Wilks' Lambda = 0.92, F (1, 20) = 1.83, p = 0.19, partial eta squared = 0.08. Interestingly, students who had adopted a proximal goal orientation tended to use fewer words over time on the Analyzing & Concluding section whereas students who had adopted a distal goal tended to use more words over time. The main effect for goal type, though, was not significant, F(1,20) = 0.09, p = 0.77.

Table 78: Comparison of mean Analyzing & Concluding section number of words by goal orientation from the 2nd inquiry task back to baseline data

	Proximal Goal			Distal Goal		
	n	Mean # of Words	SD	n	Mean # of Words	SD
Baseline	13	252	158	9	205	137
2nd Inquiry	13	245	124	9	256	142

2nd Inquiry Task Analysis – Comparison to Baseline: Effects of Gender

To understand the relationship between gender and science inquiry scores, a mixed between-within subjects ANOVA was conducted using scores from the initial and second inquiry tasks. The interaction between gender and time was not significant, Wilks' Lambda = 0.96, $F(1,53) = 2.39$, $p = 0.13$, partial eta squared = 0.04, although there was a substantial main effect for time, Wilks' Lambda = 0.89, $F(1,53) = 6.81$, $p = 0.01$, partial eta squared = 0.11. This implied that the increase in scores over time was significant for both groups of students. Although both male and female students saw their scores increase over time, female students saw the greatest increase in their scores. This difference approached significance, $F(1,53) = 2.76$, $p = 0.10$, partial eta squared = 0.05, but did not meet the significance threshold of $p < 0.05$.

Table 79: Comparison of mean student inquiry scores by gender score from the 2nd inquiry task back to baseline data

	Male			Female		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	23	12.2	3.0	32	12.8	2.1
2nd Inquiry	23	12.5	2.8	32	14.0	2.1

A similar mixed between-within subjects ANOVA was run using the score data from the Framing section of the initial and second inquiry tasks. The interaction between time and gender was not significant, Wilks' Lambda = 0.96, $F(,53) = 1.37$, $p = 0.25$, partial eta squared = 0.03. The main effect for time was not significant, Wilks' Lambda = 0.98, $F(1,53) = 1.24$, $p = 0.27$, partial eta squared = 0.02. Male scores on the Framing section did not really change much over time while female students tended to see their scores increase. The main effect for gender approached significance, F(1,53) = 2.94, p = 0.09, partial eta squared = 0.05.

Table 80: Comparison of mean student Framing section scores by gender from the 2nd inquiry task back to baseline data

	Male			Female		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	23	3.0	0.9	32	3.2	0.5
2nd Inquiry	23	3.0	0.9	32	3.4	0.5

The same analysis was run using the student scores from the Analyzing & Concluding sections. Again the interaction of time and gender was not significant (Wilks' Lambda = 0.99, F(1,53) = 0.78, p = 0.38, partial eta squared = 0.02), and the same was true for the main effect of time (Wilks' Lambda = 1.00, F(1,53) = 0.15, p = 0.70, partial eta squared = 0.003). While male scores on the Analyzing & Concluding section tended to decrease over time, female scores tended to actually increase over time (see table 81). The main effect for gender, in this case, was significant, F(1,53) = 7.29, p = 0.009, indicating that female students did indeed do better on this part of the inquiry task than their male counterparts. The partial eta squared statistic = 0.12 indicated a moderate to large effect size.

Table 81: Comparison of mean student Analyzing & Concluding section scores by gender from the 2nd inquiry task back to baseline data

	Male			Female		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	23	2.6	0.9	32	3.0	0.6
2nd Inquiry	23	2.5	1.1	32	3.2	0.9

A mixed between-within subjects ANOVA was run using word counts from the Framing sections of the baseline and 2nd inquiry tasks. The interaction of time and gender was not significant, Wilks' Lambda = 0.97, F(1,42) = 1.38, p = 0.25, partial eta squared = 0.03. The main effect for time was significant and substantial, Wilks' Lambda = 0.78, F(1,42) = 12.00, p = 0.001, partial eta squared = 0.22, which implied that both males and females used more words in the Framing section of their 2nd inquiry task. Even though female students tended to have the greater gain in terms of word use, the main effect for gender was not significant, F(1,42) = 0.87, p = 0.36, partial eta squared = 0.02.

A similar analysis was done using the word counts from the Analyzing & Concluding sections. The interaction term of time and gender was not significant, Wilks' Lambda = 0.97, F(1,42) = 1.53, p = 0.22, partial eta squared = 0.04. The same was true for the main effect of time, Wilks' Lambda = 0.98, F(1,42) = 0.85, p = 0.36, partial eta squared = 0.02. Similar to the scores in the Analyzing & Concluding section, the number of words used by females tended to increase while males tended to use fewer words over time. The main effect for gender was significant, F(1,42) =

6.11, p = 0.018. The partial eta squared statistic (0.13) indicated that this was a moderate to large effect size.

Table 82: Comparison of mean Framing section number of words by gender from the 2nd inquiry task back to baseline data

	Male			Female		
	n	Mean # of Words	SD	n	Mean # of Words	SD
Baseline	15	144	45	29	144	54
2nd Inquiry	15	187	125	29	221	94

Table 83: Comparison of mean Analyzing & Concluding number of words by gender from the 2nd inquiry task back to baseline data

	Male			Female		
	n	Mean # of Words	SD	n	Mean # of Words	SD
Baseline	15	170	86	29	233	135
2nd Inquiry	15	165	78	29	267	124

2nd Inquiry Task Analysis – Comparison to Baseline: Effects of Achievement

To assess how Total Inquiry scores for high and low achieving students changed between the initial and second inquiry task, a mixed between-within subjects ANOVA was conducted. The interaction between time and achievement did not reach statistical significance, Wilks' Lambda = 0.96, F(1,53) = 2.44, p = 0.13, partial eta squared = 0.04. The main effect for time, though, did reach statistical significance, Wilks' Lambda = 0.84, F(1,53) = 10.36, p = 0.02, which indicated that scores increased over time for both high and low achievers. The partial eta squared statistic (0.16) indicated that this was a large effect size. In addition, the main effect for achievement was also significant, $F(1,53) = 17.01$, $p < 0.0005$, which indicated that high achieving students significantly outscored their lower achieving counterparts. The partial eta squared = 0.24 statistic implied a very large effect size for the impact of prior achievement on total inquiry scores.

Table 84: Comparison of mean student inquiry scores by prior student achievement from the 2nd inquiry task back to baseline data

	High Achievers			Low Achievers		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	35	13.5	1.7	20	10.8	2.7
2nd Inquiry	35	14.1	1.9	20	12.2	2.9

A follow up analysis used a mixed between-within subjects ANOVA to look at differences in scores in the Framing section for high and low achieving students. There was no significant interaction between time and achievement, Wilks' Lambda = 0.99, $F(1,53) = 0.56$, $p = 0.46$, partial eta squared = 0.01. The main effect for time was also not significant, Wilks' Lambda = 0.96, $F(1,53) = 2.16$, $p = 0.15$, partial eta squared = 0.04, which implied that scores did not change significantly in the Framing section over time. The main effect for achievement was statistically significant, $F(1,53) = 19.76$, $p < 0.0005$, which indicated that achievement did impact student scores. This effect size was very large (partial eta squared = 0.27).

Table 85: Comparison of mean student Framing section scores by prior student achievement from the 2nd inquiry task back to baseline data

	High Achievers			Low Achievers		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	35	3.4	0.4	20	2.7	0.8
2nd Inquiry	35	3.5	0.5	20	2.8	0.9

Table 86: Comparison of mean student Analyzing & Concluding section scores by prior student achievement from the 2nd inquiry task back to baseline data

	High Achievers			Low Achievers		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	35	3.1	0.5	20	2.5	0.9
2nd Inquiry	35	3.2	0.8	20	2.4	1.3

The same analysis was run with scores for the Analyzing & Concluding section. The interaction effect between time and achievement was not statistically significant (Wilks' Lambda = 0.99, $F(1,53) = 0.45$, $p = 0.51$, partial eta squared = 0.01), and the same was true of the main effect for time (Wilks' Lambda = 1.00, $F(1,53) = 0.11$, $p = 0.74$, partial eta squared = 0.002). The main effect for achievement was significant, $F(1,53) = 12.20$, $p = 0.001$, partial eta squared = 0.19, which indicated that high achieving students continued to score higher than their lower achieving counterparts on the Analyzing & Concluding section (see table 86).

A mixed between-within subjects ANOVA was also conducted to look at the differences in Framing section word counts for high and low achieving students on the initial and second inquiry tasks. There was significant interaction between time and achievement, Wilks' Lambda = 0.91, $F(1,42) = 4.16$, $p = 0.048$, partial eta squared =

0.09. This implied that even though both high and low achieving students used more words on their second projects, one group, specifically the high achievers, had a significantly greater change in word usage than their lower achieving counterparts. In addition, the main effect for time was also significant, Wilks' Lambda = 0.81, F(1,42) = 9.91, p = 0.003, partial eta squared = 0.19 which indicated that all student word counts increased over time on the Framing section. As suspected, the main effect for achievement was significant, F(1,42) = 7.56, p = 0.009, partial eta squared = 0.15.

Table 87: Comparison of mean Framing section number of words by prior student achievement from the 2nd inquiry task back to baseline data

	High Achievers			Low Achievers		
	n	Mean # of Words	SD	n	Mean # of Words	SD
Baseline	30	150	54	14	130	40
2nd Inquiry	30	235	108	14	149	73

When the same analysis was run on the word counts from the Analyzing & Concluding section, the results indicated that not as much changed over time when compared to the Framing section word counts. The interaction between time and achievement was not significant, Wilks' Lambda = 1.00, F(1,42) = 0.02, p = 0.90, and the same was also true for the main effect of time, Wilks' Lambda = 0.96, F(1,42) = 1.72, p = 0.20, partial eta squared = 0.04. The main effect for achievement was significant, F(1,42) = 4.81, p = 0.034, which indicated that high achieving students tended to use more words than their lower achieving counterparts (see table 88). The partial eta squared statistic (0.10) indicated that this was a moderate effect size.

Table 88: Comparison of mean Analyzing & Concluding section number of words by prior student achievement from the 2nd inquiry task back to baseline data

	High Achievers			Low Achievers		
	n	Mean # of Words	SD	n	Mean # of Words	SD
Baseline	30	236	131	14	159	89
2nd Inquiry	30	256	111	14	183	128

2nd Inquiry Task Analysis – Summary

This stage of the project began with an assessment of student self-efficacy. Not surprisingly, student efficacy related to Chemistry Content increased in comparison to the initial giving of the survey. Student efficacy related to Science Inquiry, though, remained about the same and the initial administration. Even though higher achievers tended to have greater efficacy in relation to Chemistry Content, all students were very similar in their self-efficacy related to Science Inquiry. Also, there was not a significant difference between males and females on any of these self-efficacy measures.

In addition to measuring student self-efficacy, this stage of the study also sought to understand the relationship between different experimental conditions and student scores on their inquiry tasks. When scores were analyzed from this second inquiry task, it became clear that there were no distinct differences between any of the classes in this study. A further analysis indicated that only a few students had actually used the sentence stems from the experimental conditions.

Third Inquiry Task – Intervention Continuation

Chemistry Self-Efficacy Survey (CSES)

3rd Administration CSES – Validation

This phase took place approximately 26 weeks into the school year. Students had now had the chance to complete and receive feedback on two science inquiry tasks. In addition, students were continuing to receive instruction with the interventions associated with the treatment conditions. Again, the purpose of the Chemistry Self-Efficacy Survey (CSES) was to monitor student perceptions of their understandings of Chemistry Content and their skills related to Science Inquiry. This was the third time that the students took this survey.

Like previous administrations, an analysis was conducted to measure the psychometrics of this instrument. The Chemistry Content Self-Efficacy scale yielded a Cronbach's α coefficient of 0.90 with a mean inter-item correlation of 0.48 (range: 0.21 to 0.76). This analysis was repeated on the Science Inquiry Self-Efficacy scale and yielded a Cronbach's α coefficient of 0.92 with a mean inter-item correlation of 0.51 (range: 0.11 to 0.86). These Cronbach's α coefficients were both higher than the previous administrations of this survey (2nd administration Chemistry Content Self-Efficacy Cronbach's $\alpha = 0.80$; 2nd administration Science Inquiry Self-Efficacy Cronbach's $\alpha = 0.90$).

The CSES was also evaluated in order to see if the initial two-factor structure was still valid. An inspection of the correlation matrix revealed that many coefficients were 0.3 and above. The Kaiser-Meyer-Oklin value was 0.80 which exceeded the recommended value of 0.6 (Kaiser, 1970 & 1974). Bartlett's Test of Sphericity (Bartlett, 1954) reached statistical significance ($p < 0.0005$) which provided support for the existence of factors in the correlation matrix. An analysis of the screeplot indicated that a two factor solution would explain a majority of the variance. To aid in the interpretation of these two components, an oblimin rotation was performed.

A two-component solution explained a total of 57.4% of the variance. The items identified with the Science Inquiry Self-Efficacy scale explained 49.1% of the variance, and the items identified with the Chemistry Content Self-Efficacy scale explained and addition 8.3 % of the variance. While this iteration of the CSES was able to explain a greater percentage of the variance (% variance explained by 1st administration: 41.8%, % variance explained by 2nd administration: 49.5%), in this third iteration, the two scales were also more highly correlated ($r = 0.61$) than previous iterations (1st administration: $r = 0.29$, 2nd administration: $r = 0.36$).

Table 89: Pattern and structure matrix for PCA with oblmin rotation of a two factor solution of CSES items from the 3rd administration

Item	Pattern Coefficients		Structure Coefficients		Communalities
	Component 1	Component 2	Component 1	Component 2	
Question 1	.004	**.882**	.541	**.885**	.782
Question 2	-.158	**.918**	.400	**.821**	.690
Question 3	**.761**	.038	**.784**	.502	.616
Question 4	.322	**.358**	.540	**.554**	.372
Question 5	**.662**	.128	**.740**	.531	.558
Question 6	.068	**.616**	.442	**.657**	.434
Question 7	**.800**	-.070	**.757**	.417	.576
Question 8	.318	**.594**	.679	**.788**	.684

Question 9	**.490**	.000	**.490**	.298	.240
Question 10	**.502**	.289	**.678**	.595	.512
Question 11	**.537**	.275	**.704**	.602	.544
Question 12	.241	**.636**	.629	**.783**	.650
Question 13	**.549**	.258	**.706**	.593	.541
Question 14	-.065	**.847**	.451	**.807**	.654
Question 15	**.475**	.230	**.615**	.519	.411
Question 16	.016	**.577**	.367	**.587**	.344
Question 17	**.813**	.017	**.823**	.512	.678
Question 18	**.844**	-.169	**.741**	.345	.567
Question 19	.240	**.715**	.675	**.861**	.777
Question 20	**.815**	.019	**.827**	.516	.684
Question 21	**.888**	-.053	**.856**	.488	.734

Note: Major loadings for each item are in bold.
The Chemistry Content Scale contained questions: 1, 2, 6, 9, 10, 12, 14, 15, 16, 19
The Science Inquiry Scale contained questions: 3, 4, 5, 7, 8, 11, 13, 17, 18, 20, 21

Item loadings were analyzed to ensure that items from the survey loaded on the correct scale. The Science Inquiry Self-Efficacy scale was composed of eleven items. Nine of these items loaded best on this scale. The two remaining items in this scale, questions 4 & 8, loaded well on both scales. Question 4 asked students how they felt about their ability to represent data in graphical form and question 8 asked students how they felt about developing a science investigation question. Student self-efficacy for Chemistry Content (M = 4.8, SD = 1.2) was slightly lower than student self-efficacy for Science Inquiry (M = 5.0, SD = 1.2). This implied that students generally felt less efficacious in relation to these two Science Inquiry skills.

On the Chemistry Content scale, seven of the ten items had their best loading on this factor. The three items that did not load as well were questions 9, 10, and 15.

These three questions address content that was covered at various points during the year. Question 9 dealt with electrons and bonding, question 10 dealt with patterns in the periodic table, and question 15 covered stoichiometry. It was difficult to say why these questions may have loaded better on the Science Inquiry scale. Since these questions clearly cover chemistry content, the choice was made to keep these items in the scale. Again, as was stated in the previous CSES analysis, it was anticipated that as student efficacy increased as the study progressed, these scales would converge and overlap.

3rd Administration CSES – Student Score Analysis

In order to compare average scores across the classes, a one-way ANOVA was run on the class means for the Chemistry Content Self-Efficacy scale. Class means ranged from 4.5 to 5.4 on a 7 point Likert-type scale where a low score indicated low self-efficacy and a high score indicated high self-efficacy (see table 90). There was no statistically significant difference between any of the classes on Chemistry Content Self-Efficacy scale, ($F(3,53) = 1.87$, $p = 0.15$). When a similar one-way ANOVA was run on to compare the class means on the Science Inquiry Self-Efficacy scale, the results were similar. Class means ranged from 4.6 to 5.5 (see table 90) and again there was no significant difference between the groups ($F(3,53) = 1.382$, $p = 0.26$).

Table 90: Summary results from the 3rd administration of the Chemistry Self-Efficacy Survey of mean Chemistry Content Self-Efficacy and mean Science Inquiry Self-Efficacy by period

Period	Chemistry Content Self-Efficacy		Science Inquiry Self-Efficacy	
	Mean	SD	Mean	SD
2	4.9	1.3	4.6	1.5
4	4.5	0.8	5.0	1.0
5	5.4	0.9	5.5	0.9
6	4.6	1.5	4.9	1.3

The chemistry content self-efficacy data was disaggregated in order to determine the extent to which student efficacy around content had changed over the course of the year. In looking at the results, there was not a clear pattern. At the time of the survey, the students were studying the content related to questions #14 and #2 which may explain why student efficacy related to these two questions was fairly high in comparison. Also of interest, student efficacy for question #9 seemed to be fairly low. The question asked students about their ability to explain how electrons are involved in the formation of various ionic and covalent compounds. This topic had been covered much earlier in the year, and these results seemed to indicate that students did not feel confident in their abilities to explain this concept. That there was no consistent pattern in the student efficacy results was not necessarily surprising. All of the content questions with the exception of question #16 had been covered by the time that the 3rd administration of the CSES was given. Some concepts in chemistry are more difficult than others and so it is expected that student self-efficacy should vary as well.

Table 91: Summary results from the 3rd administration of the Chemistry Self-Efficacy Survey of mean Chemistry Content Self-Efficacy by order taught and by class period

Period	Chemistry Self-Efficacy Survey – Chemistry Content Self Efficacy Questions									
	#6	#12	#10	#9	#1	#15	#19	#14	#2	#16
2	4.8	5.4	5.4	3.6	5.1	4.2	4.7	5.6	5.3	4.8
4	4.1	5.3	4.5	3.7	4.7	4.4	4.7	5.3	5.1	3.9
5	5.3	6.4	6.2	3.8	6.1	4.4	5.5	6.5	5.7	4.3
6	4.0	5.7	5.3	3.5	4.9	3.7	4.7	5.5	4.5	3.8

Similar to prior administrations of the CSES, the four parts of the written inquiry were disaggregated to aid in the analysis. While students in period 5 appeared to have higher science inquiry self-efficacy scores, four ANOVA's were conducted that assessed each inquiry phase across the four classes. The results indicated that there were no significant differences between the students in various periods within each of the four inquiry phases (Framing, $F(3,53) = 1.84$, $p = 0.23$; Designing, $F(3,53) = 1.36$, $p = 0.27$; Collecting, $F(3,53) = 1.74$, $p = 0.17$; Analyzing & Concluding, $F(3,53) = 1.07$, $p = 0.37$).

Table 92: Summary results from the 3rd administration of the Chemistry Self-Efficacy Survey of mean Science Inquiry Self-Efficacy by inquiry category and period

Period	Framing Self-Efficacy		Designing Self-Efficacy		Collecting & Presenting Self-Efficacy		Analyzing & Concluding Self-Efficacy	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2	4.4	1.6	5.1	1.6	4.8	1.5	4.6	1.8
4	4.9	1.3	5.1	1.0	5.3	0.9	4.7	1.1
5	5.6	1.2	5.9	1.3	5.8	1.0	5.4	0.7
6	4.8	1.5	5.1	1.4	5.1	1.4	4.9	1.2

Student self-efficacy scores for the phases of the written science inquiry project were compared using paired-samples t-tests. These tests revealed that student self efficacy for Framing ($M = 5.0$, $SD = 1.4$) was significantly lower than that of Designing ($M = 5.3$, $SD = 1.3$), $t(56) = 2.16$, $p = 0.04$ (two-tailed). The same was also true in that student self-efficacy for Analyzing & Concluding ($M = 5.0$, $SD = 1.2$) was significantly lower than that of Designing ($M = 5.3$, $SD = 1.3$), $t(56) = 2.58$. $p = 0.01$ (two-tailed). The only other significant results was for the comparison between Analyzing & Concluding ($M = 5.0$, $SD = 1.2$) and Collecting ($M = 5.3$, $SD = 1.3$) where students indicated that they felt more efficacious about their abilities to Collect data than they did about their abilities to Communicate their results, $t(55) = 2.64$, $p = 0.01$ (two-tailed). Students felt more efficacious about Collecting ($M = 5.3$, $SD = 1.3$) than they did about Framing ($M = 5.0$, $SD = 1.4$), but this comparison only approached significance, $t(55) = 1.75$, $p = 0.09$ (two-tailed). There was no significant difference between student efficacy related to Framing ($M = 5.0$, $SD = 1.4$) and Analyzing & Concluding ($M = 5.0$, $SD = 1.2$), $t(56) = 0.32$, $p = 0.75$ (two-tailed) or between student efficacy related to Designing ($M = 5.3$, $SD = 1.3$) or Collecting ($M = 5.3$, $SD = 1.3$), $t(55) = 0.18$, $p = 0.86$ (two-tailed). These data indicated that students tended to feel more efficacious about their abilities to successfully complete the two middle phases of the science inquiry write-up and less efficacious about their abilities to successfully frame an investigation and communicate the results.

Table 93: Summary results from the 3rd administration of the Chemistry Self-Efficacy Survey of mean Science Inquiry Self-Efficacy by inquiry category

Framing Self-Efficacy		Designing Self-Efficacy		Collecting and Presenting Self-Efficacy		Analyzing & Concluding Self-Efficacy	
Mean	SD	Mean	SD	Mean	SD	Mean	SD
5.0	1.4	5.3	1.3	5.3	1.3	5.0	1.2

3rd Administration CSES – Effects of Gender

In order to assess the role of gender on both Chemistry Content Self-Efficacy and Science Inquiry Self-Efficacy, the survey results were analyzed with an independent-samples t-test. In Chemistry Content Self-Efficacy, female students (M = 4.9, SD = 1.1) tended to have higher scores than their male counterparts (M = 4.7, SD = 1.4), but this difference was not significant; t(55) = 0.44, p = 0.66 (two-tailed). In a similar manner, female students (M = 5.1, SD = 1.0) tended to score slightly higher on the Science Inquiry Self-Efficacy scale than the male students (M = 5.0, SD = 1.4). Again this difference was not significant; t(55) = 0.48, p = 0.64 (two-tailed).

Table 94: Comparison of mean student scores in Chemistry Content Self-Efficacy and Science Inquiry Self-Efficacy by gender from the 3rd administration of the CSES

	Chemistry Content Self-Efficacy		Science Inquiry Self-Efficacy	
	Mean	SD	Mean	SD
Male	4.7	1.4	5.0	1.4
Female	4.9	1.1	5.1	1.0

3rd Administration CSES – Effects of Achievement

In order to assess how this study affected high and low achieving students, the mean scores from high and low achieving students were also compared. High achieving students were categorized as those students whose final first semester grades were 85% or higher. As expected, high achieving students (M = 5.2, SD = 1.1) had significantly higher Chemistry Content Self-Efficacy scores than low achieving students (M = 4.3, SD = 1.2); t(55) = 2.860, p = 0.006 (two-tailed). The magnitude of this difference in means (mean difference = 0.88, 95% CI: 0.26 to 1.49), though, was moderate to large (eta squared = 0.13). In contrast, even though high achieving students (M = 5.2, SD = 1.2) tended to have higher Science Inquiry Self-Efficacy scores than low achieving students (M = 4.8, SD = 1.1), this difference was not significant; t(55) = 1.50, p = 0.14 (two-tailed). Table 95 summarizes these results.

Table 95: Comparison of mean student scores in Chemistry Content Self-Efficacy and Science Inquiry Self-Efficacy by prior student achievement from the 3rd administration of the CSES

	Chemistry Content Self-Efficacy		Science Inquiry Self-Efficacy	
	Mean	SD	Mean	SD
High Achievers	5.2	1.1	5.2	1.2
Low Achievers	4.3	1.2	4.8	1.1

3rd Administration CSES – Comparison to Prior Administrations: Scale Scores

A one-way repeated measures ANOVA was used to assess whether growth had taken place regarding Chemistry Content Self-Efficacy. Since all of the classes had statistically similar scores, all of the classes were grouped together as one large group for the purposes of comparison. The means from the Chemistry Content Self-Efficacy scale were compared over the first three administrations of the survey. Across the three time periods, Chemistry Content Self-Efficacy tended to increase over time.

Means and standard deviations from this test are in table 96. There was a significant effect for time, Wilks' Lambda = 0.55, $F(2,47) = 19.47$, $p < 0.0005$, multivariate partial eta squared = 0.45 indicated a very large effect size for time. Pairwise comparisons indicated that the significant differences in Chemistry Content Self-Efficacy means was found between the 1st and 2nd administration of the survey ($p < 0.0005$) and the 1st and 3rd administrations of the survey ($p < 0.0005$). The difference in Chemistry Content Self-Efficacy means between the 2nd and 3rd administrations was not significant ($p = 0.82$).

Table 96: Mean student Chemistry Content Self-Efficacy comparing 2nd and 3rd administrations to back to baseline data

	n	Mean	SD
Baseline	49	3.9	1.1
2nd Administration	49	4.7	1.0
3rd Administration	49	4.9	1.3

The same analysis was run using mean Science Inquiry Self-Efficacy scores. Again, all of the class means were assessed together as one large group for the purpose of comparison over time. Generally speaking, Science Inquiry Self-Efficacy scores generally decreased over time (see table 97). The effect for time was significant, Wilks' Lambda = 0.88, $F(2,47) = 3.27$, $p = 0.047$ and the multivariate partial eta squared (0.12) indicated that this was a moderate to large effect size. Follow up pairwise comparisons using a Bonferroni adjustment indicated that the only significant difference in mean Science Inquiry Self-Efficacy scores was between the 1st and 3rd administrations of the survey ($p = 0.039$).

Table 97: Comparison of mean student Science Inquiry Self-Efficacy scores comparing 2nd and 3rd administrations to back to baseline data

	n	Mean	SD
Baseline	49	5.5	0.9
2nd Administration	49	5.3	0.9
3rd Administration	49	5.1	1.1

A follow-up mixed between-within subjects ANOVA was conducted in order to compare the students in the control group to students in the experimental groups because it was noticed that even thought Science Inquiry Self-Efficacy scores decreased over time for all of the classes, the Science Inquiry Self-Efficacy scores seemed to drop faster for the students in the control group. The interaction between experimental condition and time was not significant, Wilks' Lambda = 0.96, $F(2,46) = 0.96$, $p = 0.39$, partial eta squared = 0.04, while the main effect for time was significant as expected, Wilks' Lambda = 0.86, $F(2,46) = 3.71$, $p = 0.032$, partial eta squared =0.14. There was no significant difference in the drop in means between the control group and the experimental group, $F(1,47) = 1.19$, $p = 0.28$, partial eta squared = 0.03.

3rd Administration CSES – Comparison to Prior Administrations: Gender

A mixed between-within subjects ANOVA was also conducted in order to assess the role of gender in self-efficacy in the science classroom over time. The interaction between gender and time was not significant for either Chemistry Content Self-Efficacy (Wilks' Lambda = 0.99, $F(2,46) = 0.32$, $p = 0.73$, partial eta squared =

0.01) or for Science Inquiry Self-Efficacy (Wilks' Lambda = 0.98, $F(2,46) = 0.41$, $p = 0.66$, partial eta squared = 0.02). The main effect for time was only significant for the Chemistry Content Self-Efficacy Scale (Wilks' Lambda = 0.55, $F(2,46) = 18.72$, $p < 0.0005$, partial eta squared = 0.45) while the main effect for time only approached significance for the Science Inquiry Self-Efficacy scale (Wilks' Lambda = 0.88, $F(2,46) = 3.14$, $p = 0.053$, partial eta squared = 0.12). Even though scores for Chemistry Content Self-Efficacy generally increased and scores for Science Inquiry Self-Efficacy generally decreased (see tables 98 & 99), there was no difference in either scale over time when gender was considered (Chemistry Content Self-Efficacy effect of gender: $F(1,47) = 0.003$, $p = 0.96$, partial eta squared = 0.00; Science Inquiry Self-Efficacy effect of gender: $F(1,47) = 0.07$, $p = 0.79$, partial eta squared = 0.001).

Table 98: Comparison of mean Chemistry Content Self-Efficacy scores by gender from the 2nd and 3rd administrations back to baseline data

	Male – Chemistry Content			Female - Chemistry Content		
	n	Mean	SD	n	Mean	SD
Baseline	20	3.9	0.9	29	3.9	1.2
2nd Administration	20	4.8	1.3	29	4.7	0.8
3rd Administration	20	4.8	1.5	29	4.9	1.1

Table 99: Comparison of mean Science Inquiry Self-Efficacy scores by gender from the 2nd and 3rd administrations back to baseline data

	Male – Science Inquiry			Female – Science Inquiry		
	n	Mean	SD	n	Mean	SD
Baseline	20	5.5	0.8	29	5.5	0.9
2nd Administration	20	5.4	0.9	29	5.2	0.9
3rd Administration	20	5.1	1.3	29	5.1	1.0

3rd Administration CSES – Comparison to Prior Administrations: Achievement

To assess the role of achievement and see how this changed over time, a mixed between-within subjects ANOVA was run for the CSES scores over the three administrations. Beginning with the Chemistry Content Self-efficacy scores, the interaction between time and achievement was not significant (Wilks' Lambda = 0.88, $F(2,46) = 3.07$, $p = 0.06$, partial eta squared = 0.12). The main effect for time was significant and substantial (Wilks' Lambda = 0.56, $F(2,46) = 18$, $p < 0.0005$, partial eta squared = 0.44). Both groups showed an increase in their Chemistry Content Self-Efficacy scores. Not surprisingly, the main effect for achievement was also significant ($F(1,47) = 8.61$, $p = 0.005$, partial eta squared = 0.16) which indicated that students who had higher achievement also tended to have high self-efficacy over time.

The same pattern held true when looking at the Science Inquiry Self-Efficacy scores over time. There was no significant interaction between time and achievement (Wilks'Lambda = 0.96, $F(2,46) = 0.89$, $p = 0.42$, partial eta squared = 0.04). The main effect for time was significant (Wilks' Lambda = 0.86, $F(2,46) = 3.84$, $p = 0.03$, partial eta squared = 0.14). Both groups showed a decrease in Science Inquiry Self-Efficacy. The main effect for achievement was also significant ($F(1,47) = 5.95$, $p = 0.019$, partial eta squared = 0.11) which implied that the self-efficacy of high achieving students did not fall as fast as the self-efficacy of low achieving students. The means for each group for this analysis are summarized in tables 100 & 101.

Table 100: Comparison of mean Chemistry Content Self-Efficacy scores by prior student achievement from the 2nd and 3rd administrations back to baseline data

	High Achievers – Chemistry Content			Low Achievers – Chemistry Content		
	n	Mean	SD	n	Mean	SD
Baseline	29	4.0	0.9	20	0.7	1.3
2nd Administration	29	5.0	0.7	20	4.2	1.2
3rd Administration	29	5.3	1.1	20	4.2	1.2

Table 101: Comparison of mean Science Inquiry Self- Efficacy scores by prior student achievement from the 2nd and 3rd administrations back to baseline data

	High Achievers - Science Inquiry			Low Achievers - Science Inquiry		
	n	Mean	SD	n	Mean	SD
Baseline	29	5.7	0.9	20	5.3	0.8
2nd Administration	29	5.5	0.8	20	5.0	1.0
3rd Administration	29	5.4	1.1	20	4.7	1.2

Third Inquiry Task

3rd Inquiry Task Analysis – Comparison of Third Inquiry Task Scores

This third and final inquiry task took place approximately 29 weeks into the school year. Again, out of the four classes, one was a control group and the other three classes were part of the experimental condition. The treatments and treatment groups were identical to the second inquiry task. Whereas the prior two tasks were more teacher directed, this third inquiry task more open. This third task took place following a unit on Acids and Bases. Students were asked to modify a titration lab that they had done during the unit in order to answer a question of their choosing. A couple examples of questions that some students asked were: How does the acidity of diet coke compare to regular coke? Is Aspirin more acidic than Ibuprofen? How does the acidity of different types of vinegar compare? Which sport drinks have the highest acid content? How does the acid content of normal orange juice compare to orange juice made from concentrate? Students were asked to make conclusions and use their data to defend their ideas.

Identical to the 2nd inquiry task, the students were split up into four groups by class period. The second period class was the control group and no changes were made to the instruction that this class received. The fourth period class reviewed the instruction they received from the 2nd inquiry task and were also asked to adopt a goal related to their project. Periods five and six also reviewed the instruction they received from the second inquiry task. Similar to the 2nd inquiry task, these two classes were also asked to write down their goals and also reflect on their inquiry task at various times during the project. Students in the 5th period class were given specific prompts to help guide their reflections whereas students in the 6th period class were only given generic reflection prompts.

Table 102: Summary results of mean inquiry task scores and mean Framing and Analyzing & Concluding section scores by period for the 3rd inquiry task

Period	Inquiry Task		Framing		Analyzing & Concluding	
	Mean Score	SD	Mean Score	SD	Mean Score	SD
2	11.9	4.4	3.1	0.5	3.2	0.4
4	13.3	4.0	3.2	0.8	3.2	0.6
5	14.0	2.8	3.3	0.8	3.1	1.1
6	12.7	4.8	3.6	0.4	3.4	0.6

A one-way ANOVA was used to compare class means for this third inquiry task. The overall scores on this third inquiry task ranged from 11.9 to 14.0 where a score of 16 would have been needed to meet the benchmark standards. Even though the treatment conditions all had higher mean inquiry task scores than the control group (see table 102), this difference was not significant ($F(3,57) = 0.595$, $p = 0.62$). A follow up independent-samples t-test was run to compare the mean inquiry scores for students control group ($M = 11.9$, $SD = 4.4$) with all of the classes in the experimental group ($M = 13.3$, $SD = 4.1$). Again, although the classes in the treatment conditions had the higher mean inquiry scores, there was no significant difference between the control and the aggregate treatment condition; $t(59) = 0.974$, $p = 0.33$ (two-tailed).

In addition to total inquiry scores, Framing section scores were also analyzed with a one-way ANOVA. The means on the Framing section scores ranged from 3.1 to 3.6 where a score of 4 would have been needed to meet benchmark standards (see table 102). Again, students in the treatment conditions all tended to receive higher scores than students in the control group. These differences though, were not significant; $F(3,53) = 1.57$, $p = 0.21$. When the Analyzing & Concluding mean scores were analyzed by class period, the means ranged from 3.1 to 3.4 (see table 102). In this case though, there was no real pattern to the data when treatment and control groups were compared. A one-way ANOVA confirmed that there were no real differences between the groups ($F(3,53) = 0.38$, $p = 0.77$.

3rd Inquiry Task Analysis – Score and Efficacy Comparison

The relationship between a student's self-efficacy for science inquiry and their actual inquiry score was investigated using Pearson's product-moment correlation coefficient. A preliminary analysis was performed in order to ensure that the assumptions of normality, linearity, and homoscedasticity were not violated. There was a moderate positive correlation between a student's self-efficacy rating and their science inquiry score, $r = 0.37$, $n = 55$, $p = 0.008$ (two-tailed) which indicated that students with a greater sense of self-efficacy for science inquiry also tended to receive higher scores on their written inquiry tasks. This correlation was slightly higher than the correlation for the initial inquiry task ($r = 0.35$), yet slightly lower than the correlation for the second inquiry task ($r = 0.38$).

3rd Inquiry Task Analysis – Word Count Study

For this third inquiry task, similar to the previous inquiry tasks, word counts in each section were analyzed with one-way ANOVA's to determine if there were differences between groups in the mean number of words that they used in Framing and Analyzing & Concluding sections. In the Framing section, student mean word counts ranged from 203 to 242. These differences were not significant (F(3,51) = 0.39, p = 0.76. A similar result was found in the mean word counts for the Analyzing & Concluding section where the means ranged from 193 to 292. Again, this difference was not significant (F(3,51) = 1.12, p = 0.35.

Table 103: Summary results of mean Framing and mean Analyzing & Concluding word count by period for the 3rd inquiry task

Period	Framing		Analyzing & Concluding	
	Mean # of Words	SD	Mean # of Words	SD
2	203	70	193	71
4	207	122	233	107
5	233	124	292	181
6	242	117	261	135

A follow-up analysis was run to look at the relationship between word count and score using a Pearson product-moment correlation coefficient. An initial scatter plot analysis ensured that no assumptions regarding normality, linearity, and homodescedasticity were violated. As was expected, the correlation between words in a section and the subsequent score was large for both the Framing ($r = 0.74$, $n = 55$, $p < 0.0005$) and the Analyzing & Concluding ($r = 0.70$, $n = 55$, $p < 0.0005$) sections. Students who tended to use more words were the same students who also tended to score higher marks. On the other hand, the number of words in one section were only modestly correlated to the number of words in the other ($r = 0.44$, $n = 55$, $p = 0.001$). The scores on each section, though, showed a large positive correlation ($r = 0.58$, $n = 57$, $p < 0.0005$). All of these correlations were significant.

Table 104: Pearson product-moment correlations between numbers of words used and section scores for the 3rd inquiry task

	1.	2.	3.	4.
1. Words in Framing Section		0.44*	0.74**	
2. Words in Analyzing & Concluding Section				0.70**
3. Framing Section Score				0.58**
4. Analyzing & Concluding Section Score				

* p < 0.005 (2-tailed) ** p < 0.0005 (2-tailed)

3rd Inquiry Task Analysis – Sentence Stem Use Study

Since student scores were similar across all of the class periods, each student work sample was assessed to see whether or not the given sentence stems were used. Out of the 46 students in the experimental group, only 7 students (15%) used the sentence stems in their Framing section while slightly more students (n = 12, % = 23) used the sentence stems in their Analyzing & Concluding section. In the Framing section, sentence stem use was split evenly between males (n = 4) and females (n = 3). In the Analyzing & Concluding section, though, females (n = 9) tended to use the sentence stems more than their male counterparts (n = 4). This difference, though, was insignificant χ^2 (1, n = 55) = 0.60, p = 0.44, phi = -0.11.

Interestingly, one of the students that used sentence stems in their Analyzing & Concluding section came from the control group. This was evidence that students in different classes may have been working together outside of the school day. Many students would often collaborate with students in other classes as they studied after school hours, so that students may have also been helping each other with their inquiry tasks would not be surprising.

Another test was run to see the distribution of sentence stem use by prior achievement. While the numbers were too low to do any statistical analysis, higher achieving students generally used the sentence stems more than their lower achieving counterparts. This difference, though, was not very large and probably would not have been significant.

Table 105: Comparison of student sentence stem use across experimental group conditions on the 3rd inquiry task

Period	Number of Students using Sentence Stems in the Framing Section	Number of Students using Sentence Stems in the Analyzing & Concluding Section
4 (n = 15)	4	5
5 (n = 14)	1	1
6 (n = 17)	2	6

Table 106: Comparison of student sentence stem use by gender on the 3rd inquiry task

	Number of Students using Sentence Stems in the Framing Section	Number of Students using Sentence Stems in the Analyzing & Concluding Section
Females (n = 15)	3	9
Males (n = 17)	4	4

Table 107: Comparison of student sentence stem use by prior student achievement on the 3rd inquiry task

	Number of Students using Sentence Stems in the Framing Section	Number of Students using Sentence Stems in the Analyzing & Concluding Section
High (n = 34)	5	8
Low (n = 21)	2	5

These results were subsequently analyzed in order to assess the extent to which there was a relationship between sentence stem use and scores. An independent samples t-test compared the Framing section scores from students who used the sentence stems to students who did not use the stems. The difference in scores between those who did use the sentence stems (M = 4.07, SD = 0.53) and those who did not use the sentence stems (M = 3.25, SD = 0.60) was significant; t(53) = 3.46, p = 0.001 (two-tailed). The magnitude of the difference (mean difference = 0.83, 95% CI: 0.35 to 1.30) was quite large (eta squared = 0.18). This was not the case though, for the relationship between sentence prompt use and score for the Analyzing & Concluding section. There was no significant difference between those who used the sentence stems (M = 3.37, SD = 0.59) and those who did not (M = 3.18, SD = 0.77); t(53) = 0.80, p = 0.43 (two-tailed).

Table 108: Comparison of student section scores by sentence stem use on the 3rd inquiry task

	Framing		Analyzing & Concluding	
	Mean Score	SD	Mean Score	SD
Did use Sentence Stems	4.07	0.53	3.37	0.59
Did not use Sentence Stems	3.25	0.60	3.18	0.77

Another independent samples t-test was run to compare the relationship between sentence prompt use and number of words used in each section. Similar to the student scores, students who did use the sentence stems tended to write more words (M = 324, SD = 108) than their counterparts who did not use the sentence stems (M = 209, SD = 106). This difference was significant; t(53) = 2.68, p = 0.01 (two-tailed). The magnitude of the differences in the means (mean difference = 115, 95% CI: 29 to 201) was moderate (eta squared = 0.12). The difference in numbers of words by sentence prompt use for the Analyzing & Concluding section was not significant; t(53) = -0.21, p = 0.84 (two-tailed). In fact, students who used the stems tended to write fewer words (M = 244, SD = 97) than those who did not use the sentence stems (M = 253, SD = 145).

Table 109: Comparison of number of words written by sentence use on the 3rd inquiry task

	Framing		Analyzing & Concluding	
	Mean # of Words	SD	Mean # of Words	SD
Did use Sentence Stems	324	108	244	97
Did not use Sentence Stems	209	106	253	145

3rd Inquiry Task Analysis – Effects of Goal Type

Similar to the 2nd inquiry task, two of the four classes were asked to adopt goals related to their science inquiry tasks. Students were asked to adopt goals related to both the lab portion of the inquiry in addition to the writing portion of the inquiry task. Before students were asked to write goals, time was spent in the class talking about the difference between distal and proximal goals. Again, it was emphasized that proximal goals tend to lead to better products because they focus in immediate actions whereas distal goals tend to be outcome focused without specifying actions necessary to reach those goals. Instruction in goal setting was given to all three classes in the experimental condition, although only two of these classes were asked to write down their goals. In addition, students were encouraged to write a goal related to using the sentence stems as they wrote their projects. The students had the freedom, though, to choose any goal they wanted. Students turned in copies of these goals when they handed in their projects.

Tables 110 & 111 summarize goal types by class for the Framing and the Analyzing & Concluding sections. Approximately equal numbers of students turned in goal sheets with their science inquiry tasks in both classes. The students were also fairly evenly split between proximal and distal goals for both the Framing and Analyzing & Concluding sections as well. While most of the students wrote goals for both the Framing and Analyzing & Concluding sections, there were a few students whose goals only pertained to the Analyzing & Concluding section.

Table 110: Comparison of student goal orientations for the Framing section on the 3rd inquiry task

Period	Proximal Goal	Distal Goal	No Goal
5 (n = 12)	6	5	1
6 (n = 15)	7	5	3

Table 111: Comparison of student goal orientations for the Analyzing & Concluding section on the 3rd inquiry task

Period	Proximal Goal	Distal Goal	No Goal
5 (n = 12)	6	6	0
6 (n = 15)	7	8	0

An independent-samples t-test was run to determine if goal type had any effect on student section scores. The mean Framing score for students who had adopted proximal goals (M = 3.4, SD = 0.6) was slightly lower than for those that had adopted distal goals (M = 3.6, SD = 0.5), but this difference was not significant, t(21) = 1.14, p = 0.27 (two-tailed). When the scores of the Analyzing & Concluding section were analyzed, there was no difference between the mean scores for either the proximal (M = 3.4, SD = 0.6) or distal (M = 3.4, SD = 0.6) goal groups, t(25) = 0.21, p = 0.83 (two-tailed). The same was true for the differences in word counts by goal type. Even though students who had chosen distal goals (M = 295, SD = 155) wrote more words in their Framing section than the students who had adopted proximal goals (M = 219, SD = 107), this difference was not significant, t(21) = 1.40, p = 0.18. In the same way, students who had adopted distal goals wrote more words in their Analyzing & Concluding section (M = 285, SD = 160) than their peers who had chosen proximal goals (M = 270, SD = 140), yet again, this difference was not significant, t(25) = 0.27, p = 0.79.

Table 112: Comparison of mean student section scores by goal orientation on the 3rd inquiry task

Goal Type	Framing				Analyzing & Concluding			
	Mean Score	SD	Mean # of Words	SD	Mean Score	SD	Mean # of Words	SD
Proximal	3.4	0.6	219	107	3.4	0.6	270	140
Distal	3.6	0.5	295	155	3.4	0.6	285	160

Following this comparison, student work was also assessed in order to see if those students who had adopted goals to use sentence stems did indeed use these stems in their writing. As instruction was given around the creation of writing goals, students were urged to adopt the goal of using three sentence stems in their Framing sections and also to use three sentence stems in the Analyzing & Concluding sections. The results indicated that most of the students did not actually use the sentence stems as they wrote their inquiry projects.

Table 113: Comparison of student goal to use sentence stems compared to actual use of sentence stems in the Framing section of the 3rd inquiry task

Period	Goal to use sentence stems and did use sentence stems	Goal to use sentence stems but did not use sentence stems	No goal to use sentence stems but did use sentence stems	No goal to use sentence stems and did not use sentence stems
5 (n = 12)	1	6	0	5
6 (n = 15)	1	5	2	7

Table 114: Comparison of student goal to use sentence stems compared to actual use of sentence stems in the Analyzing & Concluding section of the 3rd inquiry task

Period	Goal to use sentence stems and did use sentence stems	Goal to use sentence stems but did not use sentence stems	No goal to use sentence stems but did use sentence stems	No goal to use sentence stems and did not use sentence stems
5 (n = 12)	1	6	0	5
6 (n = 15)	3	3	2	7

3rd Inquiry Task Analysis – Post-Task Student Predictions

Similar to the prior inquiry task, students were again asked to reflect on their inquiry tasks prior to handing them in to be graded. On this reflection, the students were asked to score themselves against the state scoring guide on the four components of their inquiry task (Framing, Designing, Data Collecting, and Analyzing & Concluding). Again the intent was to determine the extent to which students understood the language of the scoring guide and were also able to successfully apply this understanding to critiquing their own work. The results presented in table 115 were somewhat mixed. Only half of the student students who turned in their reflection sheets were able to correctly predict their scores for the Framing section. This percentage increased slightly when students were asked to assess their Analyzing & Concluding sections. The percentage of students who were able to correctly predict their section scores decreased from the 2nd to the 3rd inquiry tasks.

Table 115: Comparison of student predicted scores to actual scores received on the 3rd inquiry task

	Correctly Predicted Section Score		Incorrectly Predicted Section Score	
	n	%	n	%
Framing	22	50.0	22	50.0
Analyzing & Concluding	28	63.6	16	36.4

To look at how motivated students were to achieve a passing grade, student predictions were sorted into one of two groups.

Group 1: Students who predicted a passing score

Group 2: students who predicted a non-passing score

Again, each section of the science inquiry task was graded on a scale of 0 to 6 where a score of 4 was considered to meet benchmark standards. It was clear from the student feedback that many of them felt that their Framing sections were adequate to meet benchmark standards which would explain the discrepancy between predicted versus actual student scores. The data from the Analyzing & Concluding section, though, was more troubling. The students were better able to assess their work against the state scoring guide, and yet just slightly under half of the students predicted that their Analyzing & Concluding would not meet benchmark standards. Even though the data implied that students were able to understand and interpret the state scoring guide, many students still turned in projects that they knew did not meet benchmark standards.

Table 116: Comparison of student score predictions on the 3rd inquiry task

	Predicted a Passing Score		Predicted a Non-Passing Score	
	n	%	n	%
Framing	33	75.0	11	25.0
Analyzing & Concluding	25	56.8	19	43.1

3rd Inquiry Task Analysis – Effects of Gender

In order to look at the effects of experimental group on student scores by gender, a two-way between-groups ANOVA was conducted. The interaction between class period and gender was not significant, $F(3,53) = 1.16$, $p = 0.34$. Neither of the main effects for gender ($F(1,53) = 0.29$, $p = 0.59$) or for experimental condition ($F(3,53) = 0.28$, $p = 0.84$) were statistically significant as well. The implication was that the experimental conditions did not have a significant effect on male or female student scores.

A follow-up independent-samples t-test was run to compare overall male and female student scores on this inquiry task. Overall, male students had a higher mean score (M = 13.2, SD = 3.5) than their female counterparts (M = 13.0, SD = 4.5) but this difference was not significant, t(59) = 0.75, p = 0.86 (two-tailed). Interestingly, female students tended to score higher (M = 3.4, SD = 0.6) than males (M = 3.25, SD = 0.7) on the Framing section and also on the Analyzing & Concluding section (females: M = 3.3, SD = 0.8; males M = 3.2, SD = 0.6). Again neither of these results were statistically significant (Framing: t(55) = 0.81, p = 0.42 (two-tailed); Analyzing & Concluding: t(55) = 0.42, p = 0.68). Independent-samples t-tests were also used to assess differences in the mean number of words that male and female students used as they wrote. In the Framing section, there was no significant difference between the mean number of words used by male (M = 216, SD = 95) and female (M = 229, SD = 123) students, t(53) = 0.41, p = 0.69 (two-tailed). The same was true of the difference in mean words in the Analyzing & Concluding section (males M = 215, SD = 114; females M = 275, SD = 143), t(53) = 1.64, p = 0.11 (two-tailed).

Table 117: Comparison of mean scores by gender on the 3rd inquiry task

	3rd Inquiry Task		Framing				Analyzing & Concluding			
	Mean Score	SD	Mean Score	SD	Mean # of Words	SD	Mean Score	SD	Mean # of Words	SD
Male	13.2	3.5	3.3	0.7	216	95	3.2	0.6	215	114
Female	13.0	4.5	3.4	0.6	229	123	3.3	0.8	275	143

3rd Inquiry Task Analysis – Effects of Achievement

In addition to gender, this study also sought to investigate how student achievement and experimental condition impacted student scores on their inquiry tasks. A two-way between subjects ANOVA was conducted to determine if experimental group condition had any impact on total inquiry scores for the 3rd inquiry task. The interaction effect between achievement and class period was not statistically significant, $F(3,53) = 0.34$, $p = 0.80$. The main effect of experimental condition was also not significant, $F(3,53) = 0.77$, $p = 0.52$. As expected, though, the main effect for achievement was significant, $F(1,53) = 18.25$, $p < 0.0005$, with a rather large effect size (partial eta squared = 0.26). This result indicated that student achievement did affect students' scores but experimental condition did not.

A number of follow up independent sample t-tests were conducted that compared the mean scores of high achieving and low achieving students on both the total inquiry score in addition to the Framing and Analyzing & Concluding section scores. As expected, high achieving students mean total inquiry scores ($M = 14.8$, $SD = 1.6$) where significantly higher than their lower achieving counterparts ($M = 10.6$, $SD = 5.2$), $t(59) = 4.49$, $p < 0.0005$. The magnitude of the mean differences (mean difference = 4.2, 95% CI: 2.30 to 6.00) was only moderate (eta squared = 0.06). In the same way, the mean differences between high achieving students ($M = 3.4$, $SD = 0.5$) and low achieving students ($M = 2.9$, $SD = 1.0$) on the Analyzing & Concluding section was significant, $t(55) = 2.56$, $p = 0.13$. Again, the magnitude of these differences (mean difference = 0.50, 95% CI : 0.11 to 0.89) was moderate (eta squared = 0.11). Interestingly, the difference in mean scores between high achievers ($M = 3.5$, $SD = 0.61$) and low achievers ($M = 3.1$, $SD = 0.76$) in the Framing section was not statistically significant, $t(55) = 1.81$, $p = 0.08$. In the same way, the difference between the mean number of words in the Framing section by achievement were not significant (high achievers M = 245, SD = 121; low achievers M = 192, SD = 92), t(53) = 1.75, p = 0.09. For the Analyzing & Concluding section, high achieving students did tend to use more words (M = 290, SD = 128) than did the low achieving students (M = 191, SD = 121) in this section. This difference was significant, t(53) = 2.84, p = 0.006.

Table 118: Comparison of mean inquiry scores by prior student achievement on the 3rd inquiry task

	3rd Inquiry Task		Framing				Analyzing & Concluding			
	Mean Score	SD	Mean Score	SD	Mean # of Words	SD	Mean Score	SD	Mean # of Words	SD
High Achievers	14.8	1.6	3.5	0.6	245	121	3.4	0.5	290	128
Low Achievers	10.6	5.2	3.1	0.7	192	92	2.9	1.0	191	121

3rd Inquiry Task Analysis – Comparison to Prior Tasks: Introduction

The purpose of these next sections was to compare student results on the 3rd inquiry task back to the prior two inquiry tasks. The intent was to see what effects, if any, were produced by the various treatment conditions and to track the effect of these treatment conditions over time. Because sample sizes in this study tended to be small and also a large number of statistics were run, the choice was made to go with the more conservative multivariate analysis of variance statistic.

3rd Inquiry Task Analysis – Comparison to Prior Tasks: Inquiry Task Scores

Total inquiry scores from the 3rd inquiry task were compared back to total inquiry scores from the 2nd inquiry task and also the baseline inquiry task. The comparison was made using a mixed between-within subjects ANOVA. There was no significant interaction between time and experimental condition, Wilks' Lambda = 0.85, F(6,100) = 1.47, p = 0.20, partial eta squared = 0.08. There was a substantial main effect for time, Wilks' Lambda = 0.79, F(2,50) = 6.53, p = 0.003, partial eta squared = 0.21 with all four groups showing changes in their total inquiry scores. Even though all of the groups in the experimental condition showed increases in total inquiry scores over time and the control group's total inquiry scores dropped, the main effect for experimental condition was not significant, F(3,51) = 0.02, p = 0.99, partial eta squared = 0.001, indicating that experimental condition did not make any difference. A mixed between-within subjects ANOVA confirmed that there was not any difference over time between the control group and the experimental groups (F(1,53) = 0.01, p = 0.92).

Table 119: Comparison of mean inquiry scores by period comparing results from the 2nd and 3rd inquiry tasks back to baseline data

	Per 2 (n = 7)		Per 4 (n = 15)		Per 5 (n = 15)		Per 6 (n = 18)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Baseline	13.4	1.1	11.9	2.8	12.1	2.2	12.8	2.9
2nd Inquiry	13.3	1.4	13.5	3.1	13.7	1.8	13.0	2.9
3rd Inquiry	13.3	1.7	14.2	1.9	14.0	2.8	13.7	2.8

Period 2 – Control Group, Period 4 - Sentence Stems Only, Period 5 - Sentence Stems + Directed Reflection, Period 6 - Sentence Stems + General Reflection

Since the main effect for time was significant, a number of paired-samples t-tests were run in order to determine if the increased scores were due to improvements in the Framing or the Analyzing & Concluding sections. Similar to the 2nd inquiry task, since all of the groups were statistically equivalent, the students were not divided into groups for this analysis. The total inquiry score increased with each successive inquiry task. The total inquiry score increase from baseline ($M = 12.7$, $SD = 2.3$) to the 3rd inquiry ($M = 14.1$, $SD = 2.1$) was significant; $t(53) = 4.97$, $p < 0.0005$ (two-tailed), eta squared = 0.32, as was the increase from the from the 2nd inquiry ($M = 13.5$, $SD = 2.3$) and the 3rd inquiry ($M = 14.1$, $SD = 2.1$); $t(53) = 2.12$, $p = 0.04$ (two-tailed), eta squared = 0.08.

Section scores were also compared. Mean scores from the Framing section increased slightly over time. From the baseline inquiry, the mean score increased from 3.2 ($SD = 0.6$) to 3.3 ($SD = 0.7$) on the 3rd inquiry task. This increase was not significant, $t(53) = 1.35$, $p = 0.18$ (two-tailed). The same was true of the comparison between the Framing section scores from the 2nd inquiry ($M = 3.3$, $SD = 0.6$) and the 3rd inquiry ($M = 3.3$, $SD = 0.7$); $t(53) = 0.29$, $p = 0.77$ (two-tailed). In the Analyzing & Concluding sections, scores generally increased as well. From the baseline inquiry to the 3rd inquiry task, scored increased from a mean of 2.9 ($SD = 0.7$) to a mean of 3.2 ($SD = 0.8$). This increase was significant ($t(53) = 3.11$, $p = 0.003$ (two-tailed), eta squared = 0.15). Even though the increase in mean score between the 2nd inquiry task ($M = 3.0$, $SD = 1.0$) and the 3rd inquiry task ($M = 3.2$, $SD = 0.8$) was not as large, the increase was still significant, $t(53) = 2.16$, $p = 0.04$ (two-tailed), eta squared = 0.08. These results implied that of the two sections that have been assessed in this study, students generally saw improvement only in the Analyzing & Concluding section. It is also worth noting that whereas in the baseline task, there was a difference between the Framing section scores and the Analyzing & Concluding section scores (t(57) = 3.24, p = 0.002 (two-tailed)), by the time of the 3rd inquiry task, these scores had moved towards parity.

Table 120: Comparison of mean inquiry scores comparing results from the 2nd and 3rd inquiry tasks back to baseline data

Time period	n	Total Inquiry		Framing		Analyzing & Concluding	
		Mean Score	SD	Mean Score	SD	Mean Score	SD
Baseline	54	12.7	2.3	3.2	0.6	2.9	0.7
2nd Inquiry	54	13.5	2.3	3.3	0.6	3.0	1.0
3rd Inquiry	54	14.1	2.1	3.3	0.7	3.2	0.8

3rd Inquiry Task Analysis – Comparison to Prior Tasks: Word Count

Since scores had generally increased with each successive inquiry task, paired-samples t-tests were conducted to see if word counts had increased as well. Again, for this initial analysis, students were not split up into their separate experimental groups. In each section, the number of words increased over time. In the Framing section, student mean word counts increased from the baseline inquiry of 145 (SD = 51) to 232(SD = 120) on the 3rd inquiry task. This increase was significant, t(42) = 4.51, p < 0.0005 (two-tailed) , eta squared = 0.33. In contrast, the mean increase in word counts between the 2nd inquiry (M = 209, SD = 107) and the 3rd inquiry (M = 232, SD = 120) was not significant, t(42) = 1.51, p = 0.14 (two-tailed). This trend was mirrored with the mean word counts from the Analyzing & Concluding sections. When the mean word counts from the baseline inquiry (M = 214, SD = 124) were compared to those of the 3rd inquiry (M = 263, SD = 140) the increase of 49 words was significant, t(42) = 2.66, p = 0.01 (two-tailed), eta squared = 0.14. The mean word increase of 28 words from the 2nd inquiry (M = 235, SD = 120) to the 3rd inquiry (M = 263, SD = 140) was not significant, t(42) = 1.71, p = 0.09 (two-tailed).

Table 121: Comparison of mean Framing and Analyzing & Concluding word counts from the 2nd and 3rd inquiry tasks back to baseline data

Time period	n	Framing		Analyzing & Concluding	
		Mean # of Words	SD	Mean # of Words	SD
Baseline	43	145	51	214	124
2nd Inquiry	43	209	107	235	120
3rd Inquiry	43	232	120	263	140

A follow-up mixed between-within subjects ANOVA was conducted to determine if experimental condition had any effect on the number of words produced in the Framing section. The general trend was that mean word counts in the control group (period 2) remained fairly flat, while mean word counts in the experimental groups tended to increase. There was no significant interaction between time and experimental condition, Wilks' Lambda = 0.88, F(6,78) = 0.86, p = 0.53, partial eta squared = 0.06. There was a substantial main effect for time, Wilks' Lambda = 0.75, F(2,38) = 6.45, p = 0.004, partial eta squared = 0.25, where all of the groups generally showed an increase in mean word counts over time. The main effect for experimental condition was not significant, F(3,39) = 0.09, p = 0.97, partial eta squared = 0.007, which suggested that there was no difference between any of the conditions and the number of words produced.

A similar comparison was run on the mean word counts from the Analyzing & Concluding section. Again, mean word counts from the students in the experimental groups tended to be higher than the mean word counts for the students in the control group, although the trend was not as clean. A mixed between-within subjects ANOVA was conducted to assess these differences. The interaction between time and experimental group was not significant, Wilks' Lambda = 0.82, $F(6,76) = 1.31$, $p = 0.26$, partial eta squared = 0.09. The main effect for time was also not significant, Wilks' Lambda = 0.89, $F(2,38) = 2.46$, $p = 0.10$, partial eta squared = 0.12 and neither was the main effect of experimental group, $F(3,39) = 0.10$, $p = 0.96$, partial eta squared = 0.008. These results suggested that there was no difference between the groups when comparing word counts from the Analyzing & Concluding section over time.

Table 122: Comparison of Framing section mean word number of words by period from the 2nd and 3rd inquiry tasks back to baseline data

	Per 2 (n = 5)		Per 4 (n = 11)		Per 5 (n = 13)		Per 6 (n = 14)	
	Mean # of Words	SD	Mean # of Words	SD	Mean # of Words	SD	Mean # of Words	SD
Baseline	185	39	153	48	148	63	120	34
2nd Inquiry	172	42	223	133	209	81	211	126
3rd Inquiry	185	71	222	130	242	124	247	128

Period 2 – Control Group, Period 4 - Sentence Stems Only, Period 5 - Sentence Stems + Directed Reflection, Period 6 - Sentence Stems + General Reflection

Table 123: Comparison of Analyzing & Concluding section mean number of words by period from the 2nd and 3rd inquiry tasks back to baseline data

	Per 2 (n = 5)		Per 4 (n = 11)		Per 5 (n = 13)		Per 6 (n = 14)	
	Mean # of Words	SD	Mean # of Words	SD	Mean # of Words	SD	Mean # of Words	SD
Baseline	209	89	191	90	212	102	235	174
2nd Inquiry	224	42	260	125	231	131	222	132
3rd Inquiry	226	60	245	113	306	181	250	140

Period 2 – Control Group, Period 4 - Sentence Stems Only, Period 5 - Sentence Stems + Directed Reflection, Period 6 - Sentence Stems + General Reflection

3rd Inquiry Task Analysis – Comparison to Prior Tasks: Sentence Stem Use

An analysis was done using a mixed between-within subjects ANOVA to assess the impact of sentence stem use and score for the Framing section scores over the three inquiry tasks. There was no significant interaction between time and prompt use, Wilks' Lambda = 0.89, F(2,49) = 2.96, p = 0.6, partial eta squared = 0.11 and the same was true for the main effect of time, Wilks' Lambda = 0.92, F(2,49) = 2.09, p = 0.13, partial eta squared = 0.08. The main effect for sentence stem use was, though, significant, F(1,50) = 8.08, p = 0.006, partial eta squared = 0.14. The students who used the stems tended to be the higher scoring students as indicated by the baseline data. Scores for those who used the sentence stems, though, increased whereas those who did not use the sentence stems saw their scores on the Framing section remain unchanged.

The same analysis was run with the data from the Analyzing & Concluding sections. The interaction between prompt use and time was not significant, Wilks' Lambda = 0.89, F(2,49) = 3.01, p = 0.06. partial eta squared = 0.11. In contrast, the main effect for time was significant, Wilks' Lambda = 0.83, F(2,49) = 4.93, p = 0.01, partial eta squared = 0.17, which suggested that the scores of both groups changed over time. The main effect for prompt use did not turn out to be significant, F(1,50) = 1.99, p = 0.16, partial eta squared = 0.04.

Table 124: Comparison of mean Framing section scores and sentence stem use from the 2nd and 3rd inquiry tasks back to baseline data

	Sentence Stem Use			No Sentence Stem Use		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	7	3.6	0.5	45	3.1	0.5
2nd Inquiry	7	3.5	0.5	45	3.3	0.6
3rd Inquiry	7	4.1	0.5	45	3.2	0.6

Table 125: Comparison of mean Analyzing & Concluding section scores and sentence stem use from the 2nd and 3rd inquiry tasks back to baseline data

	Sentence Stem Use			No Sentence Stem Use		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	13	2.9	0.5	39	2.8	0.8
2nd Inquiry	13	3.5	0.5	39	2.8	1.0
3rd Inquiry	13	3.4	0.6	39	3.2	0.8

In addition to comparing scores over time by sentence stem use, word counts were also compared over time by sentenced stem use. A mixed between-within subjects ANOVA was run to assess the impact of sentence stem use on word counts in the Framing sections over the three inquiry tasks. There was no significant interaction between prompt use and time, Wilks' Lambda = 0.90, F(2,40) = 2.23, p = 0.12, partial eta squared = 0.10. The main effect for time was substantial, Wilks' Lambda = 0.62, F(2,40) = 12.48, p < 0.0005, partial eta squared = 0.38, which indicated that mean word counts for both groups increased over time. The main effect for sentence stem use, though, was also significant, $F(1,41) = 6.48$, $p = 0.02$, partial eta squared = 0.14, which indicated that the students who used sentence stems increasingly used more words than students who did not use the stems.

The same analysis was conducted using word counts from the Analyzing & Concluding sections. There was a significant interaction between time and prompt use, Wilks' Lambda = 0.84, $F(2,40) = 3.87$, $p = 0.03$, partial eta squared = 0.16, which indicated that the word counts for the two groups did not change in the same direction over time. In looking at the data (table 126), it was clear that the number of words used in the Analyzing & Concluding section for students who did not use the stems generally increased whereas there was no trend for those who did use the sentence stems. The main effect for time was also significant, Wilks' Lambda = 0.83, $F(2,40) = 4.09$, $p = 0.02$, partial eta squared = 0.17, which indicated that word counts generally increased over time for the two groups. The main effect for sentence stem use, though, was not significant, $F(1,41) = 0.04$, $p = 0.84$, partial eta squared = 0.001.

Table 126: Comparison of mean Framing section number of words by sentence stem use from the 2nd and 3rd inquiry tasks back to baseline data

	Sentence Stem Use			No Sentence Stem Use		
	n	Mean # of Words	SD	n	Mean # of Words	SD
Baseline	7	155	36	36	143	53
2nd Inquiry	7	287	143	36	194	93
3rd Inquiry	7	324	108	36	214	115

Table 127: Comparison of mean Analyzing & Concluding section number of words by sentence stem use from the 2nd and 3rd inquiry tasks back to baseline data

	Sentence Stem Use			No Sentence Stem Use		
	n	Mean # of Words	SD	n	Mean # of Words	SD
Baseline	10	178	74	33	224	135
2nd Inquiry	10	269	118	33	225	121
3rd Inquiry	10	246	111	33	268	149

3rd Inquiry Task Analysis – Comparison to Prior Tasks: Effects of Goal Type

Since some of the students were asked to adopt goals related to their inquiry tasks, an analysis was done to compare the effect of goal type on student work. A mixed between-within subjects ANOVA was conducted with the scores from the Framing sections. The interaction between goal type and time was not significant, Wilks' Lambda = 0.78, $F(4,46) = 1.53$, $p = 0.21$, partial eta squared = 0.12. The main effect for time was significant, Wilks' Lambda = 0.77, $F(2,23) = 3.38$, $p = 0.05$, partial eta squared = 0.23, which indicated that scores for all groups generally increased over time. The main effect for goal type, though, was not significant, $F(2,24) = 0.07$, $p = 0.94$, partial eta squared = 0.005. This implied that scores generally increased over time, but goal type did not make any difference.

The same analysis was run for the data from the Analyzing & Concluding sections. The interaction between time and goal type was not significant, Wilks' Lambda = 0.85, $F(2,24) = 2.10$, $p = 0.15$, partial eta squared = 0.15. Similar to the

Framing section data, the main effect for time was significant, Wilks' Lambda = 0.62, F(2,24) = 7.26, p = 0.003, partial eta squared = 0.38, which indicated that scores in the Analyzing & Concluding section generally increased over time. Like the Framing section data, the main effect for goal type was not significant, F(1,25) = 0.63, p = 0.43, partial eta squared = 0.03, which indicated that there was no difference in mean scores by goal type.

Table 128: Comparison of mean Framing section score by goal orientation from the 2nd and 3rd inquiry tasks back to baseline data

	Proximal Goal (n = 13)		Distal Goal (n = 10)		No Goal (n = 4)	
	Mean Score	SD	Mean Score	SD	Mean Score	SD
Baseline	3.3	0.3	3.1	0.5	3.2	0.5
2nd Inquiry	3.5	0.4	3.4	0.6	3.1	1.1
3rd Inquiry	3.4	0.6	3.7	0.5	3.6	0.5

Table 129: Comparison of mean Analyzing & Concluding section score by goal orientation from the 2nd and 3rd inquiry tasks back to baseline data

	Proximal Goal (n = 13)		Distal Goal (n = 14)	
	Mean Score	SD	Mean Score	SD
Baseline	2.8	0.6	2.9	1.0
2nd Inquiry	3.3	0.8	2.8	1.1
3rd Inquiry	3.4	0.6	3.4	0.6

In addition to section scores, word counts were also compared by goal orientation. Word counts were compared across goal orientation using a mixed between-within subjects ANOVA. The interaction between time and goal type was not significant, Wilks' Lambda = 0.91, F(4,40) = 0.49, p = 0.74, partial eta squared =

0.05. The main effect for time, though, was substantial and significant, Wilks' Lambda = 0.54, F(2,20) = 8.54, p = 0.002, partial eta squared = 0.46, which indicated that all word counts, regardless of goal orientation, had increased over time. The main effect for goal type also turned out to be significant, F(2,21) = 3.78, p = 0.04, partial eta squared = 0.27. Looking at the data, students in the distal goal group had the greatest growth in their mean word counts whereas students in the proximal goal group and the no goal group were very similar in their mean word counts across all three time periods.

A mixed between-within subjects ANOVA was also conducted by goal type with the mean word counts from the Analyzing & Concluding section. There was no significant interaction between time and goal type, Wilks' Lambda = 0.82, F(2,21) = 2.38, p = 0.12, partial eta squared = 0.19. The main effect for time was also not significant, Wilks' Lambda = 0.78, F(2,21) = 3.03, p = 0.07, partial eta squared = 0.22, which indicated that there was no real change in mean word counts over time. In the same way, there was no significant effect for goal type, F(2,11) = 0.39, p = 0.54, partial eta squared = 0.02, which indicated that there was no difference in mean word counts between the goal type orientations over time.

Table 130: Comparison of mean Framing section number of words by goal orientation from the 2nd and 3rd inquiry tasks back to baseline data

	Proximal Goal			Distal Goal			No Goal		
	n	Mean # of Words	SD	n	Mean # of Words	SD	n	Mean # of Words	SD
Baseline	11	129	51	9	159	57	4	103	32
2nd Inquiry	11	189	67	9	279	125	4	165	118
3rd Inquiry	11	217	116	9	315	149	4	225	40

Table 131: Comparison of mean Analyzing & Concluding section number of words by goal orientation from the 2nd and 3rd inquiry tasks back to baseline data

	Proximal Goal			Distal Goal		
	n	Mean # of Words	SD	n	Mean # of Words	SD
Baseline	11	273	153	13	299	158
2nd Inquiry	11	236	123	13	239	143
3rd Inquiry	11	185	100	13	260	179

3rd Inquiry Task Analysis – Comparison to Prior Tasks: Effects of Gender

Interestingly, over the course of this study, female students tended to do better than their male counterparts. The relationship between gender and science inquiry scores was assessed using a mixed between within-subjects ANOVA. There was no significant interaction between time and gender, Wilks' Lambda = 0.96, $F(2,52) = 0.31$, partial eta squared = 0.05. The main effect for time was substantial, though, (Wilks' Lambda = 0.76, $F(2,52) = 8.10$, $p = 0.001$, partial eta squared = 0.24) which indicated that scores generally increased over time for both males and females. The main effect for gender, though, was not significant, $F(1,53) = 2.51$, $p = 0.12$, partial eta squared = 0.05.

Table 132: Comparison of mean student inquiry scores by gender from the 2nd and 3rd inquiry tasks back to baseline data

	Male			Female		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	23	12.2	3.0	32	12.8	2.1
2nd Inquiry	23	12.5	2.8	32	14.0	2.1
3rd Inquiry	23	13.4	3.4	32	14.2	2.3

A follow up analysis looked at the Framing and Analyzing & Concluding section scores to see if there were any differences by gender. In the Framing section, while scores for both males and females increased over the three inquiry tasks, female students tended to have the higher scores. These scores were analyzed with a mixed between-within subjects ANOVA. For the Framing section, the interaction between time and gender was not significant, Wilks' Lambda = 0.98, $F(2,51) = 0.64$, $p = 0.53$, partial eta squared = 0.03. The same was true of the main effects for time (Wilks' Lambda = 0.97, $F(2,51) = 0.90$, $p = 0.41$, partial eta squared = 0.03) and also gender ($F(1,52) = 1.72$, $p = 0.20$, partial eta squared = 0.03) which indicated that none of the differences were significant.

The results were similar when a mixed between-within ANOVA was run on with the mean scores from the Analyzing & Concluding section. Only the main effect for time was significant (Wilks' Lambda = 0.80, $F(2,51) = 6.21$, $p = 0.004$, partial eta squared = 0.20) which indicated that scores for both genders increased significantly over time. The interaction term was insignificant (Wilks' Lambda = 0.94, $F(2,51) =$

1.73, p = 0.19, partial eta squared = 0.06) as was the main effect for gender (F(1,52) = 3.78, p = 0.06, partial eta squared = 0.07).

Table 133: Comparison of mean student Framing section scores by gender from the 2nd and 3rd inquiry tasks back to baseline data

	Male			Female		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	22	3.2	0.7	32	3.2	0.5
2nd Inquiry	22	3.2	0.7	32	3.4	0.6
3rd Inquiry	22	3.2	0.7	32	3.4	0.6

Table 134: Comparison of mean student Analyzing & Concluding section scores by gender from the 2nd and 3rd inquiry tasks back to baseline data

	Male			Female		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	22	2.7	0.9	32	3.0	0.6
2nd Inquiry	22	2.7	1.0	32	3.2	0.9
3rd Inquiry	22	3.2	0.7	32	3.3	0.8

A similar analysis was done looking at section word counts by gender over time. Again, the analysis were done with a mixed between-within subjects ANOVA. For the Framing section, only the main effect for time was significant (Wilks' Lambda = 0.70, F(2,40) = 8.73, p = 0.001, partial eta squared = 0.30) indicating that both genders saw an increase in their word counts over time. The interaction between time and gender was not significant (Wilks' Lambda = 0.95, F(2,40) = 1.07, p = 0.35, partial eta squared = 0.05) and neither was the main effect for gender (F(1,41) = 0.22, p = 0.64, partial eta squared = 0.01). For the Analyzing & Concluding section, both the interaction term between time and gender (Wilks' Lambda = 0.97, F(2,40) = 0.74, p = 0.49, partial eta squared = 0.04) and the main effect for time (Wilks' Lambda = 0.89, F(2,40) = 2.55, p = 0.09, partial eta squared = 0.11) were not significant. Interestingly, the main effect for gender was significant, F(1,41) = 5.58, p = 0.02, partial eta squared = 0.12, which suggested that females consistently wrote more words than their male counterparts.

Table 135: Comparison of mean Framing section number of words by gender from the 2nd and 3rd inquiry tasks back to baseline data

	Male			Female		
	n	Mean # of Words	SD	n	Mean # of Words	SD
Baseline	14	147	46	29	144	54
2nd Inquiry	14	184	129	29	221	94
3rd Inquiry	14	232	99	29	232	130

Table 136: Comparison of mean Analyzing & Concluding section number of words by gender from the 2nd and 3rd inquiry tasks back to baseline data

	Male			Female		
	n	Mean # of Words	SD	n	Mean # of Words	SD
Baseline	14	170	86	29	233	135
2nd Inquiry	14	165	78	29	267	124
3rd Inquiry	14	205	117	29	291	144

3rd Inquiry Task Analysis – Comparison to Prior Tasks: Effects of Achievement

Finally, this 3rd inquiry task was assessed to determine the effect of prior achievement on inquiry scores. A mixed between-within subjects ANOVA was conducted that compare total inquiry scores over the three inquiries that were given as a part of this study. The interaction between achievement and time was not significant, Wilks' Lambda = 0.95, F(2,52) = 1.31, p = 0.28, partial eta squared = 0.05. As expected, the main effect for time was significant (Wilks' Lambda = 0.74, F(2,52) = 9.06, p < 0.0005, partial eta squared = 0.26) as was the main effect for achievement (F(1,53) = 18.63, p < 0.0005, partial eta squared = 0.26). Both suggested that scores increased over time but that higher achievers consistently outscored their lower achieving counterparts.

Table 137: Comparison of mean inquiry scores by prior student achievement from the 2nd and 3rd inquiry tasks back to baseline data

	High Achievers			Low Achievers		
	n	Mean Score	SD	N	Mean Score	SD
Baseline	35	13.5	1.7	20	10.8	2.7
2nd Inquiry	35	14.0	1.9	20	12.2	3.0
3rd Inquiry	35	14.8	1.6	20	12.3	3.7

To look at the effect of experimental condition, the students were split up into four groups: the control, the experimental group that did not participate in reflection, and the experimental group that did participate in reflection divided by reflection type. Each of these groups was subsequently divided into high and low achieving groups and the group means for total inquiry scores were analyzed with a mixed between-within subjects ANOVA. There was no significant interaction between group membership and time, Wilks' Lambda = 0.68, F(14,92) = 1.40, p = 0.17, partial eta squared = 0.18. The main effect for time was significant, Wilks' Lambda = 0.85, F(2,46) = 4.02, p = 0.03, partial eta squared = 0.15. Most interesting, the main effect for group membership was substantial and significant, F(7,47) = 2.92, p = 0.01, partial eta squared = 0.30. All of the groups showed an increase in their total inquiry scores except for the low achieving students in the control group and the low achieving students who were asked to reflect while completing their inquiry tasks.

Table 138: Comparison of mean student inquiry scores by prior student achievement and experimental condition from the 2nd and 3rd inquiry tasks back to baseline data

		Baseline		2nd Inquiry		3rd Inquiry	
	n	Score	SD	Score	SD	Score	SD
High Achieving Control	5	13.2	1.3	13.6	1.5	13.8	1.1
Low Achieving Control	2	14.0	0.0	12.5	0.7	12.0	2.8
High Achieving Experimental – no reflection	7	13.7	1.7	14.1	2.9	14.9	1.7
Low Achieving Experimental – no reflection	8	10.4	2.6	13.0	3.4	13.6	2.1
High Achieving Experimental – with directed reflection	11	13.2	1.8	14.3	1.3	15.2	1.8
Low Achieving Experimental – with directed reflection	4	10.3	1.7	12.0	2.2	10.8	2.6
High Achieving Experimental – with generic reflection	12	13.9	2.0	13.9	2.1	14.8	1.5
Low Achieving Experimental – with generic reflection	6	10.7	3.3	11.2	3.4	11.5	5.9

A follow up mixed between-within subjects ANOVA looked specifically at the effect of prior achievement on Framing section scores over time. While Framing scores for high achievers generally remained flat, the Framing scores for low achievers tended to increase. The interaction between time and prior achievement was not significant, Wilks' Lambda = 0.96, $F(2,51) = 1.05$, $p = 0.36$, partial eta squared = 0.04. The main effect for time was not significant either, Wilks' Lambda = 0.94, $F(2,51) = 1.77$, $p = 0.18$, partial eta squared = 0.07, which suggested that Framing scores did not significantly change much over time. Group membership, on the other hand, was significant, $F(1,52) = 16.78$, $p < 0.0005$, partial eta squared = 0.24, which indicated that high achieving students consistently outscored their lower achieving counterparts.

Table 139: Comparison of mean Framing section scores by prior student achievement from the 2nd and 3rd inquiry tasks back to baseline data

	High Achievers			Low Achievers		
	n	Mean Score	SD	N	Mean Score	SD
Baseline	35	3.4	0.4	19	2.8	0.6
2nd Inquiry	35	3.5	0.5	19	3.0	0.7
3rd Inquiry	35	3.5	0.6	19	3.1	0.8

Table 140: Comparison of mean Framing section scores by prior student achievement and experimental condition from the 2nd and 3rd inquiry tasks back to baseline data

		Baseline		2nd Inquiry		3rd Inquiry	
	n	Score	SD	Score	SD	Score	SD
High Achieving Control	5	3.6	0.5	3.3	0.4	3.1	0.7
Low Achieving Control	2	3.3	0.4	3.0	0.7	2.8	0.4
High Achieving Experimental – no reflection	7	3.6	0.4	3.4	0.7	3.3	0.8
Low Achieving Experimental – no reflection	8	2.8	0.7	3.2	0.7	3.2	1.0
High Achieving Experimental – with directed reflection	11	3.3	0.3	3.5	0.4	3.5	0.7
Low Achieving Experimental – with directed reflection	4	2.8	0.6	2.9	0.6	2.6	0.5
High Achieving Experimental – with generic reflection	12	3.4	0.4	3.6	0.5	3.6	0.4
Low Achieving Experimental – with generic reflection	5	2.7	0.4	2.7	0.8	3.6	0.4

The data was again broken down by experimental group and reflection type with the Framing section data using a mixed between-within subjects ANOVA. The interaction of time and experimental group was not significant, Wilks' Lambda = 0.0.65, F(14,90) = 1.54, p = 0.11, partial eta squared = 0.19 and neither was the main effect for time, Wilks' Lambda = 1.00, F(2,45) = 0.12, p = 0.89, partial eta squared = 0.005. The main effect for group membership was substantial and significant, F(7,46) = 2.51, p = 0.03, partial eta squared = 0.27.

Word counts from the Framing sections over the course of this study were also compared by prior achievement using a mixed between-within subjects ANOVA. The interaction between time and prior achievement was not significant, Wilks' Lambda = 0.91, F(2,40) = 2.08, p = 0.14, partial eta squared = 0.09. As expected, the main effects for time (Wilks' Lambda = 0.72, F(2,40) = 7.67, p = 0.002, partial eta squared = 0.28) and prior achievement (F(1,41) = 6.89, p = 0.01, partial eta squared = 0.14) were both substantial and significant. These results suggested that even though the word counts increased over time for both groups (see table 141), students in the higher achieving group tended to write more words than their lower achieving counterparts.

Table 141: Comparison of mean Framing section number of words by prior student achievement from the 2nd and 3rd inquiry tasks back to baseline data

	High Achievers			Low Achievers		
	n	Mean # of Words	SD	N	Mean # of Words	SD
Baseline	29	151	55	14	130	40
2nd Inquiry	29	238	109	14	149	73
3rd Inquiry	29	253	124	14	189	101

The mean word counts from the Framing sections were also compared across experimental conditions by prior achievement. There was no significant interaction between time and experimental group assignment, Wilks' Lambda = 0.61, F(14,68) =

1.37, p = 0.19, partial eta squared = 0.22. The main effects for time (Wilks' Lambda = 0.81, F(2,34) = 4.03, p = 0.03, partial eta squared = 0.19) was significant, which indicated that scores did indeed change over time. The main effect of experimental condition, though, (F(7,35) = 1.22, p = 0.32, partial eta squared = 0.20) was not significant. These results suggested that even though word counts changed over time, these changes were not necessarily due to different experimental conditions. In addition, the small sample sizes that were compared made it difficult to come to any definitive conclusions.

Table 142: Comparison of mean Framing number of words by prior student achievement and experimental condition from the 2nd and 3rd inquiry tasks back to baseline data

		Baseline		2nd Inquiry		3rd Inquiry	
	n	Mean # of Words	SD	Mean # of Words	SD	Mean # of Words	SD
High Achieving Control	4	191	42	154	16	188	81
Low Achieving Control	1	163	0	242	0	177	0
High Achieving Experimental – no reflection	6	179	28	270	155	219	124
Low Achieving Experimental – no reflection	5	122	50	166	86	224	152
High Achieving Experimental – with directed reflection	10	149	71	238	67	276	117
Low Achieving Experimental – with directed reflection	3	145	32	112	29	126	60
High Achieving Experimental – with generic reflection	9	119	32	254	130	278	148
Low Achieving Experimental – with generic reflection	5	124	41	133	75	193	60

A similar set of analyses were conducted using the data from the Analyzing & Concluding sections. Students were divided into two groups based on prior achievement and their Analyzing & Concluding scores were then compared using a mixed between-within subjects ANOVA. The interaction between time and achievement was not significant (Wilks' Lambda = 0.99, F(2,51) = 0.14, p = 0.87, partial eta squared = 0.005). In contrast, both of the main effects for time (Wilks' Lambda = 0.84, F(2,51) = 4.97, p = 0.01, partial eta squared = 0.16) and prior achievement (F(1,52) = 11.33, p = 0.001, partial eta squared = 0.18) were both significant. These results indicated that Analyzing & Concluding scores tended to increase over time, but the higher achieving students consistently outscored the lower achieving students.

Table 143: Comparison of mean Analyzing & Concluding section scores by prior student achievement from the 2nd and 3rd inquiry tasks back to baseline data

	High Achievers			Low Achievers		
	n	Mean Score	SD	n	Mean Score	SD
Baseline	35	3.1	0.5	19	2.5	0.9
2nd Inquiry	35	3.2	0.8	19	2.6	1.2
3rd Inquiry	35	3.4	0.5	19	2.9	1.0

Analyzing & Concluding scores were also assessed by high and low achievers within experimental groups to see if there were specific conditions that favored high or low achieving students. The comparison was conducted using a mixed between-within subjects ANOVA. The interaction between time and experimental condition was not significant, Wilks' Lambda = 0.75, F(14,92) = 0.48, p = 0.48, partial eta squared = 0.13. The main effect for time was also not significant, Wilks' Lambda = 0.88, F(2,45) = 3.07, p = 0.06, partial eta squared = 0.12, which indicated that scores did not significantly change over time. The main effect for experimental condition, though, was significant, F(7,46) = 2.60, p = 0.02, partial eta squared = 0.28, which suggested that scores were influenced by achievement and experimental condition.

Table 144: Comparison of mean Analyzing & Concluding section scores by prior student achievement and experimental condition from the 2nd and 3rd inquiry tasks back to baseline data

		Baseline		2nd Inquiry		3rd Inquiry	
	n	Score	SD	Score	SD	Score	SD
High Achieving Control	5	2.8	0.9	3.2	0.3	3.3	0.3
Low Achieving Control	2	3.5	0.7	2.8	0.4	3.2	0.9
High Achieving Experimental – no reflection	7	3.0	0.3	3.2	1.1	3.3	0.4
Low Achieving Experimental – no reflection	8	2.6	0.7	3.0	1.2	3.1	0.8
High Achieving Experimental – with directed reflection	11	3.0	0.5	3.3	0.7	3.5	0.6
Low Achieving Experimental – with directed reflection	4	2.3	0.3	2.1	1.3	1.9	1.3
High Achieving Experimental – with generic reflection	12	3.3	0.5	3.1	0.8	3.4	0.6
Low Achieving Experimental – with generic reflection	5	2.3	1.4	2.2	1.3	3.2	0.8

Word counts from the Analyzing & Concluding sections were also compared by prior achievement using a mixed between-within subjects ANOVA. The interaction between time and achievement was not significant, Wilks' Lambda = 0.99, F(2,40) = 0.13, p = 0.88, partial eta squared = 0.006. Even though the number of words in the Analyzing & Concluding section tended to increase over time, this increase was not significant, Wilks' Lambda = 0.88, $F(2,40) = 2.68$, $p = 0.08$, partial eta squared = 0.12. The main effect for group membership, though, was significant, $F(1,41) = 6.10$, $p = 0.02$, partial eta squared = 0.13, which indicated that students with higher prior achievement tended to write more words than the students with lower prior achievement.

Table 145: Comparison of mean Analyzing & Concluding number of words by prior student achievement from the 2nd and 3rd inquiry tasks back to baseline data

	High Achievers			Low Achievers		
	n	Mean # of Words	SD	n	Mean # of Words	SD
Baseline	29	240	131	14	159	89
2nd Inquiry	29	260	110	14	183	128
3rd Inquiry	29	294	138	14	199	127

Finally, the mean number of words from the Analyzing & Concluding sections were compared by high and low achievers within each experimental group. This test was done with a mixed between-within subjects ANOVA. The interaction between time and group membership was not significant, Wilks' Lambda = 0.66, $F(14,68) = 1.13$, $p = 0.35$, partial eta squared = 0.19. The same was true of the main effects for time (Wilks' Lambda = 0.92, $F(2,34) = 1.41$, $p = 0.26$, partial eta squared = 0.08) for achievement within each experimental condition ($F(7,35) = 1.32$, $p = 0.27$, partial eta squared = 0.21). Again, with these results, it was difficult to draw any conclusions due to the small samples sizes of some of the groups.

Table 146: Comparison of mean Analyzing & Concluding section number of words by prior student achievement and experimental condition from the 2nd and 3rd inquiry tasks back to baseline data

		Baseline		2nd Inquiry		3rd Inquiry	
	n	Mean # of Words	SD	Mean # of Words	SD	Mean # of Words	SD
High Achieving Control	4	230	87	240	21	240	59
Low Achieving Control	1	125	0	157	0	170	0
High Achieving Experimental – no reflection	6	202	92	288	97	238	113
Low Achieving Experimental – no reflection	5	178	97	226	158	254	125
High Achieving Experimental – with directed reflection	10	237	99	275	117	370	149
Low Achieving Experimental – with directed reflection	3	131	79	89	46	91	79
High Achieving Experimental – with generic reflection	9	274	197	234	139	270	143
Low Achieving Experimental – with generic reflection	5	163	109	200	132	214	144

Final Assessment – Post-Study Survey

Chemistry Self-Efficacy Survey (CSES)

4th Administration CSES – Validation

This final phase took place approximately 34 weeks into the school year. Students, at this point, had completed and received feedback on the three science inquiry tasks that were a part of this study. In addition, there were approximately 2 weeks left in the school year when this survey was given, so the students had essentially completed one year's worth of chemistry content. Again, the purpose of this survey was to monitor student perceptions of their understandings of Chemistry Content and their abilities related to Scientific Inquiry. This was the fourth and final time that these students took the Chemistry Self-Efficacy Survey (CSES).

Table 147: Pattern and structure matrix for PCA with oblmin rotation of a two factor solution of CSES items from the 4th administration

Item	Pattern Coefficients		Structure Coefficients		Communalities
	Component 1	Component 2	Component 1	Component 2	
Question 1	.144	**.641**	.515	**.724**	.538
Question 2	.333	**.435**	.585	**.628**	.468
Question 3	**.734**	.134	**.811**	.558	.670
Question 4	**.507**	.209	**.628**	.502	.423
Question 5	**.932**	-.222	**.803**	.317	.678
Question 6	-.013	**.774**	.434	**.766**	.587
Question 7	**.801**	.039	**.824**	.503	.680
Question 8	.296	**.548**	.613	**.719**	.576
Question 9	-.299	**.897**	.220	**.724**	.584
Question 10	.289	**.582**	.626	**.750**	.618
Question 11	**.609**	.281	**.771**	.633	.647
Question 12	.141	**.711**	.552	**.792**	.641
Question 13	**.636**	.352	**.839**	.720	.787
Question 14	.259	**.612**	.613	**.762**	.625
Question 15	.089	**.723**	.507	**.775**	.605
Question 16	.090	**.762**	.530	**.813**	.667
Question 17	**.887**	.086	**.937**	.599	.883
Question 18	**.814**	.052	**.844**	.523	.714
Question 19	.129	**.748**	.562	**.823**	.688
Question 20	**.841**	.017	**.851**	.503	.724
Question 21	**.855**	.052	**.885**	.547	.786

Note: Major loadings for each item are in bold.
The Chemistry Content Scale contained questions: 1, 2, 6, 9, 10, 12, 14, 15, 16, 19
The Science Inquiry Scale is contained questions: 3, 4, 5, 7, 8, 11, 13, 17, 18, 20, 21

As with the prior three administrations, an analysis was conducted to assess the psychometrics of the instrument. The Chemistry Content Self-Efficacy scale yielded a Cronbach's α coefficient of 0.92 with a mean inter-item correlation of 0.54 (range: 0.20 to 0.74). The same analysis was done with the Science Inquiry Self-Efficacy scale. In this case the Science Inquiry Self-Efficacy scale yielded a Cronbach's α coefficient of 0.95 with a mean inter-item correlation of 0.61 (range: 0.30 to 0.87). These Cronbach's α coefficients were both higher than the previous administrations of this survey (3^{rd} administration Chemistry Content Self-Efficacy Cronbach's $\alpha = 0.90$; 3^{rd} administration Science Inquiry Self-Efficacy Cronbach's $\alpha = 0.92$).

The CSES was also evaluated in order to see if the initial two-factor structure was still valid. An inspection of the correlation matrix revealed that many coefficients were 0.3 and above. The Kaiser-Meyer-Oklin value was 0.84 which exceeded the recommended value of 0.6 (Kaiser, 1970 & 1974). Bartlett's Test of Sphericity (Bartlett, 1954) reached statistical significance ($p < 0.0005$) which provided support for the existence of factors in the correlation matrix. An analysis of the screeplot indicated that a two factor solution would explain a majority of the variance. To aid in the interpretation of these two components, an oblimin rotation was performed.

A two-component solution explained 64.7% of the variance. The items identified with the Science Inquiry Self-Efficacy scale explained 55.5% of the variance. In contrast, items identified with the Chemistry Content Self-Efficacy scale only explained 9.2% of the variance. Each successive iteration of the CSES has been able to explain a greater percentage of the variance (% variance explained by 1^{st}

administration: 41.8%, % variance explained by 2nd administration: 49.5%, % variance explained by the 3rd administration: 54.7%). There was a strong positive correlation between the two scales ($r = 0.58$). Although stronger than the first two administrations (1st administration: $r = 0.29$, 2nd administration: $r = 0.36$), this correlation was similar to the previous administration of the survey ($r = 0.61$).

Item loadings were then assessed to ensure that survey items loaded on the correct scale. For the Science Inquiry Scale, ten of the eleven items loaded best on this scale. The only exception was question #8. This question asked students how competent they felt about their ability to develop a scientific question. Student self-efficacy for Chemistry Content ($M = 5.0$, $SD = 1.2$) was slightly lower than student self-efficacy for Science Inquiry ($M = 5.4$, $SD = 1.2$). The implication was that students did not feel as efficacious about their abilities to construct a scientific question when compared to the other skills involved in Scientific Inquiry. This was similar to prior administration of this survey, so the decision was made to keep this question in the scale. On the Chemistry content scale, all ten items loaded as expected, so no changes were made.

4th Administration CSES – Student Score Analysis

To determine if experimental condition had any effect on student Chemistry Content Self-Efficacy scores, a one-way ANOVA was conducted on student scores by class period. Class means ranged from 4.6 to 5.7 on a 7 point Likert-type scales where a low score indicated low self-efficacy and a high score indicated high self-efficacy.

There was no significant difference between the classes for their scores on the Chemistry Content Self-Efficacy scale, $F(3,54) = 2.31$, $p = 0.09$. For the Scientific Inquiry scale, class means were slightly higher and ranged from 5.1 to 6.0. These differences were not significant, $F(3,54) = 1.56$, $p = 0.21$.

Table 148: Mean Chemistry Content Self-Efficacy and mean Science Inquiry Self-Efficacy from the 4th administration of the CSES

Period	n	Chemistry Content Self-Efficacy		Science Inquiry Self-Efficacy	
		Mean	SD	Mean	SD
2	10	4.9	1.5	5.1	1.5
4	16	4.6	1.1	5.3	1.0
5	14	5.7	0.9	6.0	1.1
6	18	5.0	1.2	5.3	1.2

Student self-efficacy scores in Chemistry Content were arranged by the order taught out in order to assist in the analysis of the survey. The goal was to determine how student efficacy specific to chemistry content changed over time. Similar to the 3rd administration of the survey, there was no real pattern to the data. At the time of the 4th administration of the survey, all of the content on the survey had been covered. Again, these results were not surprising. It was anticipated that students would feel more efficacious about some content and less efficacious about other content. One thing that is of interest is that some topics such as periodicity in questions #12 and #10 seemed to remain fairly stable whereas student efficacy on other content such as #6 and #9 on moles and bonding seemed to fluctuate over time.

Table 149: Summary results from the 4th administration of the CSES of mean Chemistry Content Self-Efficacy by order taught and by class period

Period	Chemistry Self-Efficacy Survey – Chemistry Content Self Efficacy Questions									
	#6	#12	#10	#9	#1	#15	#19	#14	#2	#16
2	4.6	5.5	5.3	3.6	5.6	4.2	4.1	5.6	4.3	5.2
4	3.9	5.2	5.1	3.9	5.0	4.3	4.6	4.8	4.8	4.5
5	4.9	6.4	6.6	4.9	6.1	5.2	5.9	6.1	5.4	5.4
6	3.8	5.7	5.4	4.2	5.3	4.6	5.0	5.4	5.1	4.9

Student efficacy scores for the four phases of science were also disaggregated in order to determine if there were differences in efficacy by phase and class period. Four ANOVA's were conducted to assess these differences. The differences in student efficacy in the first three phases were not significant (Framing, $F(3,54) = 1.45$, $p = 0.24$; Designing, $F(3,54) = 2.00$, $p = 0.13$; Collecting, $F(3,54) = 0.62$, $p = 0.60$). The difference in student self-efficacy scores in the Analyzing & Concluding phase, though, did turn out to be significant, $F(3,54) = 3.17$, $p = 0.03$. The effect size, calculated using eta squared, was quite large at 0.15. Post-hoc comparisons were conducted using the Tukey HSD test. The tests indicated that the only difference that was significant was between period 2 ($M = 4.7$, $SD = 1.8$) and period 5 ($M = 6.2$, $SD = 1.0$), $p = 0.03$.

Table 150: Summary results from the 4th administration of the CSES of mean Science Inquiry Self-Efficacy by inquiry category and period

Period	Framing Self-Efficacy		Designing Self-Efficacy		Collecting and Presenting Self-Efficacy		Analyzing & Concluding Self-Efficacy	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2	5.1	1.7	4.8	2.2	5.8	1.1	4.7	1.8
4	5.3	1.2	5.3	1.3	5.6	1.3	5.1	1.0
5	6.0	1.0	6.2	1.2	6.0	1.0	6.2	1.0
6	4.9	1.7	5.5	1.2	5.5	1.1	5.3	1.2

The four phases were also compared to each other to see if there were differences in how efficacious students felt about each phase. Phases were compared using a paired-samples t-test. Only two of the contrasts turned out to be significant. Students felt less efficacious about their abilities in Framing (M = 5.3, SD = 1.5) and Analyzing & Concluding (M = 5.4, SD = 1.3) than they did about their abilities related to Collecting (M = 5.7, SD = 1.1), $t(57) = 3.27$, $p = 0.002$ (two-tailed) and $t(57) = 3.10$, $p = 0.003$ (two-tailed) respectively. These results were similar to prior administrations of the survey.

Table 151: Summary results from the 4th administration of the CSES of mean Science Inquiry Self-Efficacy scores by inquiry category

Framing Self-Efficacy		Designing Self-Efficacy		Collecting and Presenting Self-Efficacy		Analyzing & Concluding Self-Efficacy	
Mean	SD	Mean	SD	Mean	SD	Mean	SD
5.3	1.5	5.5	1.5	5.7	1.1	5.4	1.3

4th Administration CSES – Effects of Gender

To assess the effect of gender on both Chemistry Content self-efficacy and Science Inquiry self-efficacy, results were analyzed with an independent-samples t-test. In Chemistry Content self-efficacy, males (M = 5.3, SD = 1.2) tended to have greater self-efficacy than females (M = 4.9, SD = 1.2). Although the gap was much smaller for Science Inquiry self-efficacy, males (M = 5.5, SD = 1.0) again tended to have greater self-efficacy than females (M = 5.4, SD = 1.4). Neither of these differences, though, was significant (Chemistry Content self-efficacy, $t(56) = 1.28$, $p = 0.20$ (two-tailed); Science Inquiry self-efficacy, $t(56) = 0.16$, $p = 0.88$ (two-tailed)).

Table 152: Mean student scores in Chemistry Content Self-Efficacy and Science Inquiry Self-Efficacy by gender from the 4th administration of the CSES

	n	Chemistry Content Self-Efficacy		Science Inquiry Self-Efficacy	
		Mean	SD	Mean	SD
Male	24	5.3	1.2	5.5	1.0
Female	34	4.9	1.2	5.4	1.4

4th Administration CSES – Effects of Achievement

To determine the effect of prior achievement on efficacy scores, another independent-samples t-test was used. On the measure of Chemistry Content self-efficacy, high achieving students (M = 5.5, SD = 1.1) tended to have greater self-efficacy than their lower achieving counterparts (M = 4.5, SD = 1.2). This difference was significant ($t(56) = 3.17$, $p = 0.002$ (two-tailed) and the magnitude of the difference in means (mean difference = 0.95, 95% CI: 0.35 to 1.55) was large (eta squared = 0.15). Although the difference in Science Inquiry self-efficacy was smaller for high achievers (M = 5.8, SD = 1.1) in comparison to lower achievers (M = 5.0, SD = 1.2), this difference was still significant, t(56) = 2.46, p = 0.02 (two-tailed). The magnitude of the differences in these means (mean difference = 0.76, 95% CI 0.14 to 1.39) was only moderate (eta squared = 0.10).

Table 153: Mean scores in Chemistry Content Self-Efficacy and Science Inquiry Self-Efficacy by prior student achievement from the 4th administration of the CSES

	n	Chemistry Content Self-Efficacy		Science Inquiry Self-Efficacy	
		Mean	SD	Mean	SD
High Achiever	34	5.5	1.1	5.8	1.1
Low Achiever	24	4.5	1.2	5.0	1.2

4th Administration CSES – Comparison to Prior Administrations: Scale Scores

In order to compare scores back to previous administrations of the survey, a one-way repeated measures ANOVA was used. Since there was no significant difference between the mean class scores on the Chemistry Content Self-Efficacy scale, all of the periods were grouped together as one large group for the purpose of comparison. As can be seen in table 154, student self-efficacy in Chemistry Content increased with each administration of the survey. The effect for time was significant, Wilks' Lambda = 0.51, F(3,44) = 14.17, p < 0.0005, partial eta squared = 0.49. The mean scores from each time period were compared to each other using a Bonferroni adjustment. The results of this comparison are given in table 154. While the mean

Chemistry Content self-efficacy scores from the 2nd, 3rd, and 4th administration were significantly larger than the baseline administration, neither of these three were significantly different from each other. The implication was that while student self-efficacy in Chemistry Content did grow during the course of the year, student efficacy seemed to plateau about two-thirds of the way through the school year.

Table 154: Mean Chemistry Content Self-Efficacy scores across all CSES administrations

	N	Mean	SD
Baseline	47	3.8	1.1
2nd Administration	47	4.7	1.0
3rd Administration	47	4.9	1.3
4th Administration	47	5.0	1.3

Table 155: Mean differences in Chemistry Content Self-Efficacy scores across all CSES administrations

	1	2	3	4
1. Baseline				
2. 2nd Administration	0.9*			
3. 3rd Administration	1.1*	0.2		
4. 4th Administration	1.2*	0.3	0.1	

* mean difference is significant at the $p < 0.0005$ level

The same analysis was conducted with the mean scores from the Science Inquiry Self-Efficacy scale. Again, all of the class periods were grouped together for the purposes of comparison. The general trend was that self-efficacy generally dropped over time, but then rebounded at the end of the year. It is also worth noting that the student self-efficacy in Science Inquiry was generally higher than their self-efficacy in Chemistry Content. The effect for time was significant, Wilks' Lambda = 0.84, $F(3,44) = 2.78$, $p = 0.05$, partial eta squared = 0.16. Even though the main effect for time was significant, a following pairwise comparison using a Bonferroni adjustment did not reveal any significant differences between individual survey administrations.

Table 156: Mean Science Inquiry Self-Efficacy scores across all CSES administrations

	N	Mean	SD
Baseline	47	5.5	0.87
2nd Administration	47	5.3	0.92
3rd Administration	47	5.2	1.16
4th Administration	47	5.4	1.25

Table 157: Mean differences in Science Inquiry Self-Efficacy scores across all CSES administrations

	1	2	3	4
1. Baseline				
2. 2nd Administration	- 0.2			
3. 3rd Administration	- 0.3	- 0.1		
4. 4th Administration	- 0.1	0.2	0.2	

* mean difference is significant at the p < 0.05 level

4th Administration CSES – Comparison to Prior Administrations: Gender

A mixed between-within subjects ANOVA was used to compare gender in relation to Chemistry Content self-efficacy across all four administrations. The was no significant interaction between time and gender, Wilks' Lambda = 0.96, F(3,43) =

0.55, $p = 0.65$, partial eta squared = 0.04. The main effect for time, though, was significant, Wilks' Lambda = 0.50, $F(3,43) = 14.20$, $p < 0.0005$, partial eta squared = 0.50. This implied that Chemistry Content self-efficacy scores showed a large increase over time. The main effect for gender was not significant, $F(1,45) = 0.16$, $p = 0.70$, partial eta squared = 0.003, which indicated that there was no real difference between male and female students in terms of Chemistry Content self-efficacy. A post-hoc analysis was done using a Bonferroni adjustment and found that for male students, the gain in efficacy from the initial administration to the remaining three administrations was significant (p values ranged from 0.0005 to 0.009). There was no difference in mean male Chemistry Content self-efficacy between the three final administrations of the survey. The same was true for the mean female Chemistry Content self-efficacy scores. Mean scores from the initial administration were significantly lower the three subsequent administrations (p values ranged from 0.002 to 0.006). Again, there was no difference between female mean Chemistry Content self-efficacy scores for the final three administrations.

The same analysis was conducted with the Science Inquiry self-efficacy data. The trend in this case was a general drop in Science Inquiry self-efficacy followed by a rebound with the final survey. The results from the mixed between-within subjects ANOVA revealed that the interaction between time and gender was not significant, Wilks' Lambda = 0.94, $F(3,43) = 0.98$, $p = 0.41$, partial eta squared = 0.06. In the same way, neither the main effect for time (Wilks' Lambda = 0.86, $F(3,43) = 2.34$, $p =$

0.09, partial eta squared = 0.14) nor the main effect for gender (F(1,45) = 0.05, p = 0.83, partial eta squared = 0.001) were significant.

Table 158: Mean Chemistry Content Self-Efficacy scores across all CSES administrations by gender

	Male – Chemistry Content			Female - Chemistry Content		
	n	Mean	SD	N	Mean	SD
Baseline	19	3.8	1.0	28	3.9	1.3
2nd Administration	19	4.9	1.2	28	4.6	0.8
3rd Administration	19	4.9	1.5	28	4.9	1.1
4th Administration	19	5.1	1.2	28	4.9	1.3

Table 159: Mean Science Inquiry Self-Efficacy scores across all CSES administrations by gender

	Male – Science Inquiry			Female – Science Inquiry		
	n	Mean	SD	n	Mean	SD
Baseline	19	5.5	0.9	28	5.5	0.9
2nd Administration	19	5.5	0.9	28	5.2	0.9
3rd Administration	19	5.2	1.3	28	5.1	1.1
4th Administration	19	5.3	1.0	28	5.5	1.4

4th Administration CSES – Comparison to Prior Administrations: Achievement

To examine the effect of prior achievement on Chemistry Content self-efficacy scores, a mixed between-within subjects ANOVA was conducted. The interaction between time and prior achievement was not significant, Wilks' Lambda = 0.90, F(3,43) = 1.52, p = 0.22, partial eta squared = 0.10. The main effect for time, though, was substantial and significant, Wilks' Lambda = 0.53, $F(3,42) = 12.90$, $p < 0.0005$, partial eta squared = 0.47, which indicated that scores did indeed significantly increase over time. As expected, the main effect for prior achievement was also significant, $F(1,45) = 7.92$, $p = 0.01$, partial eta squared = 0.15, which indicated that high achieving students generally had higher Chemistry Content self-efficacy when compared to their peers. Post-hoc comparisons were made using a Bonferroni adjustment. These comparisons indicated that the initial mean Chemistry Content self-efficacy scores for high achieving students were significantly lower than the Chemistry Content self-efficacy scores of the other three administrations (all p values were below 0.0005). For the high achieving students, though, there was no significant difference between the mean Chemistry Content self-efficacy scores for the 2nd, 3rd, and 4th administrations of the survey. In contrast, there were no significant differences in the mean Chemistry Content self-efficacy scores for the low achieving students between any of the survey administrations.

Table 160: Mean Chemistry Content Self-Efficacy scores across all CSES administrations by prior student achievement

	High Achievers Chemistry Content			Low Achievers Chemistry Content		
	n	Mean	SD	N	Mean	SD
Baseline	28	4.0	1.0	19	3.7	1.4
2nd Administration	28	5.0	0.7	19	4.3	1.2
3rd Administration	28	5.3	1.1	19	4.2	1.3
4th Administration	28	5.4	1.1	19	4.5	1.4

Table 161: Mean Chemistry Content Self-Efficacy scores across all CSES administrations by prior student achievement and experimental condition

		Baseline		2nd Administration		3rd Administration		4th Administration	
	n	Mean	SD	Mean	SD	Mean	SD	Mean	SD
High Achieving Control	4	4.1	1.2	4.8	1.0	5.3	1.1	5.2	1.5
Low Achieving Control	3	4.7	1.8	5.0	1.0	4.0	1.7	4.7	2.6
High Achieving Experimental – no reflection	5	3.2	0.6	4.7	0.6	5.0	0.5	4.9	0.8
Low Achieving Experimental – no reflection	6	2.8	0.7	4.0	1.0	4.3	1.1	4.0	1.3
High Achieving Experimental – with directed reflection	9	4.0	1.0	5.2	0.7	5.6	1.0	6.0	0.9
Low Achieving Experimental – with directed reflection	4	4.9	1.3	4.7	1.8	4.7	0.5	5.0	0.8
High Achieving Experimental – with generic reflection	10	4.2	0.9	5.1	0.7	5.2	1.4	5.2	1.0
Low Achieving Experimental – with generic reflection	6	3.3	1.1	3.9	0.9	3.9	1.7	4.4	1.2

Another mixed between-within subjects ANOVA was conducted with the Chemistry Content self-efficacy scores, but this time, the students were grouped both by prior achievement and experimental condition. Again there was no significant interaction between time and grouping, Wilks' Lambda = 0.54, $F(21,107) = 1.20$, $p = 0.26$, partial eta squared = 0.18. The main effect for time was significant, Wilks' Lambda = 0.54, $F(3,37) = 10.40$, $p < 0.0005$, partial eta squared = 0.46, which again indicated that scores generally increased over time. The main effect for group membership by experimental condition, though, only approached significance, F(7,39)

= 2.17, p = 0.06, partial eta squared = 0.28. High achieving students tended to have higher Chemistry Content self-efficacy scores regardless of experimental condition.

A similar mixed between-within subjects ANOVA was conducted using the Science Inquiry self-efficacy data. The interaction between time and achievement was not significant, Wilks' Lambda = 0.96, $F(3,43) = 0.57$, $p = 0.64$, partial eta squared = 0.04. Like the content data, the main effect for time was also significant, Wilks' Lambda = 0.84, $F(3,43) = 2.84$, $p = 0.05$, partial eta squared = 0.17, which indicated that Science Inquiry self-efficacy scores did indeed change over time. Even though the Science Inquiry self-efficacy score patterns were similar for both high and low achievers, the main effect for prior achievement was significant, $F(1,45) = 5.11$, $p = 0.03$, partial eta squared = 0.10, which indicated that high achievers also tended to have higher Science Inquiry self-efficacy.

Table 162: Mean Science Inquiry Self-Efficacy scores across all CSES administrations by prior student achievement

	High Achievers - Science Inquiry			Low Achievers - Science Inquiry		
	n	Mean	SD	n	Mean	SD
Baseline	28	5.6	0.9	19	5.2	0.8
2nd Administration	28	5.5	0.8	19	5.0	1.1
3rd Administration	28	5.4	1.9	19	4.7	1.2
4th Administration	28	5.7	1.1	19	5.0	1.3

The student scores were then grouped by achievement and experimental condition. Science Inquiry self-efficacy scores were then compared using a mixed between-within subjects ANOVA. There was no significant interaction between time and group membership, Wilks' Lambda = 0.58, $F(21,107) = 1.03$, $p = 0.43$, partial eta squared = 0.16. Like the other comparisons, the main effect for time was significant, Wilks' Lambda = 0.81, $F(3,37) = 2.86$, $p = 0.05$, partial eta squared = 0.19, which again indicated that the changes in Science Inquiry Self-Efficacy were significant. The main effect for group membership, though, was not significant, $F(7,39) = 1.41$, $p = 0.23$, partial eta squared = 0.20, which indicated that there was not much difference in Science Inquiry Self-Efficacy scores between the groups.

Table 163: Mean Science Inquiry Self-Efficacy scores across all CSES administrations by prior student achievement and experimental condition

		Baseline		2nd Administration		3rd Administration		4th Administration	
	n	Mean	SD	Mean	SD	Mean	SD	Mean	SD
High Achieving Control	4	5.7	0.9	4.9	1.2	5.0	1.2	4.7	1.5
Low Achieving Control	3	5.1	0.9	4.8	1.0	4.1	1.5	5.3	2.4
High Achieving Experimental – no reflection	5	5.7	0.9	5.5	0.3	5.7	0.6	6.0	0.4
Low Achieving Experimental – no reflection	6	5.0	0.8	4.4	1.1	4.5	1.0	4.6	1.1
High Achieving Experimental – with directed reflection	9	5.6	1.1	5.5	0.8	5.8	0.8	6.3	1.0
Low Achieving Experimental – with directed reflection	4	5.6	0.8	5.8	0.8	5.2	1.4	5.4	1.3
High Achieving Experimental – with generic reflection	10	5.5	0.8	5.7	0.7	5.2	1.4	5.4	1.1
Low Achieving Experimental – with generic reflection	6	5.3	0.9	5.3	0.9	4.9	1.2	5.0	1.2