DISSERTATION

COOPERATIVE LEARNING INSTRUCTION AND SCIENCE ACHIEVEMENT FOR SECONDARY AND EARLY POST-SECONDARY STUDENTS: A SYSTEMATIC REVIEW

Submitted by

Christopher C. Romero

School of Education

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2009

COLORADO STATE UNIVERSITY

February 24, 2009

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY CHRISTOPHER C. ROMERO ENTITLED COOPERATIVE LEARNING INSTRUCTION ON SCIENCE ACHIEVEMENT FOR SECONDARY AND EARLY POST-SECONDARY STUDENTS: A SYSTEMATIC REVIEW BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Greg Florant, Ph.D.

Jean Lehmann, Ph.D.

Advisor, R. Brian Cobb, Ph.D.

Co-Advisor, Paul Kennedy, Ph.D.

Director, School of Education, Tim Davies, Ph.D.

ABSTRACT OF DISSERTATION

COOPERATIVE LEARNING INSTRUCTION AND SCIENCE ACHIEVEMENT FOR SECONDARY AND EARLY POST-SECONDARY STUDENTS: A SYSTEMATIC REVIEW

A systematic review of 2,506 published and unpublished citations identified in a literature search on science outcomes associated with cooperative learning in secondary and early post-secondary science classrooms between 1995 and 2007 was conducted. The goal of this review was to determine what impact cooperative learning had on science achievement of students compared to traditional instruction.

A tri-level screening and coding process was implemented and identified 30 original, empirical studies that met the inclusionary criteria while yielding an overall effect size estimate. The minimum methodological criteria for inclusion were as follows: (a) the study utilized a treatment/control design, (b) cooperative learning was the intervention, and the control group experienced traditional instruction, (c) the subjects in included studies were secondary or early-post-secondary students, (d) the study was performed in a science classroom, and (e) student achievement was the outcome measure. This meta-analysis describes the main effect of cooperative learning; additionally, a variety of moderator analyses were conducted in order to determine if particular study and participant characteristics influenced the effect of the intervention.

The results of this review indicate that cooperative learning improves student achievement in science. The overall mean effect size was .308, a medium effect (Cohen, 1988). Moderator analyses on study participant characteristics gender and ability level were inconclusive based on the small number of studies in which data on these characteristics were disaggregated. If the intervention was structured in a particular fashion, the effect on student achievement was greater than that for an unstructured intervention. The intervention showed a greater effect on student achievement in biology classes than in other science disciplines. Studies performed using cluster randomized or quasi-experimental without subject matching methodologies showed a greater effect on student achievement in science than studies that used the quasi-experimental with subject matching methodology. Implications for teacher education policy and recommendations for improvements in methodological practices and reporting are given.

Christopher Charles Romero
School of Education
Colorado State University
Fort Collins, CO 80523
Spring 2009

ACKNOWLEDGEMENTS

A large number of people have assisted me throughout my doctoral experience and without whom I would not have been able to complete this lifelong goal of mine. I would like to thank Dr. Brian Cobb and Dr. Paul Kennedy for their assistance and support as my advisor and co-advisor throughout this educational experience. Dr. Cobb has provided me with candid and constructive input regarding my analytical and writing skills that have allowed me to develop more fully as a scholar and educator. Dr. Kennedy has provided me with the financial support and continual academic support which have both allowed me to complete my degree. I would like to thank Dr. Jean Lehmann and Dr. Greg Florant for their participation on my committee and their work in providing me with input that also allowed me to finish my degree.

I would like to thank Barbara Patterson, my colleague and friend, who was instrumental in assisting me in the data analysis phase of my dissertation. Without her assistance, I would not have been able to complete the research. I would like to thank the other CLTW fellows for their emotional and intellectual support throughout this experience. We have had a lot of great experiences in the past 5 years, and I will always remember them.

I would like to thank my colleagues at Front Range Community College, Larimer Campus, Jim Butzek and Shashi Unnithan, for allowing me the flexibility in my job to get this degree completed without disrupting my employment or my family life.

There are three people that I need to thank, and without whom, I would never have gotten into the graduate program at CSU. Dr. Christine Jones and Dr. Ed Geary were the original

principal investigators of the CLTW grant. I had known them through other projects and in the spring of 2004, they mentioned this program to me and encouraged me to apply. Without their encouragement, I would never have considered going back to school, so I am indebted to them. In addition, Dr. Rick Ginsberg was the Director of the School of Education at that time and was my first advisor. He was very supportive of me and helped me get started in the program.

Last but certainly not least, I would like to thank my wife Claudia and our children for their support and understanding through this long and often stressful process. They have stood by me and endured my long hours at the computer while I worked to finish my dissertation. I would also like to thank my parents, Art and Ginny, who have always supported my educational endeavors and who brought me into this world.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

"Let us put our minds together...and see what life we can make for our children."

--Sitting Bull

A Brief History of Cooperative Learning

Cooperative learning (CL) is one of the most recognized and distinguished types of instructional practice and intervention, and has a rich history spanning many centuries. Some of the earliest examples of cooperative learning or activity appear in the Talmud, a book of rabbinic Jewish laws dating as far back as 500 B.C. (Johnson & Johnson, 1999). This type of behavior is even thought to have evolved in various organisms (Pennisi, 2005). One of the earliest appearances of this intervention in American education was during the Common School Movement of the early 1800s. Educational theorists such as Dewey and Lewin began discussing cooperation in their writings in the early 1900s. Research on cooperation in education began to gain momentum and recognition in the 1960s. Cook (1969) examined the impact of cooperation on relationships between black and white college students; Madsen compared children's preferences for cooperative and competitive interaction across cultures and ages; Kagan, one of Madsen's students, performed a series of studies on cooperation and competition in children (Johnson & Johnson, 1999). The Johnson brothers began training teachers in the implementation of cooperative learning in the late 1960s and developed research reviews on cooperation and competition in the 1970s. A number of theorists and researchers continued this type of

work from the 1970s to the 1990s, including Devries and Edwards (1974), Aronson (1978), Slavin (1980), and Slavin, Leavey, and Madden (1982). To date, at least seven meta-analyses of the types and effects of cooperative learning have been completed (Bowen, 2000; Howard, 1996; Johnson, Johnson & Stanne, 2000; Kulik & Kulik, 1982; Qin, Johnson, & Johnson, 1995; Scott, Tolson, Schroeder, Lee, Huang, Hu, & Bentz, 2005; Springer, Stanne, & Donovan, 1999). The primary conclusions to be drawn from research on cooperative learning are that it is an effective instructional intervention that has a positive effect on student achievement.

Theoretical Roots of Cooperative Learning

Three general theoretical perspectives regarding cooperative learning have guided its practice and the research regarding its effectiveness—social interdependence, cognitive-developmental, and behavioral learning theories. Social interdependence theory is considered the most influential. In the 1920s and 1930s, Lewin (1935) theorized that the essence of a group is the interdependence among its members so that any change in membership changes the dynamics of the group as a whole. In addition, an internal state of tension within group members motivates the accomplishment of common goals. In the 1940s, Deutsch (1949) refined Lewin's notions, formulating a theory of cooperation and competition, indicating that interdependence can be either positive, through cooperation, or negative, through competition. Johnson and Johnson (1974) extended Deutsch's work into social interdependence theory, which states that the way social interdependence is structured will determine how individuals interact with each other. This in turn affects group outcomes. Cooperation, which is positive interdependence, results in promotive interaction during which individuals facilitate each

other's learning efforts. Competition, negative interdependence, usually results in oppositional interaction, during which individuals obstruct each other's learning efforts. Without interaction, individuals work independently without any exchange of ideas. This situation typically leads to decreased achievement and negative relationships, while positive interaction does just the opposite.

The cognitive-development perspective is based primarily on the theories of Vygotsky (1978) and Piaget (1950). Moreover, Johnson and Johnson (1979) discuss the theory of academic controversy and its relationship to this perspective. To Piaget, cooperation involves striving to achieve common goals while "coordinating one's own feelings and perspective with a consciousness of others' feelings and perspectives" (in Johnson & Johnson, 1999; p. 187). When individuals cooperate, conflict often occurs which creates cognitive dissonance. This in turn encourages different perspective-taking and cognitive development, accelerating a student's intellectual development by forcing them to reach consensus with other students whose points of view differ regarding the educational task at hand. Vygotsky (1978) claimed that human mental functions and accomplishments have their origins in social relationships, and that knowledge is socially constructed through cooperative efforts to learn and solve problems. A central concept of his theory is the zone of proximal development, which is the zone between what an individual can achieve independently and what the student can achieve when working collaboratively with more knowledgeable peers or under the guidance of an expert, such as a teacher. If students do not work cooperatively, Vygotsky states that students will not grow intellectually and that time spent on independent work should be minimized.

Behavioral learning theory states that students will work hard on tasks from which they will secure some sort of reward and will fail on those tasks that provide no reward or yield punishment (Bandura, 1977; Skinner, 1968). Cooperative learning is designed to offer incentives to group members to achieve a group task since individuals are not intrinsically motivated to help their classmates toward a common goal in a non-cooperative activity.

The aforementioned theories provide a triangulation that validates cooperative learning from a theoretical perspective. All three predict that cooperative learning will promote higher achievement among students than competition and have inspired research on cooperation and cooperative learning, which will be addressed subsequently.

Essential Characteristics of Cooperative Learning

Cooperative learning (CL) is defined as "the instructional use of small groups so that students work together to maximize their own and each other's learning" (Johnson & Johnson, 1999, p. 5). According to Johnson and Johnson (1999), CL has five essential characteristics. The first is positive interdependence, where the success of the group depends on the success of each member of the group. When positive interdependence is clearly understood by a group, members understand that their individual efforts are required and necessary for group success, and that each member has a unique and valuable contribution to make to the cooperative effort because of her or his role, responsibilities, or resources. The second essential characteristic of CL is individual accountability/personal responsibility, which exists when the performance and effort of each individual in a group is assessed and brought back to the individual and the group and compared against some measure of standard performance. The member is held

responsible for contributing to the group's success. The third essential characteristic of CL is face-to-face promotive interaction, which occurs when group members facilitate and encourage each other's efforts and success so that the group may be successful. The fourth characteristic is interpersonal and small group skills. Students must be taught these social skills so that groups will function effectively. Higher achievement can be expected of groups if teachers devote adequate attention to training their students in these skills. The fifth essential characteristic of CL is group processing. Group members must reflect on a group session in order to determine what member activities were helpful and unhelpful and then make decisions about what activities remain or need to be changed. Kagan (1994) identifies four basic principles of CL. The first is simultaneous interaction, where all members of a group are involved and engaged in the discussion taking place. This is similar to the third essential characteristic as identified by Johnson and Johnson (1999). The second principle is positive interdependence, which is essentially the same as is stated by Johnson and Johnson (1999). The third principle is individual accountability, which is also the same as stated by Johnson and Johnson (1999). The fourth is equal participation, which is also similar to the third essential characteristic as identified by Johnson and Johnson (1999).

Study Context

Research on Cooperative Learning/Meta-Analysis

During the past 25 years, at least eight meta-analyses addressing the effects of cooperative learning on student achievement have been reported. Johnson, Maruyama, Johnson, Nelson, and Skon (1981) performed a meta-analysis comparing the effectiveness of cooperation, cooperation with intergroup competition, interpersonal

competition, and individualistic goal structures in promoting achievement among North American students from elementary through college. In all cases, based on 122 identified studies, cooperative work regardless of competition increased student achievement. This study examined moderator variables including sex, ethnicity, and subject area, but based on the information reported in the individual studies, no conclusions regarding the impact of cooperation based on these variables could be made. This particular study was methodologically sound and reported a positive impact of cooperation; however, it was done in 1981, and many studies have been done since. In addition, it did not disaggregate data based on student characteristics or class subject. Additionally, literature search keywords were not reported, but study inclusion criteria were stated. Kulik and Kulik (1982) performed a meta-analysis examining the effects of ability-grouping in secondary schools on four student outcomes, one of which was achievement. Based on 52 identified studies, they reported a small but significant effect on student achievement in favor of grouping when comparing classes that were grouped with those that were not. Grouped classes showed an increase in achievement on examination scores from the 50^{th} percentile to the 54^{th} percentile, an increase of one-tenth standard deviations. While this study reported positive effects of grouping, some concerns warrant mentioning. First, the study is from 1982, and others have been performed since. Methodologically, search terms for the literature search were not listed, nor were the criteria for the selection of included studies, so duplication of the study would be difficult. Moreover, only three studies were performed in a science discipline, calling into question results related to science achievement, and the demographics of the subjects were not disaggregated (except for ability level). Qin, Johnson, and Johnson (1995) compared the impacts of cooperation

and competition on problem solving. They identified four different types of problem solving and did meta-analyses of each one, examining student achievement on these tasks. The age of the subjects of the studies ranged from preschool through college. From the 46 studies that were identified, effect sizes were positive, in favor of cooperation over competition, regardless of study quality, problem-solving type, and age of subjects in the study. Methodologically, this study was of good quality; however, search terms and inclusion criteria were not given. In addition, the disciplines in which the CL took place were not delineated. Howard (1996) performed a meta-analysis of the effects of "scripted" cooperative learning (in which student groups are given explicit instructions for a task) on student achievement in comparison to individualized or traditional learning activities. Inclusion criteria were reported, but no rubric for rating studies was included. Additionally, no search terms for the literature search were listed, and no descriptions of the subjects from the studies were given. Positive effect sizes were reported based on an analysis of 13 studies, but concluding that the CL model increased student achievement based on a sample size this small is suspect. Johnson, Johnson, and Stanne (2000) published a meta-analysis in which they identified ten particular types of cooperative learning from 158 studies and analyzed their effect on student achievement as measured by a variety of assessments. When compared to competitive learning, effect sizes of the ten types of CL ranged from .18 to .85, exhibiting positive effects (Cohen, 1988). When compared to individualistic learning, effect sizes for the ten types of CL, effect sizes ranged from .13 to 1.04, again exhibiting positive effects (Cohen, 1988). This particular study employed highly reliable search methods and inclusion criteria, as well as widely accepted effect size calculation methodology.

However, this particular study did not identify the scholastic disciplines in which the CL took place. Springer, Stanne, and Donovan (1999) conducted a meta-analysis on the effects of small-group learning (cooperative learning) on undergraduates in science, math, engineering, and technology (SMET) classes focusing on literature dating 1980-1996. In this review, they disaggregated data related to achievement in each of the disciplines and calculated a mean effect size (*ES*) of .42 for science based on nine identified studies, indicating that small-group learning has a positive impact on science achievement. Two methodological shortcomings of this review were apparent upon examination of the report in detail. First, the researchers did not list the keywords or search terms they used in the literature searching. And second, they did not include detailed criteria or a protocol that explained inclusionary or exclusionary criteria that set the boundaries establishing the studies that were included in the review. Similarly, Bowen (2000) conducted a meta-analysis on the effects of cooperative learning on the achievement of college and high school students in chemistry classes. He identified 15 studies and from these studies, 30 effect sizes were calculated with a mean effect size of .37. Bowen's study supports the notion that cooperative learning has a positive effect on chemistry achievement. This study also exhibited methodological shortcomings. First, no keywords or search terms were listed. Second, only four specific peer-reviewed journals were searched; there was no mention of electronic database searching. And finally, there were no detailed criteria or a protocol that explained inclusionary or exclusionary criteria that set the boundaries establishing the studies that were included in the review. These two reviews, then, constitute the most recent meta-analytic evidence of research conducted solely on cooperative learning and its effect on science achievement.

Scott, Tolson, Schroeder, Lee, Huang, Hu, and Bentz, (2005) published a meta-analysis with the intent of describing and analyzing teaching methodologies utilized in secondary science classrooms that have been shown to improve student achievement in science. In this report, they analyzed 61 different studies and identified eight teaching strategies from these studies. For cooperative learning, they calculated an average effect size of .958 based on three studies that were among the 61 in their review. Consistent with the aforementioned studies, the Scott, et al. (2005) study also exhibited significant methodological shortcomings. First, in their list of literature search keywords and terms, cooperative learning and small-group learning were not listed. This oversight alone would suggest the mean effect size from the three reported studies is probably highly unreliable. No other meta-analyses addressing the impact of cooperative learning on science achievement have been published since Bowen (2000), and his only focused on chemistry achievement. Therefore, there have been no meta-analyses on the impact of cooperative learning on science achievement (including all science disciplines) since the Springer, et al. (1999) study. A significant opportunity to engage in a new and more thorough meta-analysis on the impact of cooperative learning on science achievement now exists.

Research on Cooperative Learning/Original Studies

Chang and Mao (1999b) performed a study assessing the effects of cooperative learning in comparison to traditional instruction on ninth grade students' achievement in earth science. In this quasi-experimental study involving 8 teachers, 20 classes of students were divided into a control group and an experimental group (10 classes in each). Prior to participating in the study, the teachers attended a 15-hour workshop on

cooperative learning. Additionally, the students practiced cooperative learning strategies on topics other than those used in the study. The control group experienced traditional, lecture/discussion instruction, while the experimental group experienced a modified group investigation method with three types of learning activities: small group discussion designed to clarify concepts, group projects based on hands-on activities and group discussion (with individual and group accountability), and individual presentations to the entire class addressing group projects and discussion. The measurement of achievement was a test using items selected from two national examinations, which was given as both a pre-test and post-test to both groups of students. A team of experts assessed the correspondence of the items to textbook contents. In addition, the Kuder-Richardson reliability coefficient was calculated at .777. Consequently, the measurement tool can be deemed reliable. The team of experts also classified the test items into three categories based on Bloom's taxonomy as knowledge items, comprehension items, and application items. ANCOVA tests performed on the post-test scores revealed no significant difference in overall achievement between treatment and control groups. Moreover, there were no differences in achievement between treatment and control groups on knowledge or comprehension items. By contrast, the experiment group showed a significant difference in achievement scores on the application items in comparison to the control group ($F(1,17) = 4.63, p < .05$). Based on the findings of this study, cooperative learning increases student achievement on application of concepts, but not on knowledge or comprehension of concepts. This study was methodologically sound with regard to its overall design and the instrument used to assess student achievement. The pre-study training of the teachers made an effort to ensure that classroom instruction would be

consistent for the duration of the experiment. The only criticisms that could be made is that the experimental and control groups were not necessarily equivalent based on student characteristics and that multiple teachers were involved in the study. Controlling for differences in teacher characteristics was not done, with the exception that all teachers attended the same training on cooperative learning.

Shachar and Fischer (2004) performed a study comparing the effects of the Group Investigation (GI) method of cooperative learning on student achievement in 11th grade chemistry in comparison to students who experienced traditional instruction. In the GI method, groups of students develop research projects based on prior exposure to course concepts. Then, the groups plan their work, carry out their investigation, compile findings, and then report their findings to their classmates and teacher. The teacher and the other students then assess the group's work. During the first year of the study, the teachers involved participated in workshops devoted to understanding cooperative learning and the GI method. In addition, all of the teachers conducted trial lessons under the observation of the researchers to determine if the teachers could in fact teach students using these interventions. In the second year of the study, teachers were assigned to the experimental or control groups. Pre-test and post-test assessment instruments were used to assess student achievement. The post-test instrument dealt specifically with the content taught during the experiment, while the pre-test dealt with content that had been learned previously by the students. The teachers in the study developed the assessment instruments using questions from an item bank from a national administered examination. The instruments were then reviewed by the teachers and an expert. Degree of agreement (Miles & Huberman, 1994) between the teachers and the expert was calculated at 88.7%,

so the instruments may be deemed reliable. Scores on the pre-test revealed a broad range of scores, so it was decided that the effects of the intervention would be assessed using student achievement levels, rather than overall achievement of all students. A multivariate analysis of variance of the 2 teaching methods, 3 levels of student achievement, and pre-test and post-test scores was performed. The results of this analysis revealed a main effect for teaching method; students in experimental classes outperformed students in the control group. Moreover, there was a main effect for students' achievement level, stating that students at different levels of achievement were affected differentially by the two teaching methods. Also, a two-way interaction was noted between students' achievement level and method of instruction such that the GI method affected students from different achievement levels differentially. Low achieving students in both experimental and control groups improved their scores, with those in the experimental group showing the highest improvement compared to students of different achievement levels. The scores of middle achievers in the control group declined somewhat while the scores in the experimental group improved significantly. The scores of high achieving students in both the control and experimental groups declined, but this was more pronounced in the control group. Post hoc analyses of the main effect of method were conducted as well, with teaching method and time (pre-test and post-test) as the factors. This analysis showed that low-level achieving students in GI classes significantly outperformed their peers who experienced traditional instruction. These same results were seen in middle-level achieving students. Lastly, all high-achieving students experienced a significant decline in scores in both control and experimental groups. The findings of this study are multifaceted. First, the GI method of instruction

has a significant positive effect on the achievement of low and middle level achieving students. These results support the idea that cooperative learning can improve student achievement. Despite the finding that the performance of high-achieving students declined, cooperative learning can bring up the achievement of lower performing students and is an effective intervention. This study was very thorough methodologically. All the teachers were trained in the GI method prior to the study. The instruments used in the study were reliable, and the statistical analysis of the data was thorough and appropriate. The only criticism that could be made would be that controlling of student characteristics was not completely taken into account; however, after the pre-test was scored, further analysis based on student achievement level was done, thus partially mitigating this issue.

Bilgin and Geban (2006) conducted a study in which they examined the effects of collaborative learning based on conceptual change conditions on student achievement in comparison to traditional instruction in 10^{th} grade chemistry students. The conceptual change condition model allows students to address their own prior knowledge and potential misconceptions regarding a concept and work to modify or change them if necessary. In this study, whole class direct instruction of course concepts was conducted, and then groups of students were given specific tasks to complete that had both individual and group rewards for completing the tasks. The subjects included 2 classes of 10^{th} grade chemistry students both taught by the same teacher. The control group experienced traditional instruction while the experimental group experienced cooperative learning based on conceptual change conditions. Both groups were given a standardized chemical concepts pre-test, which had a reliability coefficient of .82. In addition, both groups were given a standardized science process skills assessment as a pre-test to control its effect as

a covariate and to reveal any contribution to student conceptual understanding and achievement. This instrument had a reliability coefficient of .85. Finally, all students were given a standardized chemical achievement assessment as a post-test. This instrument had a reliability coefficient of .81. Based on reliability statistics, all instruments were deemed reliable indicators of student performance. Prior to the treatment, the teacher was trained in cooperative learning based on conceptual change conditions, and the researcher provided materials to the teacher for use in the classroom. In addition, the researcher observed the teacher in both classes in order to control for teacher behavior or bias. No teacher bias was observed. Students were organized into heterogeneous groups based on prior achievement levels with 1 high-achieving, 2 middle-achieving, and one low-achieving. Both groups were given both pre-tests prior to the treatment, with equal numbers of boys and girls. The control group was instructed based on the traditional approach where the teacher presented concepts, allowed time for whole-class discussions and then assigned worksheets for students to complete individually, which were scored by the teacher and returned to the students. The experimental group was instructed using cooperative learning based on conceptual change conditions. The instructor explained the objectives of each unit of content and then assigned problem sets for the teams to discuss and complete. Each student received discussion statements prior to group meetings and were asked to decide if the agreed or disagreed with the statements and record their explanations. In the groups, students were required to come to consensus regarding the discussion statements and whether they agreed or disagreed with the statements. Students then were required to develop analogies related to the concepts and then complete various teacher-constructed

computational assignments in their groups. Students took individual quizzes during the treatment and the top three groups in order of success on the quizzes were given a reward to encourage in-group discussion. At the end of the treatment time, both the control and experimental groups were given the post-tests. Analysis of the two pre-test scores indicated no significant difference in students' understanding of chemical concepts but did indicate a difference in science process skills between the control and experimental groups. Consequently, the science process skills score was used as a covariate in the subsequent analysis of post-test scores. A multivariate analysis of covariance (MANCOVA) revealed a significant main effect for the treatment, indicating a significant difference in achievement between the treatment and control groups, with the treatment group exhibiting higher achievement. Subsequent analysis of covariance on both post-test scores indicated statistically significant differences between students understanding of chemical concepts and achievement related to chemical computation problems, with the treatment group exhibiting higher scores on both assessments. This study had few methodological flaws, and utilized appropriate statistics based on what was being analyzed. The only issue of note is that no mention of population differences was addressed between the control and experimental groups. The science process skills pre-test was utilized to assess at least one variable between the groups, and it was found that this particular skill did have an effect on student understanding of chemical concepts and achievement.

Bilgin (2006) conducted a study that compared achievement in chemistry between first-year undergraduate students who all used Polya's problem solving technique, with the experimental group using pair problem solving and a control group who experienced

traditional instruction. Methods used in the study were similar to those mentioned above. Results of the study showed a statistically significant increase in student achievement in the experimental group when compared to the control group. Since the findings of this study are similar to those of the aforementioned studies, details will not be presented.

All of the studies mentioned above support the notion that cooperative learning techniques improve student achievement. However, this is not always the case. Hanze and Berger (2007) conducted a quasi-experimental study in which they investigated the affect of the "jigsaw" method of cooperative learning on student achievement in 12^{th} grade physics classes. In the "jigsaw" method used in this study, students were assigned to "expert" groups in which they were assigned specific course concepts that they were asked to investigate and learn using prepared lesson plans and experiments that they could design. They were required to learn the assigned material and then present it to their "jigsaw" group which consisted of members from different "expert" groups. Each student in the "jigsaw" group presented what they learned to the other members of the group. Eight physics classes were randomly assigned to either the control or experimental group. The control groups experienced traditional instruction while the experimental groups experienced the "jigsaw" method. All students were given a teacher-constructed pre-test prior to the experiment. This instrument was tested for reliability, and received a Cronbach's alpha value of .51. At the end of the experiment, all students were given a teacher-constructed post-test which was different from the pre-test. The Cronbach's alpha value for the post-test was .56. A multivariate analysis of variance (MANOVA) comparing pre-test and post-test scores showed a significant main effect of method of instruction. However, when univariate tests were done, no significant

effect of method of instruction on academic achievement was observed. Methodologically, this study was relatively sound in design and statistical analysis. However, group differences were not controlled for except in the MANOVA where pre-test scores were used as the covariate. In addition, the reliability scores of both assessment instruments were below .7, which is considered by most researchers to be the threshold that describes the internal reliability of an assessment instrument. This being the case, drawing definitive conclusions regarding the relative effect of the "jigsaw" method of cooperative learning is suspect.

Statement of Problem

Research on the effectiveness of cooperative learning has revealed its value as an instructional intervention that can increase student achievement at all levels of education. Various meta-analyses that have been performed in the past 30 years support this notion reporting positive effect sizes regarding the intervention. In addition, original research conducted in the past 10 years echoes this same conclusion. Unfortunately, the most recent meta-analysis focusing specifically on the effects of cooperative learning on student achievement was that of Johnson, Johnson, and Stanne (2000) and it did not disaggregate data based on academic discipline. The most recent meta-analysis on the effects of cooperative learning in a science discipline was performed by Bowen (2000) and focused specifically on chemistry achievement. Moreover, it has significant methodological flaws that call into question the conclusion regarding the effects of this particular intervention on student achievement. Original research on the effects of cooperative learning on student achievement in science disciplines has been conducted in recent years, and provides reliable conclusions as to the value of this intervention in

improving student achievement. However, since no meta-analyses on the effect cooperative learning has on science achievement have been conducted in the past 8 years, a gap in the research exists which could be filled by such a study.

The purpose of the study proposed here is to conduct a systematic review and meta-analysis of the effect of cooperative learning on science achievement of secondary and early post-secondary students. This type of study adds to the body of research that exists on cooperative learning by providing a quantitative synthesis of available research conducted between 1995 and 2007.

Research Questions

The research question for this study is as follows:

1. What is the effect of cooperative learning on student achievement in science disciplines in comparison to traditional instruction?
    a. Additionally, does cooperative learning have a differential effect on student achievement based on student demographics such as gender, ethnicity, socioeconomic status, and ability level?

Definition of Terms

Academic controversy – "when one person's ideas, information, conclusions, theories and opinions are incompatible with those of another, and the two seek to reach an agreement" (Johnson & Johnson, 1999, p. 234)

Achievement – a student's performance on an academic task (Wikipedia, 2008)

Assessment – "the process of documenting, usually in measurable terms, knowledge, skills, attitudes and beliefs" (Wikipedia, 2008)

Competition – "a social situation in which the goals of the separate participants are so linked that there is a negative correlation among their goal attainments" (Johnson & Johnson, 1999, p. 234)

Cooperation – working together to accomplish shared goals (Johnson, Johnson, & Holubec, 1993)

Cooperative learning – "students working together to accomplish shared learning goals and maximize their own and their groupmates' achievement" (Johnson & Johnson, 1999, p. 234)

Effect size – "statistical measure of the size of an effect of an intervention, or of the relationship between two or more variables (Petticrew & Roberts, 2006)

Group accountability – "the overall performance of the group is assessed and the results are given back to all group members to compare against a standard of performance" (Johnson & Johnson, 1999, p. 236)

Group processing – "reflecting on a group session to (a) describe what member actions were helpful and unhelpful and (b) make decisions about what actions to continue or change" (Johnson & Johnson, 1999, p. 236)

Individual accountability – "the measurement of whether or not each group member has achieved the group's goal [;] assessing the quality and quantity of each member's contributions and giving the results to all group members

Intervention – "term used to refer to an action intentionally undertaken to bring about some beneficial outcome" (Petticrew & Roberts, 2006, p. 280)

Meta-analysis – "quantitative synthesis of study findings as a part of a systematic review" (Petticrew & Roberts, 2006, p. 280)

Moderator variable – "variable that affects the relationship between two other variables" (Petticrew & Roberts, 2006, p. 281)

Outcome – "the effects of an intervention" (Petticrew & Roberts, 2006, p. 281)

Promotive interaction – "actions that assist, help, encourage, and support the achievement of each other's goals" (Johnson & Johnson, 1999, p. 239)

Quasi-experimental design – "studies that benefit from the use of one or more control groups, but without random allocation of participants" (Petticrew & Roberts, 2006, p 63)

Randomized controlled trial (RCT) – "experimental study in which participants are allocated randomly to receive an intervention of interest to the researcher, or to a comparison group (who may have received a different intervention, or none at all) (Petticrew & Roberts, 2006, p 282)

Social interdependence – "when each individual's outcomes are affected by the actions of others" (Johnson & Johnson, 1999, p. 239)

Zone of proximal development (ZPD) – the zone between what a student can do independently and what she or he can achieve while working in cooperation with other students or under the guidance of an instructor (in Johnson & Johnson, 1999)

Delimitations of the Study

A variety of delimitations will be placed on this study. In a meta-analysis, delimitations are referred to as "inclusion criteria" (Petticrew & Roberts, 2006, p. 63) that identify which studies to include in the analysis and which studies to disregard. For this study, the delimitations include the timeframe for studies to be included. This timeframe will be 1995-present. This span of time was chosen since the most recent meta-analyses on cooperative learning were reported in 1999 and 2000, and will allow inclusion of

some of the studies used in those analyses as well as the most recent ones. Additionally, only studies that report using cooperative learning in science disciplines will be included. Studies that identify student achievement in a science discipline with reliability analysis of assessment instruments will be included. The demographics of the subjects in the included studies will also be specified. Only studies identifying the subjects as secondary and early post-secondary (first two years at a university or any junior or community college students) will be included. Studies using both randomized controlled trial and quasi-experimental methodology will be included. Studies published only in English will be included since the author does not possess the resources to translate studies published in languages other than English. Since meta-analysis uses the effect size statistic to report the effect of an intervention, only studies that report parametric statistics which allow for the calculation of effect sizes will be included.

Limitations of the Study

As with any type of research study, a systematic review and meta-analysis has its limitations. In this type of methodology, one of the most common uncontrollable sources of error is that the researcher may not find all of the written research regarding the intervention being assessed. That is certainly the case in this study. Despite this limitation, meta-analysts often make the assumption that the population of published and fugitive studies is representative of all the studies that have been performed on a particular intervention (Lipsey & Wilson, 2001). That same assumption will be made in this study. Moreover, this study will assume that the actual implementation of cooperative learning in the studies included in the meta-analysis is the same as what was described in the study.

Significance of the Study

A review of the research on the effects of cooperative learning on science achievement reveals that this particular intervention is effective in increasing student achievement in various academic disciplines in comparison to traditional instruction. Additionally, recent original research studies have shown that cooperative learning has a positive effect on science achievement, but no systematic reviews or meta-analyses assessing the impact of this intervention specifically have been done since 2000. While two of the studies of this type that were published in 1999 and 2000 reported positive effects of this intervention on science achievement, the study by Bowen (2000) was methodologically flawed and the Springer, Stanne, and Donovan (1999) study was completed almost ten years ago. A review of existing research reveals a significant gap in the knowledge base regarding cooperative learning and its impact on science achievement on a larger scale. The study fills that gap and provides an updated, thorough review for educational researchers and practitioners. In addition, a systematic review of cooperative learning will also provide a great deal of clarity on the various types of CL in use and the protocols by which it is implemented in classrooms. My experience as a science educator has provided me with a broad knowledge base regarding instructional interventions, and a review such as the one that is being proposed here would be a very valuable tool and reference for me. I have used various types of cooperative learning in my own classroom for a number of years and believe in its effectiveness in improving student learning. However, my attempts to support its worth and large-scale implementation have typically gone unnoticed. A review of the research on CL would provide me with the type of evidence that I could use to convince other educators of the

value of CL as a tool for improving student learning and possibly help me increase its appeal and distribution. I plan to pursue professorial employment at the university level upon completion of my doctoral degree. Completing a study such as this one would provide me with at least two skill sets I could present to potential employers regarding my expertise in educational research and practice. First, I will increase my knowledge regarding meta-analytic research, a technique that is gaining in popularity and validity in social science research. Second, I will become an expert in the theory behind CL and its implementation, thus giving me the ability to teach the intervention to both pre-service students, in-service practitioners, and other professors.

In the 15 years that I have been an educator, I have had considerable experience with a variety of instructional interventions. One of my favorites and one that I use regularly in my own teaching is cooperative learning. Consequently, I have a high degree of bias regarding the use and potential benefits of this particular intervention. That being said, I am biased toward reporting that CL has a positive effect on student achievement in science. Despite this particular bias, I plan to review any and all research regarding CL that I encounter and will report positive, neutral, and negative effects where appropriate.

CHAPTER 2: REVIEW OF LITERATURE

What is Cooperative Learning?

Cooperative learning (CL) is an instructional strategy with a rich history in education and a multitude of definitions provided by a number of authors. The primary description of CL includes students working together on specific tasks assigned by the teacher (Johnson, Johnson, & Holubec, 1993; Murray, 2002; Slavin, 1983; Slavin, 1986). While this description indicates interaction among students, far more detail is needed in order to truly characterize this instructional structure. Johnson and Johnson (1999) state that in CL, students work together to maximize each other's learning and their own learning. Slavin (1995) adds that CL methods are designed to encourage students to work together to master material that has been initially presented by the teacher. Goal-sharing is another characteristic of CL which is important to its structuring and effectiveness. Johnson, et al. (1993) state that "cooperation is working together to accomplish shared goals" (p. 15). Slavin (1983, 1995) and Kagan (1994) also discuss the importance of shared goals in the implementation and success of CL. In order to encourage students to work together successfully, individual and group accountability must be part of the assessment structure. In addition, students must be taught group interaction strategies for CL to be successful in affecting their learning (Murray, 2002).

Teachers must therefore be educated on these various strategies so that they can facilitate the CL activities adequately and successfully in the classroom.

Elements of Cooperative Learning

Cooperative learning has a number of basic elements that must be described in order to fully understand this instructional intervention. Kagan (1994), Johnson and Johnson (1999), and Slavin (1995) all describe different numbers of these elements, but they are all similar. Due to these similarities, a synthesis of their ideas will be presented here. First, positive interdependence is necessary for cooperative learning to be successful. In positive interdependence, each member of a team of students is equally important in the success of the group and the success of the other individuals in the completion of the assigned academic task. Kagan (1994) describes three forms of positive interdependence. In its weakest form, each team member's success may contribute to other member's success and the success of the team is likely to be facilitated by the success of each individual member. Intermediate forms of positive interdependence occur when the success of each member contributes to the success of all members, but any member could also succeed on their own. Moreover, the success of the team is facilitated by the success of each member, but the team could succeed without the contribution of every member. In its strongest form, the success of every team member is not possible without the contribution and success of the other members and the success of the group is not possible without the contributions of each member. Structuring positive interdependence is important to student success in CL and involves three steps (Johnson & Johnson, 1999). The first step is assigning each group a clear and measurable task. Group members must know exactly what they to accomplish. The second step involves

structuring positive goal interdependence. For this to occur successfully, members must believe that they can achieve their own individual academic goals if and only if their teammates attain their goals. Members know that they cannot succeed unless all members of the group succeed also. The third step requires the teacher to supplement positive goal interdependence with other types of positive interdependence. This step can be accomplished in a number of ways. When a group successfully completes an academic task at an acceptable level of proficiency, the teacher may organize some type of celebration. In addition, teachers may also add bonus points to each group members' grade when everyone in the group scores above a certain level or criterion on other classroom assessments. Teachers may also provide nonacademic rewards for students, such as extra free time, stickers, food, or extra recess time. According to Kagan (1994), structuring positive interdependence is similar to what Johnson and Johnson (1999) describe. Members of a team have the same goal, team rewards are based on the contributions of all members, tasks are structured so that they cannot be completed alone, resources are available, and each team member has a specific role.

The second element of CL is individual accountability. Individual accountability exists when the performance of each individual is measured against a standard of performance and the individual is held responsible by team members for contributing her or his fair share to the group's success. Individual accountability results in team members understanding that they cannot rely on the performance or work of other members and must continue to strive to meet group goals (Johnson & Johnson, 1999). Individual accountability can be structured and assessed in a variety of ways and take different forms. In terms of group make up, individual accountability is easier to measure

when groups are small. Small groups prevent individuals from "free riding" (Kagan, 1994) and not putting forth individual effort to achieve group goals. With regard to assessment of learning, group members can all be required to complete individual assessments which focus on learning that took place during completion of the group task. Additionally, individuals can be assigned specific tasks involved in completing the group task and then be assessed independently of the other members based on their own performance on their assigned portion of the group task. Team scores can also be based on these individual scores. Teachers can also give oral examinations to individuals at random to determine what they learned. Group observations are a helpful tool in assessing the contributions of each member. One member of a group can be assigned the role of checker, who asks other members to explain the reasoning and understanding that underlie group answers. Lastly, students can be asked to teach what they learned in their groups to other students (Johnson & Johnson, 1999; Kagan, 1994).

Group accountability exists when the overall performance of the group is assessed and the results are reported back to the group to compare against some standard of performance and most often results from the development of individual accountability (Johnson & Johnson, 1999). When individual group members understand that their own performance can have a significant performance of other group members, a greater sense of accountability develops, and groups are more likely to succeed in completing academic tasks.

A third element of CL is face-to-face promotive interaction/simultaneous interaction. Students should work in their groups face-to-face and work together to promote each other's success (Johnson & Johnson, 1999). All groups in a classroom will

work on group tasks simultaneously, making the teacher more of a facilitator of the interactions rather than a purveyor of information (Kagan, 1994). Structuring these interactions is relatively simple. In contrast to traditional, didactic instruction where the teacher does most of the talking in a classroom, and typically the only students interacting are those asking questions or being called on, teachers schedule specific time during a class period in which groups are allowed to work on their assigned tasks. In small groups, all students are interacting with their team members. Teachers can also monitor and observe group interactions and encourage promotive interaction by asking groups how they are progressing in their task and also encouraging the groups to ensure that all members are involved in the interactions (Johnson & Johnson, 1999; Kagan, 1994).

A fourth element of CL involves interpersonal and small group skills/equal participation of group members. In order for groups of students to function effectively in a cooperative environment, they must be taught the social and teamwork skills required for quality collaboration and be motivated to use them so that their group will be effective. Students must get to know and learn to trust each other, communicate clearly and effectively with each other, support one another, and resolve conflicts in a constructive manner (Johnson & Johnson, 1999). Participation in cooperative group tasks is essential for student learning and success (Slavin, 1995) and relates directly to social and teamwork skill development. Equal participation by students ensures that each member has an active role in completing the cooperative task and that each member has equal opportunity to learn from the task and interactions. Structuring equal participation in cooperative groups facilitates the development of social skills as well. In order to

structure equal participation effectively, a number of strategies may be employed by classroom teachers. First, each student can be given turns in speaking during cooperative activities. In addition, tasks can be divided among group members so that each member has specific duties to complete that will contribute to the completion of the group task. Many specific forms of CL rely heavily on a division of labor of tasks (Kagan, 1994).

The final element of cooperative learning is group processing (Johnson & Johnson, 1999). Effective and meaningful cooperative work is influenced directly by whether or not groups engage in active reflection on how they are functioning collectively. Group processing can be defined as reflecting on a group session in order to describe what member actions were helpful and not helpful, and make decisions regarding what actions to continue or change. The value in group processing lies in the ability of a group to clarify and improve the overall effectiveness of each member in contributing to the achievement of the group's goals. Structuring group processing has five steps that a teacher can utilize. The first step is to assess the quality of the interactions among group members by observing the groups as they work together. This allows the teacher to determine if each member is functioning in such a way as to maximize each other's learning. The teacher can use a formal observation checklist to gather specific information on each group. By doing this, a teacher can gain better insight into what students do and do not know about the subject matter being studied in the cooperative work. Additionally, teachers can appoint student observers in each group (a role which can be rotated among members) to assess group dynamics. Another technique would involve having each group member fill out a checklist that assesses the frequency with which they engaged in task completion.

The second step in examining group processing is to give each group feedback regarding their group skills. Teachers must allocate time at the end of each class in which cooperative learning was utilized in order to give groups feedback so the groups can further process how they functioned. Data from the checklists that the teacher completed and that the students completed can be used to focus discussions among groups. Small group processing allows each group to celebrate their successes, maintain good working relationships, and hold each member accountable.

The third step in examining group processing is for groups to set goals related to improving their effectiveness. Group members can suggest ways to improve teamwork and then the group can decide as a whole which suggestions to adopt. By receiving feedback and reflecting on group functioning, the group will function better in the future.

The fourth step is for the teacher to process how effectively the whole class is functioning. At the end of a given class period, the teacher can conduct a whole-class processing discussion during which the teacher shares the results of her or his observations with the class. In addition, if each group had a student observer, that individual could report their own results. Sharing observations among the whole class may provide differing viewpoints regarding group processing that were not discussed in individual groups.

The final step in group processing is to conduct whole-class and small-group celebrations. Feeling respected, appreciated, and successful helps students build commitment to learning and working in cooperative groups.

The Psychological Basis of Cooperative Learning

Social Interdependence Theory

The most influential theorizing on cooperative learning focused on the notion of social interdependence, that groups were dynamic entities in which the interdependence among members could vary in particular ways. This idea was proposed by Koffka in the early 1900s and further refined by Lewin in 1935. Lewin stated that the essence of any group was based on the interdependence among members which resulted in groups being a dynamic whole so that any change in the state of any one group member changes the state of the group. He also postulated that any intrinsic tension within the group motivates the group toward accomplishment of group goals. Deutsch (1949) refined Lewin's ideas and developed a theory of competition and cooperation noting that interdependence can be either negative (competition) or positive (cooperation). Johnson and Johnson (1974, 1989) extended Deutsch's work into social interdependence theory.

Social interdependence theory states that the way that social interdependence is structured determines how individuals interact. This in turn determines the outcome of group interactions. Cooperation (positive interdependence) results in promotive interaction, during which individuals facilitate and encourage each other's learning efforts. Competition (negative interdependence) results in oppositional interaction during which individuals discourage each other's learning efforts. In the absence of interdependence, there is no interaction and individuals work independently without any cooperation. Promotive interaction results in increased efforts to achieve, positive interpersonal relationships, and improved psychological health. Oppositional or no

interaction typically results in decreased efforts to achieve, psychological maladjustment, and negative interpersonal relationships.

Slavin (1995) asked the question "What's wrong with competition?" (p. 3). Educators have long known about the negative effects of competition in the classroom when structured poorly. But what aspects of competition result in negative affects on students? Competition is defined as working against other individuals to achieve a goal that only one or a few students can attain (Johnson & Johnson, 1999). In competitive classroom situations, individuals seek goals or outcomes that are beneficial to themselves and usually detrimental to others. This type of learning focuses student efforts on performing faster and more accurately than other students, with students perceiving that they can only succeed if others fail (Deutsch, 1972; Johnson & Johnson, 1989). Pepitone (in Slavin, Sharan, Kagan, Lazarowitz, Webb, & Schmuck, 1985) reported that student performance in cooperative groups was significantly higher than performance when students were competing against one another. In addition, interpersonal interactions between students, including task-oriented behaviors, were significantly higher in the cooperative situations in comparison to the competitive situations. This should come as no surprise since competitive learning precludes the use of any interpersonal interactions.

Individualistic learning, in contrast to competitive learning, involves students working toward similar goals and ensuring their own learning independently from the efforts, learning or achievement of other students (Johnson & Johnson, 1999). Johnson and Johnson (in Slavin, et al., 1985) summarized a number of studies that support the notion that cooperative learning results in great academic achievement than individualistic learning. They also reported much greater interpersonal interactions

among students, which is not surprising since individualistic learning is similar to competitive learning and does not involve any interactions between and among students.

Cognitive-Developmental Theory

In addition to social interdependence theory proposed initially by Lewin (1935) and further developed by other authors, cognitive-developmental theory has also had a significant impact on cooperative learning from a psychological and developmental standpoint. The cognitive developmental perspective is primarily based on the theories of Jean Piaget (1950) and Lev Semenovich Vygotsky (1978), as well as cognitive science, and academic controversy (Johnson & Johnson, 1979, 1995). According to Piaget, cooperation can be described as "striving to attain common goals while coordinating one's own feelings and perspective with a consciousness of others' feelings and perspectives" (Johnson & Johnson, 1999, p. 187). Piaget also postulated that social-arbitrary knowledge (language, morality, and symbol systems (such as mathematics and reading) can only be learned through interactions with other individuals (Slavin, 1995). From this perspective comes the premise that when individuals cooperate and socio-cognitive conflict arises, that creates cognitive disequilibrium; this stimulates alternative perspective taking and cognitive development. Teachers working within the Piagetian perspective may place students into pairs in which each student holds an opposing viewpoint on a particular subject or where the students disagree on the solution to a particular problem. The students must work together until they can agree or come to a common answer (Murray, in Thousand, Villa, & Nevin, 2002). CL in the Piagetian tradition accelerates a student's intellectual development by forcing her or him to develop

consensus with students who hold contrasting points of view regarding the academic task that has been assigned.

Vygotsky and other theorists claim that distinctively human mental functioning has its origins in social relationships (Johnson & Johnson, 1999; Murray, in Thousand, Villa, & Nevin, 2002). This mental functioning can be defined as the "internalized and transformed version of the accomplishments of a group" (Johnson & Johnson, 1999, p. 187). Knowledge is socially constructed from cooperative efforts to learn and solve problems. One of Vygotsky's central concepts is the zone of proximal development, the zone between what a student can do independently and what she or he can achieve while working in cooperation with other students or under the guidance of an instructor. In this view, cooperative activity among children promotes intellectual growth because they will likely be operating within each other's zone of proximal development (Slavin, 1995). If students do not work cooperatively, they will not grow and develop intellectually, so from this perspective, the amount of time that students spend working on tasks independently should be reduced.

Cognitive Science

Cognitive science involves modeling, scaffolding, and coaching (Johnson & Johnson, 1999; Murray, in Thousand, Villa, & Nevin, 2002). The learner must restructure information and rehearse it in a cognitive manner for it to be retained and incorporated into existing knowledge. One of the most effective ways of accomplishing this goal is reciprocal teaching, in which the students and teacher take turns as teacher. In teaching reading, the teacher and the student both read a particular passage and the teacher demonstrates the process of formulating questions regarding the passage

(modeling). Then the teacher coaches the student in comprehension skills and works with the student until the student achieves the skill. Over time, the student will develop a new conceptual model for the skill being taught (Murray, in Thousand, Villa, & Nevin, 2002). Peer tutoring is a similar strategy which can be employed that is collaborative and enhances learning. In this situation, a peer explains material being learned to a collaborating student, which is essentially another form of cooperative learning.

Academic Controversy/Controversy Theory

Academic controversy or controversy theory (Johnson & Johnson, 1979, 1995) states that when students are confronted with opposing points of view, uncertainty or conceptual conflict is created. In turn, this conflict encourages students to search for new information regarding the conflict and then reconceptualize what they are learning into a more thoughtful and refined conclusion. Structuring this type of activity involves placing students into groups of four, where one pair is given the pro position on an issue and the other is given the con position on the issue. Each pair is then instructed to define their position by seeking out information regarding the issue. Each pair of students then presents and advocates their position to the other pair. The pairs then engage in discussion during which disagreement occurs. Students then attempt to find a synthesis of the two viewpoints on which all members can agree, without advocating their original position.

Behavioral Learning Theory

Behavioral learning theory posits that students will work hard on tasks that secure some type of reward and will typically fail on tasks that yield no reward or yield some type of punishment (Bandura, 1977; Skinner, 1968). Cooperative learning should

therefore be designed to provide incentives for group members to participate in group efforts. From a motivational perspective, it is important for cooperative learning to focus primarily on reward structures under which students operate, such as individual rewards and group rewards. Cooperative goal structures create a situation where the only way individuals can attain their own personal goals is if the group is successful in attaining group goals or completing group tasks. Individuals must work collaboratively and help other members of the group if they are to be successful (Murray, in Thousand, Villa, & Nevin, 2002; Slavin, 1995). Group reward structures encourage group members to help each other be successful in completing academic tasks. Methods for structuring group rewards have been mentioned previously.

Critiques of Cooperative Learning

Despite the fact that the motivational and cognitive theories support the benefits of cooperative learning on student achievement, one major criticism of this instructional method is the "free rider" (Slavin, 1995, p. 19) effect, in which some group members do most of the work while others do not, yet still receive grades reflecting participation. This particular situation usually occurs when a group has a single specific task to complete rather than a task with multiple parts. Assignments such as these can allow some students to do all the work, leaving others out of the picture. In addition, some students in a group may appear to be more skilled academically, and will often complete to work independent of the other, possibly less-skilled group members. This is a problem called "diffusion of responsibility" (Slavin, 1995, p.19), which can be detrimental to the achievement and learning effects of cooperative learning. The "free rider" effect and "diffusion of responsibility" effect can be dealt with in specific ways. One of the ways to

deal with these issues is to structure the cooperative learning assignments very clearly so that multiple tasks must be completed, rather than a single task. Another is to make each group member responsible for a specific part of the group task. Lastly, making students individually accountable for their own performance forces each student to complete their assigned task so that they can receive credit for their efforts.

History of Cooperative Learning

Cooperative learning has been part of educational philosophy and delivery for a long period of time. Educational philosophers such as John Dewey, Kurt Lewin, and Lev Vygotsky all referenced cooperation as a means to learning in the early 1900s. Dewey argued that if humans are to live cooperatively, then they must experience this living process in the schools (Schmuck in Slavin, et al., 1985). The philosophies of Lewin and Vygotsky have already been reported in this document. Psychologists such as B.F. Skinner and Jean Piaget both referenced and discussed the importance of cooperation in the development of social and intellectual skills in the 1950s and 1960s. Morton Deutsch engaged in research on the effects of competition and cooperation on learning in the 1940s, which was continued by David and Roger Johnson in the 1970s and into the present. Cook, Madsen, and Kagan all engaged in research on cooperation and competition in children and their effects on learning in the 1960s (Johnson & Johnson, 1999). All of these previous studies and writings led to the implementation of cooperative learning structures into teacher professional development and subsequently into classrooms in the United States. David Johnson first began training teachers in the use of cooperative learning at the University of Minnesota in 1966 and was joined by his brother Roger in 1969. Many educators began to develop more varied and specific types

of cooperative learning structures and methods, all of which will be described in more detail later. Johnson and Johnson developed the "Learning Together and Alone" method of cooperative instruction in the mid-1960s. In 1973, DeVries and Edwards combined an instructional games approach with intergroup competition, a technique which would come to be known later as "Teams-Games-Tournaments (TGT)." Slavin began to develop cooperative curriculum in the mid-1970s. In 1976, Sharan and Sharan developed a technique called "Group Investigation." Aronson developed the "Jigsaw" structure in 1978. Slavin developed the "Student Teams Achievement Divisions (STAD) structure in the late 1970s and in the early 1980s, developed the structure known as "Team Accelerated Instruction (TAI)." Cohen developed the "Complex Instruction" structure in the early 1980s. During the same time period, Kagan developed the "Co-op co-op" structure which places teams in cooperation with each other. In 1985, the American Educational Research Association (AERA) and the Association for Supervision and Curriculum Development (ASCD) formed special interest groups to examine cooperative learning and its affects on instruction and learning. According to Johnson and Johnson (1999), cooperative learning began to gain popularity among educators in the early 1990s.

Cooperative Learning and its Delivery in the Classroom

Cooperative Learning Groups

Cooperative learning can be defined by the types of groups being utilized for organization and delivery of group tasks (Johnson & Johnson, 1999). Formal cooperative learning groups typically work together for a minimum of one class period or up to several weeks and may have rotating membership. Any academic assignment or task can

be assigned to this type of group. This kind of group ensures that students are actively involved in the academic work that is assigned. Informal cooperative learning groups are temporary groups that last from a few minutes to one class period and also have rotating membership. These types of groups can be used to help direct instruction and to focus students' attention on specific material they are responsible for learning. Cooperative base groups are long-term groups that can last up to a full school year that have stable membership. The purpose of these types of groups is to provide members with support, encouragement, and assistance as each member progresses academically.

Organizing students into groups can be done in a variety of ways (Johnson & Johnson, 1999). Random assignment is the simplest and involves dividing the number of students in a class by the size of the group the teacher desires. Then the students number off and those with the same number find each other. Other grouping strategies involve the teacher selecting specific student characteristics, such as achievement level, and then organizing the groups using the student characteristics. The least recommended grouping procedure involves allowing students to self-select their group members. This type of grouping often leads to homogeneous groups with similar students working together. Many authors suggest that the student makeup of groups be heterogeneous in order to provide the greatest potential for learning (Johnson & Johnson, 1999; Kagan, 1994; Slavin, in Slavin, et al., 1985; Slavin, 1995). Heterogeneity of groups can vary considerably in terms of how it is defined. A number of student characteristics can be considered when structuring heterogeneous groups. In order to facilitate equal learning among all students, teachers can group them according to achievement level. Students can be ranked in a list according to their prior or current achievement level in a class.

Then the teacher can construct groups of four students with one high achiever, one low achiever, and two middle achievers. Using this type of grouping provides learning opportunities for all students in a group because the high achievers can help the other members of the group to learn concepts related to the group task. High achievers learn better because they become "teachers" of the other students. Another student characteristic that can be considered when forming heterogeneous groups is cultural background (Johnson & Johnson, in Thousand, Villa, & Nevin, 2002; Sapon-Shevin, Ayres, & Duncan, in Thousand, Villa, & Nevin, 2002; Slavin, 1995). Grouping students using this characteristic helps develop a positive sense of diversity and community in a classroom and also promotes inclusion of students in the learning process whose background differs from the mainstream. Gender can also be considered so that males and females work with each other and not just with the same gender.

Types of Cooperative Learning Structures

A multitude of CL structures exist and are practiced at all educational levels, from K-12 to higher education. Johnson, Johnson, and Stanne (2000) identified ten specific structures in a meta-analysis of the effectiveness of CL on student achievement.

Learning Together and Alone

The first structure that will be described is known as "Learning Together and Alone" (Johnson & Johnson, 1975/1999). This particular structure involves grouping students into heterogeneous groups in various ways (as previously mentioned) in order for them to complete academic tasks. Groups can be structured as informal, which typically stay together no longer than a given class period, formal, which can stay together for up to several weeks, or cooperative base groups, which may stay together for

up to an entire school year. Most of the characteristics of this structure have been mentioned in the descriptions above. Students are held individually accountable for their own performance and groups are held accountable for the performance of the group. Group performance can be rewarded by giving bonus points to individual students on individualized assessments, such as exams which assess material learned in the groups.

Student Teams-Achievement Divisions

"Student Teams-Achievement Divisions" (STAD) was developed by Slavin in 1978 and is one of the simplest CL structures for teachers to start using if they are unfamiliar to the cooperative approach (Slavin, 1995). "STAD consists of five major components: class presentations, teams, quizzes, individual improvement scores, and team recognition" (p. 71, Slavin, 1995). Class material is initially introduced to the students by the teacher in a lecture-discussion format. These presentations differ from usual lecture in that they are focused specifically on the STAD unit. Students will therefore focus their attention during the presentation because they know that they will be assessed on their knowledge of the material and that their quiz scores determine their team scores. Teams are composed of four or five students and are heterogeneous in terms of achievement, gender, and cultural background. The major function of the team is to make sure that all team members are learning in order to prepare each member for the quizzes that will be taken at the end of the STAD unit. After the teacher presents the material, the teams work together on worksheets or other assignments. Team functions most often include discussing questions related to the worksheets, problem-solving, comparing answers on assignments, and correcting any misconceptions regarding the material. The team is the most important feature of STAD. Emphasis is placed on

members doing their best for the team and on the team doing its best to help each of the members. After approximately one or two sessions of teacher presentation and one or two sessions of team practice, students complete individual quizzes during which they are not allowed to help other team members. As a result, each member is responsible for knowing the material themselves. Individual improvement scores are designed to give students a performance goal that can be reached if she or he works harder and performs better on quizzes than in the past. Each student is given a "base" score derived from averages of grades on previous quizzes. Students then earn improvement points for their teams based on the degree to which their quiz scores exceeded their base scores. Teams are recognized with certificates or other reward for their performance if their average scores exceed a certain criterion defined by the teacher. In addition, team scores can often be used to determine up to 20 percent of a student's individual grade.

Teams-Games-Tournaments

"Teams-Games-Tournaments" (TGT) was developed by DeVries and Edwards in 1974 and is very similar to STAD in structure and delivery. The class presentations and the teams are structured the same way that they are in STAD. The games that are played in TGT are composed of content-related questions that are designed to test the knowledge that the students learned from the class presentations and the team practice. The games are played at tables of three students who each represent a different team. Each table has numbered questions on a worksheet and cards with numbers that correspond to the numbers on the worksheet. A student picks a numbered card and attempts to answer the question on the worksheet. Players are allowed to challenge each other's answers. The tournament is the structure in which the games at each table take place and is usually held

at the end of a week or unit after the class presentation and team practice. In the first tournament, the teacher assigns the highest three students based on past performance to table 1, the next three highest students to table 2, and so forth until all the students are at a table. This equal competition makes it possible for students at all achievement levels to contribute as much as they can to their team if they do their best. After the first tournament, students switch tables depending on their performance in the previous tournament. The winner at each table advances to the next higher table, the second scorer stays at the same table, and the third scorer moves to the next lowest table. Table movement accounts for any students who may have been misplaced in the first table assignment since they will eventually move up or down depending on their performance. At the end of the entire tournament, team scores are compiled and team recognition is conducted in the same manner as in STAD.

Group Investigation

"Group Investigation" (GI) is a form of CL that dates back to John Dewey (1970), but has been refined by Sharan and Sharan (1976, 1992) (Slavin, 1995). GI requires an educational environment that supports interpersonal dialogue, so initial efforts required of the teacher involve training the students in communication and social skills, and team building. These skills have been reported by many authors and can be developed in a number of ways. Group Investigation involves students acquiring, analyzing, and synthesizing information in order to solve a multi-faceted problem assigned by the teacher. Generally, the teacher designates a broad topic for study, which the students break down into subtopics that they may choose from for their investigation. The students then conduct their investigation using a variety of sources and then evaluate and

synthesize information in order to produce a group product. Group products may include written reports or oral presentations. Implementation of GI involves six stages through which students progress. During the first stage, the teacher presents a problem to the entire class and asks the students what they would like to know about the problem. Students meet in informal groups where each student expresses their interests and ideas about what to investigate. Recorders in each group keep track of the ideas, which progresses into larger groups, and ultimately, a final list of topics is compiled which represents the interests of all the students. Then the teacher presents this list to the entire class so that students can choose the topic they are interested in studying. Groups are formed based on these interests, but the teacher may want to limit the number of students in each group and should ensure that each group is heterogeneous in its composition. In stage two, each of the groups plans their investigation and how members will contribute to the final product, including how the work will be divided. This process can be student-driven but should be monitored by the teacher. During stage three, each group carries out their investigations, which is typically the longest stage. Students may gather, analyze, and evaluate information individually or in pairs. When individuals or pairs complete their parts of the investigation, the group reassembles and shares their new knowledge. Groups can compile a written summary or the entire investigation, or each member can provide a summary of their part of the investigation. The fourth stage involves preparing the final report, and is essentially a transition from the information gathering stage to the stage where the group presents its report to the class. Each group also plans a presentation that will be engaging and informative. At the end of the investigation stage, the teacher asks each group to identify a representative for membership on a committee

which will hear each group's presentation ideas and determine if these are feasible. This committee will also coordinate a presentation schedule for the groups. The teacher should act as an adviser to this group to ensure that the presentation ideas and schedule are feasible and meet the requirements of the assignment. The fifth stage of GI is where the groups make their final presentations to the entire class. The teacher can identify guidelines for presentations which may include the use of audiovisual equipment, speaking clearly, learning stations where classmates can perform specific tasks, and visual displays. The final stage of GI involves the teacher evaluating student achievement. Due to the nature of the investigations that students perform, evaluating student achievement in GI can be a challenge. Teachers should evaluate students on their higher-level thinking regarding the subject they studied rather than just the basic facts and concepts that were learned. Constant assessment of student progress is possible through observations of group work during the investigation. Teachers may opt to test students on their knowledge, but should take into account the different levels and types of learning exhibited by the students when developing assessments such as these. Another possible form of assessment is peer evaluation, where the teacher and the students develop an assessment tool comprised of questions from each group that relate to the most important ideas presented by each group. Each group would then be given other students' answers and would correct them.

Jigsaw

The "Jigsaw" structure is another CL strategy and was developed in 1978 by Aronson, Blaney, Stephan, Sikes, and Snapp. Jigsaw is similar to GI in that students perform research on a selected topic. However, the structure for this particular method

differs from other previously mentioned in a variety of ways. Slavin (1995) believes that the original Jigsaw format was more difficult to implement, so he developed the "Jigsaw II" format (1986) as an adaptation of the original. Both formats will be discussed here. The structure of student groups in this CL strategy should be heterogeneous as seen in the previous structures (e.g., STAD) and should contain 3-6 students. In the original Jigsaw formats, members of the "home" group are given unique reading assignments related to the material being studied. Each member reads their section of material and then new "expert" groups are convened, containing members of each "home" group with the same reading assignment. The "expert" groups then discuss their reading assignment and may complete worksheets or other assessments. Once the "expert" groups are finished with their discussions, the "home" groups reconvene and each member teaches the others about their specific topic. The "home" groups then prepare a team report as in GI and present it to the class. Teachers can then administer tests or quizzes on the material that the students complete individually. Team recognition and reward for this strategy can be done in the same fashion as in STAD. Jigsaw II is very similar to the original format. The main difference is that in the original format, each student gets a different reading assignment that the other members do not see, while in Jigsaw II, all members get all the readings. In this format, all students get to see all the information, but they are responsible primarily for their own specific section which they will discuss in their "expert" groups.

Constructive Controversy

The next CL format is called "Constructive Controversy" and was developed by Johnson and Johnson in 1979 and further reported in 1995 and 1999. This structure

requires a cooperative context and is actually an advanced form of CL. This structure should be used by teachers who have used other forms of CL and whose students are familiar with cooperative methods. In this format, teachers must choose a topic that is manageable by the students on which at least two documented positions (pro and con) can be developed. The instructional materials (readings, etc.) should be organized into pro and con packets for the students. Heterogeneous groups of four students will be constructed. These groups will be further divided into pairs with each pair being assigned pro and con packets. Each pair is then assigned tasks related to learning its position and the supporting information and arguments, and researching relevant information related to their position. Each pair will also prepare a series of arguments the support their position and prepare a persuasive presentation to be given to the opposing pair from their original group. Each pair then presents their point of view to the other pair, with each pair taking notes on the other's position. After presentations, the teacher should allow the students to discuss the issues by freely exchanging information and asking for support of each other's position. Students are free to compare strengths and weaknesses of each position. Teachers can encourage argument by taking sides when a pair is in trouble, play devil's advocate, and generally stir the debate. Next, the teacher can have the pairs switch positions and argue the other point of view, but without the benefit of the notes prepared by the other pair. Lastly, the pairs will stop advocating their position and the groups will try to reach agreement through consensus. After this discussion, the groups can write a group report that includes the joint position, with supporting arguments. In addition, teachers can test the students on both positions individually. Team reward and recognition can be structured as in STAD.

Complex Instruction

"Complex Instruction" is a CL technique developed by Cohen in 1986 and is similar to GI. This technique focuses on discovery-oriented projects, particularly in mathematics, science, and social studies. The task that is created should be a multiple ability task which has more than one answer or way to solve a problem, is intrinsically interesting and challenging to the students, allows different students to make different contributions, requires reading and writing, and requires a variety of behaviors and skills. Groups for this strategy should be heterogeneous. The teacher will introduce the primary concepts with a presentation of some type that presents a variety of topics for the students to investigate. After the presentation, teachers distribute written instructions which provide students with information on the task itself, and how to complete it. Instructions can have differing levels of specificity based on the age of the students with whom the teacher is working. Students in each group are assigned specific roles to ensure equal participation. In groups of four students, the roles are facilitator, reported, checker, and materials manager. The facilitator manages the group, keeps the members on task, and is responsible for finding out answers to questions that the group has. The reporter takes notes and will report the group's findings to the rest of the class. The checker ensures participation of the members and makes sure that everyone has finished their part of the assignment. The materials manager obtains any and all materials that the group needs to complete the task. In order to evaluate student performance, groups are required to complete a report and presentation on their particular topic. Teachers can grade the presentations using a variety of criteria and assign a group grade. Peer evaluation can also be implemented, with students evaluating the presentations of the other groups.

Testing can also be done on the topics that were presented by each group. Rewards can be offered as in the STAD method.

Team-Accelerated Instruction

"Team-Accelerated Instruction" (TAI) was developed by Slavin, Leavey, and Madden in 1986, and was originally known as "Team-Assisted Individualization." This particular structure was developed to assist students in learning mathematics principles and involves cooperative work and allows for individualized instruction by the teacher. Teams are set up as in STAD or TGT. Prior to the group work, students are given a pretest to assess their prior knowledge. This helps the teacher decide what particular mathematics lessons each student should be doing. The curriculum materials for the groups include guide pages and practice sheets. Following the pretest, the teacher teaches the lesson and the students are then given a starting point in their individualized mathematics unit. Students then form their teams and begin working on their materials, asking teammates and the teacher for help if necessary. The teacher can provide very specific individualized instruction while assisting each group as needed. Once a student has completed a set number of problems, they take a ten-item formative test individually to assess their skill level. A peer grades the test and if the student scores eight or more, they move on to the next set of problems. If the student does not get at least eight problems correct, the teacher intervenes and provides individualized instruction to help the student understand the problems and further develop their math skills. Students continue to take formative tests until they have completed all of the assigned problems. At this point, a student will take a unit test which is scored by a peer monitor. At the end of each week of instruction, the teacher computes a team score based on the average

number of units completed and the score on the unit tests. Criteria are established for team performance, and rewards are given accordingly. Students also complete facts tests twice per week to reinforce these concepts. After every three weeks, the teacher stops the TAI program and teaches specific skills necessary for new units.

Co-op Co-op

"Co-op Co-op" was developed by Kagan in 1992 and is similar to GI and allows students to learn about a topic in small groups and then share their understanding with the rest of the class. This technique involves nine steps. First, teachers can provide an initial set of lectures, materials, or experiences to introduce particular topics to the students. Once this has been done, a student-centered class discussion can be conducted in which students discuss the various topics so they may gain some understanding of the topics so they can choose what they want to learn about. Step 2 involves the selection of student teams as in STAD. In step 3, the teams select a topic for study. During this time, the teacher should circulate around the classroom and act as a facilitator. In step 4, each team divides their topic into mini-topics, dividing labor among the group so that each person covers one aspect of the team topic. During step 5, students work individually on their mini-topics using a variety of resources made available by the teacher. In step 6, the students present what they learned about their mini-topics to their group. This is similar to what happens in Jigsaw. Following the presentations, the groups discuss the team topic based on the mini-topic presentations. During this time, roles may be assigned within the group so that each group has a note taker, facilitator, and checker. In step 7, groups integrate the mini-topic information and prepare a team presentation. Team presentations are given in step 8. Each team may include a variety of activities during

their presentation. It can be useful for the teacher to lead a feedback session after each team presentation so that each team can learn presentation techniques from the other teams. The final step, step 9, is where the teacher evaluates student learning. This can be accomplished using peer evaluation of the team presentation by the whole class, having teammates evaluate each other's contributions to the team presentation, or an individual write-up or presentation of the mini-topic by each student, which is evaluated by the teacher.

Cooperative Integrated Reading and Composition

The last CL strategy to be discussed is "Cooperative Integrated Reading and Composition" (CIRC), which was developed by Stevens, Madden, Slavin, and Farnish in 1987. This strategy was developed specifically for reading and writing. CIRC consists of three primary elements: basal-related activities, direct instruction in reading comprehension, and integrated writing and language arts. If reading groups are already being used in a class, students are then assigned to two or three reading groups according to their reading level. Students are paired within their reading groups, with the pairs then assigned to teams composed of partnerships from two reading groups or levels. Team members receive points based on individual performance on a variety of assessments, and these points form a team score. Team rewards are based on whether teams meet certain performance criteria. Stories are introduced to the students in teacher-led reading groups in which the teacher sets the purpose for reading, introduces new vocabulary, reviews old vocabulary, and discusses the story after the students have read it. After the stories are introduced, students are given a packet which contains a series of activities for them to complete in their teams when they are not working with the teacher in a reading group.

As students complete these activities, their partners check them to see that they have been completed accurately. Students have daily expectations for the number of activities they complete, but they are allowed to go at their own pace. At the end of three class periods, students are given a comprehension test on the story, asked to write sentences using the vocabulary, and asked to read the word list aloud. Students are not allowed to help each other on these tests. The test scores are a major component of weekly team scores. One day per week students receive direct instruction from the teacher on specific comprehension skills as a whole class. After each lesson, students work on comprehension activities as a team, assess each other's knowledge, and discuss any problems that may occur. In order to teach writing, teachers us curriculum designed specifically for CIRC. This program uses "writers' workshops," where students write on topics of their choice, and teacher-directed lessons on skills such as writing comparative paragraphs, letters, and mystery stories. On all of the writing assignments, students write drafts after working with their teammates on their topic, and the work with their teammates to revise what they have written. Lastly, students are also asked to read a book of their choice at home every evening and then write a report on the book. By completing book reports on a weekly basis, students contribute bonus points to their teams. For regularly assigned book reports that are required, the students also earn team points.

Outcomes Addressed in Cooperative Learning Research

The primary student outcome that has been measured in research regarding the effectiveness of CL is student achievement. Johnson, Johnson, and Stanne (2000) noted that many research studies assessing CL have been "efficacy studies" (studies of short-

term effects; p. 4), rather than "effectiveness studies" (studies describing how CL is delivered and what the outcomes are like; p. 4). In their review of ten CL strategies, they examined only the "effectiveness studies" in order to determine the effects that CL has on student achievement. In studies such as these, two types of assessment tools have been used to measure student achievement. Standardized assessments (Popham, 1999) are tools that are typically developed by national or international groups of educators and scholars for large-scale distribution and have undergone rigorous reliability and validity testing. These types of assessments are typically criterion-referenced and measure student achievement in comparison to that of other students. Teacher-constructed assessments are also widely used in CL research. These types of assessments are created by outside researchers not directly involved in instruction, or by educators conducting research who are directly involved in instruction. One major issue regarding these types of assessments is reliability of the tool and how it measures student achievement. In many cases, these instruments undergo reliability and validity testing to ensure that they are measuring what the researcher is interested in assessing. Results from studies that do not perform reliability testing to teacher-constructed assessment instruments may be subject to criticism and not necessarily support positive effects, if reported.

Another outcome of research regarding the effectiveness of CL in increasing student achievement is student grades in a course. This outcome is reported in some studies (Bowen, 2000), but may be subject to criticism. A student's overall grade in a course may be influenced by any number of confounding variables that the researcher may not be able to control. For example, the cooperative learning component of a course may constitute only a small portion of the instruction that takes place in a classroom.

Some students may respond better to another type of instructional method, which could influence their grades as well.

A third outcome that is often cited in CL research is student performance on local or national standardized tests, such as the Scholastic Aptitude Test (SAT) or American College Testing Program (ACT). This outcome is reported in very few studies and would come under the same criticism as student grades as an outcome.

Research on Cooperative Learning and Student Achievement

Original Research

Primary studies conducted on the affect of CL on student achievement in science disciplines have shown similar results. In a study performed in Taiwan, Chang and Mao (1999) reported no effect on student achievement in earth science for cooperative learning in comparison to control groups for overall achievement ($F = .13, p < .05$), knowledge-level test items ($F = .12, p < .05$), or comprehension-level test items ($F = .34, p < .05$) but reported a statistically significant difference in student performance on application-level test items ($F = 4.63, p < .05$) for students in cooperative groups. Bilgin and Geban (2006), in a study performed in Turkey, reported a statistically significant difference in students' understanding of chemical equilibrium in cooperative groups compared to control groups (multivariate analysis of covariance results: Wilk's lambda = .483; $F(2,83) = 44.344, p < .05$) with treatment groups showing higher achievement. Bilgin (2006), in another study performed in Turkey, reported a statistically significant difference in student performance in chemistry for cooperative groups compared to control groups (analysis of covariance results: $F(1,86) = 65.289, p < .05$). Using the GI method of cooperative learning in Israel, Shachar and Fischer (2004) reported a

statistically significant main effect of cooperative learning compared to the control groups on student achievement in chemistry (multivariate analysis of variance results: $F(1,162) = 28.6, p < .001$.

Meta-analyses

The research base addressing the impact of CL on student achievement is very broad and dates back to the 1960s. A number of meta-analyses of the effect of CL on student achievement have been published in the past thirty years. One of the first of these studies was performed by Johnson, Maruyama, Johnson, Nelson, and Skon (1981). In this study they analyzed research that addressed differences in student achievement between cooperative learning and competitive and individualistic learning at all grade levels, K-12 and post-secondary. They used three meta-analytic techniques in their analysis: the voting method, the effect size method and the *z*-score method. In the voting method, raters read each study and determined if the findings by the original authors reported negative, no difference, or positive results. In the effect size method, data was extracted from the studies and the standard mean difference effect size was calculated. In the *z*-score method, *p*-values from the studies were used to calculate *z*-scores, which were then referenced in the appropriate statistical table to determine if any differences found in the studies were due primarily to chance. They analyzed 122 studies and found that each of the three methods showed an overall positive effect of CL on student achievement when compared to competitive or individualistic learning. That is to say, students who worked cooperatively on academic tasks scored higher on assessments than students working in competitive or individualistic educational environments. They also subdivided studies based on a variety of CL structures as reported in the original

research. When they analyzed studies in which groups did not compete versus studies in which groups did compete, there was no difference between these methods. When comparing cooperation versus competition, cooperation showed increased student achievement, with an effect size of .78. Cooperation with group competition was showed a smaller effect than cooperation with interpersonal competition (*ES* = .37). Cooperation in comparison to individualistic learning showed an effect size of .78. Cooperation with group competition showed an increase in achievement in comparison to individualistic learning (*ES* = .5). When comparing competition and individualistic learning, no difference in student achievement was found. When comparing other moderating variables, other results were noted. Sex or ethnicity of the subjects were not shown to have any difference because 90% of the studies pooled males and females and 93% of the studies did not identify ethnicity of the subjects. Subject area also showed no difference in effect. They reported that the smaller a group is, the greater the effect on achievement in cooperative efforts. They did not report an overall mean effect size nor did they analyze research methodology in their moderating variable analysis. Lastly, some authors (Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H., 2009) view the vote counting method as the least reliable meta-analytic technique. However, since Johnson, et al. (1981) used multiple methods to assess study inclusion, the results of their meta-analysis may still be viewed as valid.

In 1995, Qin, Johnson, and Johnson conducted a meta-analysis assessing the difference between cooperative and competitive efforts and the quality of problem solving that students undertake and their subsequent achievement. In this study, they examined 46 studies published between 1929-1993 in which the independent variables

were compared. They reported an overall mean effect size of .55, supporting the notion that cooperative efforts lead to higher student achievement than competitive efforts. Additionally, they analyzed a number of independent variables and calculated effect sizes based on these variables. The first variable was the type of problem-solving task that was utilized in a study. Linguistic problems are those that require written or oral efforts by the student. Non-linguistic problems are those that require the student to represent their responses with pictures, graphs, mathematical formulas, symbols, or motor activities. Non-linguistic problems showed a larger effect (.72) than linguistic problems (.37). Ill-defined problems are those that do not have a clearly specified goal, while well-defined problems do have a specified goal. Ill-defined problems showed a somewhat larger effect (.60) than well-defined problems (.52). The age of students was not determined to have any significant effect on student achievement. The quality of the methodology of studies analyzed showed varying effects. Studies of high quality had an effect of .68, medium quality studies had an effect of .34, and low quality studies had an effect of .47 though these differences were not significant. Lastly, the length of the study showed no significant difference in effect.

Johnson, Johnson, and Stanne (2000) performed a meta-analysis on cooperative learning methods and their affect on student achievement. In this study, they identified ten specific CL methods, which have all been described previously. They reported findings from 164 studies that investigated eight of the ten CL methods. The two methods that did not have empirical studies related to student achievement were "Complex Instruction" and "Co-op co-op." For the other eight methods, all had a significant positive effect on student achievement in comparison to competitive or

individualistic learning and instructional methods. They did not report an overall mean effect size, but those they reported showed varying, yet positive, effects on student achievement based on different types of CL. The authors separated the effect size data into 2 main categories comparing CL with competition, and CL with individualistic learning. The "Learning Together" method had the highest effect sizes, .85 and 1.04, respectively. The other methods, in order when comparing with competition, were as follows: Constructive Controversy, .67; STAD, .51; TGT, .48; GI, .37; Jigsaw, .29; TAI, .25; and CIRC, .18. When comparing to individualistic learning, effect sizes were as follows: Constructive Controversy, .91; GI, .62; TGT, .58; TAI, .33; STAD, .29; CIRC, .18; and Jigsaw, .13. The other variable they considered in their study was the structure of the CL method. They ranked studies based on whether the structure of CL was more direct, having more proscribed instructions, or more conceptual, with less proscribed instructions. They found that the more conceptual the structure of CL, the greater the achievement of the students.

The aforementioned studies represent some of the main meta-analyses that have been performed on the affect of CL on student achievement. None of these studies identified the specific academic discipline that the original research examined. Consequently, it is difficult to determine if CL has differential effects on achievement based on the academic discipline in which it is implemented. The focus of this dissertation is to determine what affect CL has on student achievement in science disciplines, and since 1999, only three meta-analyses on this issue have been published.

Springer, Stanne, and Donovan (1999), Bowen (2000), and Scott, Tolson, Schroeder, Lee, Huang, Hu, and Bentz, (2005) all published meta-analyses exhibiting

positive effects of cooperative learning instruction on academic achievement in science. The Springer, et al. (1999) study examined student achievement in undergraduate Science, Mathematics, Engineering, and Technology (SMET) disciplines and reported a positive mean effect size ($d = 0.51$, based on 37 studies analyzed) which shows a positive effect for cooperative learning on student achievement in these disciplines in comparison to individualistic or competitive learning. They also addressed various moderator variables of CL. One moderator variable that was addressed examined student performance in the different disciplines. For science, they reported an overall mean effect size of .42 from 9 included studies. Another moderator variable was methodology employed in the study. Studies in which the teacher was also the investigator reported significantly higher effect sizes ($d = .73$) than studies in which the teacher was not the investigator ($d = .41$). Studies that compared experimental and control groups showed a higher effect ($d = .57$) than studies that used a one-group, pretest/posttest design. Studies completed at 4-year institutions showed a larger effect ($d = .54$) than those done at 2-year institutions ($d = .21$), which was reported as non-significant. No difference among the different SMET disciplines was reported. No difference in achievement between males and females was evident. Groups of predominantly African American or Hispanic students exhibited significantly higher effects ($d = .76$) than predominantly White groups ($d = .46$) or heterogeneous groups ($d = .42$). No effect was seen for majors, nonmajors, or first year students. No significant difference was observed when comparing type and implementation of cooperative learning. Lastly, no significant difference was observed when comparing the amount of time students spent in cooperative activities. One

potential methodological issue regarding this meta-analysis was that they did not report specific literature search strategies, including the search terms or databases they used.

Bowen (2000) examined the effect of cooperative learning and chemistry achievement of both high school and undergraduate students. He reported a positive mean effect size as well ($d = .37$, based on 15 studies analyzed) but did not report on any moderator effects. Additionally, Bowen only searched specific journals and did not appear to do a more complete database search using keywords.

The Scott, et al. (2005) study reviewed the effectiveness of a variety of interventions, including cooperative learning. They reported a mean effect size of .95 based on 3 studies on cooperative learning, suggesting that cooperative learning has a positive effect on student achievement in science. However, the sample size of studies is very small, calling any conclusions into question. In addition, they did not use "cooperative learning" or any synonyms in the database which could be why they only used three studies in the analysis. Moreover, they did not do moderator analyses on each of the interventions that were studied, so none can be reported here.

No other meta-analyses specifically addressing the impact of cooperative learning on science achievement have been published since Bowen (2000), and he only focused on chemistry achievement. Scott, et al. (2005) reported effects of cooperative learning, but their study focused on other interventions as well and only cited 3 studies on cooperative learning. Therefore, there have been no meta-analyses focusing solely on the impact of cooperative learning on science achievement (including all science disciplines) since the Springer, et al. (1999) study and they only included nine studies that focused on science achievement.

Summary

The theoretical perspectives regarding cooperation and social interdependence have driven the development of cooperative learning strategies and research over the past few decades. An overwhelming number of empirical studies on cooperative learning support the notion that this instructional strategy has a positive effect on student achievement in comparison to traditional instructional methods, regardless of educational discipline or level. Meta-analytic research on the effectiveness of CL has also supported this notion. A number of moderator variables have also been analyzed with varying findings as reported above. The similarity of the results of original research and meta-analyses exhibit the consistency with which CL can increase student achievement.

Meta-analysis: Theoretical Background

The term "meta-analysis" was first coined by Glass (1976), based on his desire to distinguish primary analysis, the original analysis of data from a given research study, from meta-analysis, the reanalysis of data to answer new research questions. Although Glass was the first to use the term meta-analysis, this type of research technique actually began in the 1930s when a number of researchers combined statistical data from different studies (Petitti, 2000). The need for research techniques like meta-analysis increased in the 1970s in the social sciences when large numbers of studies on the same topic existed. The development of this technique was driven by the perception that narrative reviews of published studies were subjective in how they chose and weighted studies (Lipsey & Wilson, 2001; Hunter & Schmidt, 2004). Meta-analysis has been used extensively in the social sciences since the late 1970s when a number of researchers popularized the technique, further developed statistical methods for its application, and attempted to

systematically identify studies to be combined in a review. In addition, these same individuals made the estimation of effect sizes the primary aim of meta-analysis (Pettiti, 2000). An analysis of the PsycINFO database of articles dated 1974-2000 showed that meta-analysis appeared in zero titles in 1974 and then increased to over 850 titles in 2000 (Hunter & Schmidt, 2004). A similar trend occurred in medicine, with zero publications listing meta-analysis in the title in 1975 and over 1000 publications listing meta-analysis in the title in 1997, according to the MEDLINE database (Pettiti, 2000).

The quality and validity of the meta-analytic method has been widely reported (Lipsey & Wilson, 2001; Hunter & Schmidt, 2004; Pettiti, 2000). First, meta-analysis is considered one of the most effective methods by which to "summarize, integrate, and interpret selected sets of scholarly works..." (Lipsey & Wilson, 2001, p. 2). Second, meta-analysis presents key research findings in a manner which is more sophisticated than other types of review procedures that rely on qualitative summaries of statistical significance. Since effect sizes are calculated from the data reported in selected research studies, meta-analysis presents findings that are of different strengths across these studies. Third, the meta-analytic technique can find relationships or effects that are often blurred in qualitative or narrative summaries of similar research. The systematic coding and calculation of effect sizes from selected studies allows "an analytically precise examination of the relationships between study findings." Lastly, meta-analysis provides an organized technique for working with a potentially large number of studies in a review (Lipsey & Wilson, 2001).

The meta-analytic technique is useful in the development of theories related to a particular research area. In the social sciences, research studies often involve smaller

sample sizes, a fact that can make broad generalizations from their findings problematic. Meta-analysis of similar research studies allows the analyst to combine the findings from these studies, thus increasing the sample size from which statistical data can then be analyzed. Subsequent effect sizes from selected studies are calculated using a variety of formulas that use sample sizes, variance, and other factors to calculate a weighted effect size that corrects for heterogeneous sample sizes and statistical tests. The weighted effect sizes are then pooled and a mean effect size is calculated, typically with a confidence interval (Lipsey & Wilson, 2001). These statistical techniques allow the meta-analyst to calculate the most accurate effect size, from which relationships between variables in selected studies can be examined and theory regarding the research findings can be generated.

Criticisms of Meta-analysis

One of the primary criticisms of the meta-analytic technique is that is often seeks to combine dissimilar studies and form a conclusion based on these potentially very different studies (Petticrew & Roberts, 2006). This issue can be overcome by ensuring that studies that are pooled are similar enough in methodology, intervention, and outcome measurement. The meta-analytic technique relies primarily on a systematic review of reported research in a given area. Since most of the research that is published is typically statistically significant (Lipsey & Wilson, 2001; Hunter & Schmidt, 2004; Pettiti, 2000), a potentially large body of research in a given field may not be reported due to its lack of statistical significance. The primary issue in this case is that any effect sizes calculated only from published research studies could be inaccurate and skewed to show a greater effect than may actually exist (Lipsey & Wilson, 2001; Hunter & Schmidt, 2004). In

addition, the literature search procedures that are conducted in a systematic review of this kind often do not find all of the published and unpublished studies that might by available, often called "grey literature." The issue in this case is similar to that stated previously, that effect sizes may be skewed to show a greater effect than actually exists since results of a potentially significant number of studies are not included (Hopewell, McDonald, Clarke, & Egger, 2002). Despite these particular limitations of this type of study, most experienced meta-analysts report that these issues are part of the process and do not typically cause enough skew in effect sizes to warrant significant changes to the process (Lipsey & Wilson, 2001; Hunter & Schmidt, 2004). These same individuals recommend literature search techniques that can overcome, to some degree, the problem that any systematic review faces in terms of what studies are actually discovered and those that are not discovered in the review.

In recent years, the development of procedures which allow the synthesis of qualitative research has grown. While this methodology is not well-defined yet, there are a number of issues to consider. First, qualitative methodology differs greatly, thus making it difficult to categorize and therefore synthesize data. Second, qualitative research operates within a different analytic paradigm than quantitative research, again making synthesis of data problematic. Third, qualitative reviews may have to become much longer than quantitative reviews, which is a logistical issue in light of how time consuming quantitative reviews can be. Despite these particular issues, and as Petticrew and Roberts (2006, p. 191) state, "not everything that counts can be counted," qualitative data often provides the story to go along with the statistics which can assist in describing

the effect of a particular intervention which has been shown to have a significant effect in empirical studies.

The theoretical grounding of this study is based upon the viewpoints of three different organizations: The Institute of Education Sciences (IES), The Campbell Collaboration (C2), and The Cochrane Collaboration (CC). The Institute of Education Sciences is an office of the United States Department of Education whose mission is to "provide rigorous evidence on which to ground education policy and practice" (IES, 2006). The IES provides a hierarchy entitled the "Levels of Evidence for Assessing Program Effectiveness" (Cobb, personal communication, May 22, 2006) which ranks and grades different types of research methodologies. This hierarchy gives the systematic review a grade of "A" with regard to its effectiveness is assessing research findings, its highest rating in comparison to other types of research methodologies. C2 is an international organization composed of education experts whose "objectives are to prepare, maintain and disseminate systematic reviews of studies of interventions" in order to impact decision making about the effects of interventions in the social, educational, and behavioral fields (www.campbellcollaboration.org, 2006). The Cochrane Collaboration is the sibling organization of C2, with essentially the same objectives; however, its focus is on the development of systematic reviews of interventions in healthcare. All three of these organizations view the systematic review as the highest level of research that can be used to develop theories regarding interventions in various fields. The primary purpose of this study is to determine what effect cooperative learning has on student achievement in science disciplines. Based on the international acceptance

of the systematic review and meta-analytic methodologies and their usefulness, they will be used in this proposed study.

CHAPTER 3: METHOD

Research Design and Rationale

The research methodology to be employed in this study is of a quantitative design. The rationale for this approach is simple: a systematic review of the literature will be conducted and a meta-analysis will be performed in which effect sizes will be calculated using reported statistics from the selected studies. This research technique therefore necessitates a quantitative approach.

The purpose of this study is to conduct a systematic review and meta-analysis of the recent research into cooperative learning and its effect on student achievement in science at the secondary level and the first two years of post-secondary education (i.e., community or junior college). This study will involve an exhaustive review of published and non-published research in this area so that a meta-analysis can be performed in which effect sizes can be calculated based on the statistics reported in the selected studies. The effect size statistic can be analyzed so that a hypothesis regarding the actual effect of various curricular and instructional interventions can be developed and their impact assessed. In recent years, the effect size statistic has become more widely used as a manner with which to report the statistical effect of different types of treatments in educational settings (American Psychological Association, 2005; Morgan, Gliner, & Harmon, 2006). In addition, the American Psychological Association (APA) has

stipulated in their publication guidelines that all quantitative studies that report statistical results also report effect sizes (American Psychological Association, 2005). This change in policy by the APA is of great significance to meta-analysts, whose primary focus is to calculate mean effect sizes from research that reports various forms of statistical data. In many cases, meta-analysts are forced to take reported statistical values (such as means, *F*, *p*, and *t* values, to name a few) in the absence of reported effect sizes and then manipulate those values statistically so that effect sizes can be calculated. While this approach can be time consuming and difficult to complete, it is still one of the main methods employed by meta-analysts.

As mentioned in Chapter 2, the quality and validity of the meta-analytic method has been widely reported (Lipsey & Wilson, 2001; Hunter & Schmidt, 2004; Pettiti, 2000). First, meta-analysis is considered one of the most effective methods by which to "summarize, integrate, and interpret selected sets of scholarly works..." (Lipsey & Wilson, 2001, p. 2). Second, meta-analysis presents key research findings in a manner which is more sophisticated than other types of review procedures that rely on qualitative summaries of statistical significance. Since effect sizes are calculated from the data reported in selected research studies, meta-analysis presents findings that are of different strengths across these studies. Third, the meta-analytic technique can find relationships or effects that are often blurred in qualitative or narrative summaries of similar research. The systematic coding and calculation of effect sizes from selected studies allows "an analytically precise examination of the relationships between study findings" (p. 6). Lastly, meta-analysis provides an organized technique for working with a potentially large number of studies in a review (Lipsey & Wilson, 2001).

The meta-analytic technique is useful in the development of theories related to a particular research area. In the social sciences, research studies often involve smaller sample sizes, a fact that can make broad generalizations from their findings problematic. Meta-analysis of similar research studies allows the analyst to combine the findings from these studies, thus increasing the sample size from which statistical data can then be analyzed. Subsequent effect sizes from selected studies are calculated using a variety of formulas that use sample sizes, variance, and other factors to calculate a weighted effect size that corrects for heterogeneous sample sizes and statistical tests (details of these calculations are reported in the Data Analysis and Form of Results section). The weighted effect sizes are then pooled and a mean effect size is calculated, typically with a confidence interval (Lipsey & Wilson, 2001). These statistical techniques allow the meta-analyst to calculate the most accurate effect size, from which relationships between variables in selected studies can be examined and theory regarding the research findings can be generated.

This research project would add to this field of study by providing an exhaustive summary of the type of research that has been conducted and the effect of cooperative learning on student achievement in science in this area since 1995. A summary of this kind would provide a functional description of CL techniques that educators could use as a reference for changing their curriculum and instructional techniques so that student achievement in science may be affected. In addition, a summary of this kind could be utilized by professional developers as a basis from which they could design and deliver professional development workshops and institutes for teachers with whom they work. The instructional activities discovered in this research project could be demonstrated in

various types of professional development activities and provide many teachers with the tools to work toward impacting student achievement in science.

Meta-analysis, by nature, is very easy to replicate. The methods that are utilized in this technique are widely published and available to educators. In addition, all of the research studies that are selected for analysis in a review such as this are listed in the references section so that anyone can obtain them and repeat the analysis.

Theoretical Population, Sampling Frame, and Final Sample

Theoretical Population

In a review such as this, it is important to identify the theoretical population to which the results can be generalized. The theoretical population of this study will be identified by the following characteristics, and are based on the inclusion criteria that will be described subsequently. Depending on the country of origin of studies, the results will most likely generalize to the Western educational culture, since only studies published in English are being included. However, if multiple studies from non-Western educational systems are included in the final analysis, they will be grouped as such and analyzed accordingly. Secondly, since the educational level criterion is secondary and early post-secondary students, the results may be generalized to all students at these educational levels. Thirdly, the results of this review could apply to any cooperative learning strategy that has been described whether by name or simply by description. Lastly, since only science disciplines are considered in this review, the results will generalize to any science discipline and the use of cooperative learning as an instructional strategy.

Sampling Frame

In meta-analyses, the sampling frame (or participants) in the review is actually published or unpublished studies that have been identified through database and other types of library searching. In order for studies to be included in the final data extraction for calculation of effect sizes, they must meet specific criteria described in a coding protocol. These inclusion criteria are developed *a priori* to the review and revised as needed after a few studies are coded in a "practice run" using the protocol. The primary criteria for inclusion are as follows:

1. Cooperative learning is identified as the intervention
2. The research design of the study is either randomized controlled trial (RCT) or quasi-experimental
3. A science discipline is identified as part of the setting of the study
4. Student achievement in science is identified as the outcome
5. The date of publication of the study falls between 1995-2007
6. The subjects are identified as secondary (middle, junior high, or senior high) or early post-secondary students (community college or first two years of university study)
7. The study is published in English
8. The study provides enough data for an effect size statistic to be calculated

Data Collection

The data in a systematic review and meta-analysis can be described in multiple ways. The first "data set" includes the articles that are ultimately selected for the meta-analysis through coding studies based on specific criteria found in a search of available

literature. This is the first method that will be described. The first step in a systematic review is an online database search for relevant research studies. Although many authors suggest searching a large number of online databases, five were chosen for this review based on their inclusion of various types of social science studies and the advice of the author's graduate adviser (R.B. Cobb), Merinda McClure, Research Librarian at Colorado State University, and Anne Wade, Manager/Information Specialist for the Centre for the Study of Learning and Performance at Concordia University and Information Retrieval Specialist for the Campbell Collaboration. The databases utilized in this search were: ERIC/OCLC, PsycINFO, Digital Dissertations (ProQuest), Web of Science, and Education Abstracts. These databases are most commonly used in social science research.

A total of seven searches using varying combinations of keywords and Boolean logic terms were performed. The year range for the searches was 1995-2007 and the language chosen was English (if provided as an option).

The keywording and Boolean logic used for the searches that were conducted included the following:

1. "collaborative learning" OR "cooperative learning" OR "group activities" AND "science"
2. "collaborative learning" OR "cooperative learning" OR "group activities" AND "science" AND "higher education" OR "secondary education" OR "high schools"
3. "collaborative learning" OR "cooperative learning" OR "group activities" AND "science" AND "higher education" OR "secondary education" OR "high schools" OR "middle schools"

4. "collaborative learning" OR "cooperative learning" OR "group activities" AND "science" AND "higher education" OR "secondary education" OR "high schools" OR "middle schools" OR "middle school students" OR "high school students"
5. "collaborative learning" OR "cooperative learning" OR "group activities" AND "science" OR "science education" AND "higher education" OR "secondary education" OR "high schools" OR "middle schools" OR "middle school students" OR "high school students"
6. "collaborative learning" OR "cooperative learning" OR "group activities" AND "science" OR "science education" OR "high school science" OR "middle school science" AND "higher education" OR "secondary education" OR "high schools" OR "middle schools" OR "middle school students" OR "high school students"
7. "collaborative learning" OR "cooperative learning" OR "group activities" AND "science" OR "science education" OR "high school science" OR "middle school science" OR "college science" AND "higher education" OR "secondary education" OR "high schools" OR "middle schools" OR "middle school students" OR "high school students"

After the literature search was completed, all of the citations, titles, and abstracts were downloaded into EndNote (v. 9) software for further analysis.

Coding of the studies for germaneness to the review is the next process in a meta-analysis, and includes the following processes. The first step in coding citations obtained in the database search is to code the titles and abstracts. A protocol for this part of the process was developed using an adaptation of the inclusion criteria mentioned previously (see Appendix A). The criteria for the title and abstract coding are more general in nature

in order to ensure that as many studies as possible may be included. Moreover, in the event that any of the criteria were difficult to assess in a dichotomous fashion, a code of "Yes" was assigned for that study, which is included. If any criteria were assigned a code of "No," the study was excluded. In order to avoid bias on the part of the author, it is necessary to have at least one other individual code the studies found in the database search. In addition to the author, one other coder (Barbara Patterson, doctoral candidate, School of Education, Colorado State University) was chosen to participate in this portion of the process. This second coder was trained in the use of the title and abstract coding protocol prior to completing any abstract coding. The rationale behind multiple coders is well-documented (Lipsey & Wilson, 2001; Pettiti, 2000). The interrater reliability of the coding protocol can be assessed only if more than one individual codes the studies found in the literature search. A random sample of fifty studies was selected from the database search and each coder coded them using the title and abstract coding protocol. Percent agreement statistics were calculated to determine the reliability of the coding protocol. In the event that one hundred percent agreement is not achieved, the coders will meet to discuss any discrepancies regarding inclusion or exclusion of studies and come to consensus regarding these studies. Once the percent agreement statistics and consensus meeting have been completed, the author will code the titles and abstracts of the remaining studies.

The second step in the coding process is to analyze the studies that remained after the title and abstract coding process. These remaining studies were obtained in full-text from the Colorado State University library directly, or through Interlibrary Loan. These studies were then coded using an intermediate screening protocol, which is a more

detailed version of the title abstract coding protocol. The intermediate protocol (see Appendix B) has more defined criteria that a study must meet in order to progress to the final full-text screening and data extraction. Intermediate coding was completed by the author.

The third step in the coding process is to analyze the studies that remained after the intermediate screening. These studies will be further coded using the final full-text coding protocol (see Appendix C) to determine their inclusion in the meta-analysis. The final full-text coding protocol adds more detail to the intermediate screening protocol so that all inclusion criteria can be assessed and studies can be kept for data extraction or eliminated from further analysis. This part of the study coding was completed by the author and Barbara Patterson. Reliability statistics (percent agreement) will be calculated as described for the title and abstract coding, and any discrepancies will be discussed between the coders.

Studies that are included in the final meta-analysis will go through the described tri-level coding process and will then go through data extraction. The title and abstract coding process essentially excludes studies that do not come close to meeting the general criteria for inclusion. The intermediate coding process identifies studies that meet more specific criteria but may have some shortcomings that could lead to exclusion. The final full-text coding process is the most specific and excludes studies based on the criteria listed in the full-text screening protocol.

All included studies will be tabulated, listing the country in which the study was conducted, the participant characteristics, and the study characteristics.

Once the final pool of studies is identified, data extraction occurred. After data was extracted from each study, it was entered into a database created in Comprehensive Meta-Analysis® (CMA) software for effect size calculation and further analysis. Data may be reported in many different formats in different studies. Fortunately, CMA allows entry of multiple types of data for effect size calculation.

Measures

The dependent measure that is being analyzed in this review is student achievement in science disciplines. This must be the outcome measurement in any of the studies that are identified for inclusion in the meta-analysis. This criterion is listed in the final full-text screening protocol. While many studies may identify student achievement in science as an outcome, this criterion must be analyzed in detail. If an assessment tool is used, the study must identify the assessment tool by which student achievement was measured. The final screening protocol lists two types of assessment tools that meet inclusionary criteria. First, standardized (nationally-developed or locally-developed) assessment tools that measure student science achievement reliably measure student achievement in science since they typically go through rigorous reliability and validity testing. Second, teacher-constructed assessments that are statistically analyzed or expert-reviewed for reliability and validity are considered to reliably measure student achievement in science. If a study in which a teacher-constructed assessment tool is used does not describe statistical reliability and validity analysis, the study will still be included, but this fact will be noted for potential further analysis. Additionally, student grades in a course will also be included as outcome measures.

Moderator Analyses

A number of independent measures were analyzed in this review provided the studies that are included report disaggregated information. A number of characteristics of the participants in included studies were analyzed and are as follows:

1. Gender
2. Ethnicity
3. Socioeconomic status
4. Educational level
5. Student ability level
6. Language proficiency
7. Disability

Research methodology of included studies will also be analyzed. For this analysis, the assignment of groups to conditions will be considered (randomized-controlled trials or quasi-experimental design). Science discipline will be analyzed as an independent measure, as will the type of CL intervention, as described previously. The country in which the study was conducted will be analyzed also. The last independent measure that will be analyzed is whether the researcher(s) reported reliability testing on the assessment instrument used to measure student science achievement.

Data Analysis

After the final pool of studies is selected, statistical analysis of their findings will be conducted. The primary statistic of value to the meta-analyst is the effect size. Three primary types of effect sizes can be calculated from selected studies depending on the methodology in a given study (including one-variable, two-variable and multiple variable

relationships) and the statistics reported. These different effect size statistics are the mean difference (unstandardized or standardized), the correlation coefficient, and the odds-ratio.

For studies that present results that compare two groups on measures that have a continuous underlying distribution (Lipsey & Wilson, 2001), two types of the mean difference effect size statistics can be calculated. The two groups in a given study may be treatment and control, active independent variables, or naturally occurring groups such as gender, an attribute independent variable. If the same measurement procedures and numerical scale for the dependent variable across samples are used in a given study, the unstandardized mean difference effect size can be used. If the operationalization of the dependent variable differs across samples in a study, the standardized mean difference effect size is used. This particular effect size statistic is more commonly used. Calculation of these particular effect size statistics require descriptive statistics from a given study, mean values and standard deviations of a given measurement, and sample size. In studies that do not report means but report *t*-values, *F*-ratios, gain score estimates of the mean difference, covariate adjusted estimates of the mean difference, or regression coefficient estimates of the mean difference, algebraic formulas can be used to calculate the standardized mean difference effect size statistic. The same can be done for studies that do not report the standard deviations of the means (Lipsey & Wilson, 2001). All algebraic equations for the above statistical values can be found in Lipsey and Wilson (2001), but for the sake of brevity, will not be reported here. For the two above effect size statistics, the strength and direction of the effect is interpreted using Cohen's (1988) suggested rules, with higher positive values indicating a stronger effect.

The second effect size statistic of interest to meta-analysts is the correlation coefficient. Many studies report data directly using this statistic, which eases the coding process. Correlation coefficients are typically reported in studies in which two variables are being measured, with both on a continuous scale. The typical correlation statistic reported in many studies is the Pearson product-moment correlation coefficient (r), in which the standard formula is used in the calculation. For studies describing correlations between variables that do not report r, but report joint frequency distributions, means and standard deviations, results of independent t-tests, results of one-way analysis of variance (F-ratios), results of chi-square tests, various formulas exist that allow the calculation of r using these values. Values for this effect size statistic range from -1 to +1, with a stronger relationship indicated by values further from zero (Lipsey & Wilson, 2001).

The third type of effect size statistic is the odds-ratio. This statistic compares two groups in terms of the relative odds of an event and is applicable to study findings that use dichotomous variables presented in proportions or frequencies. This statistic is based on the frequencies found in a 2 X 2 contingency table, constructed from data in the study of interest. Interpretation of this effect size is similar to mean difference effect sizes, with larger values indicating stronger relationships between variables (Lipsey & Wilson, 2001)

The effect size that will be calculated in this study is the standardized mean difference effect size since one of the inclusion criteria states that only studies that compare two groups (control and treatment) will be considered. The specific effect size statistic that will be calculated is Hedges's g (Hedges, 1981), which is a type of standardized mean difference effect size. This particular statistic considers the sample

size of both the control and treatment groups for a data set and weights the effect size based on these sample sizes. Since sample sizes in included studies may vary widely, this particular statistic improves the fidelity of the analysis.

After the standardized mean difference effect sizes are calculated for each study, a forest plot of effect sizes calculated from included studies will be generated to show a median value, confidence intervals, and any outliers. If extreme outliers are found, they may be discarded from the data set or considered in moderator analysis. An overall mean effect size of all the studies will then be calculated with a confidence interval. The mean effect size is computed by weighting each of the individual study effect sizes by the inverse of their variances, and is calculated as the sum of the individual study effect sizes multiplied by their inverse variances. This value is then divided by the sum of all of the inverse variances. The confidence interval for the mean effect size indicates the range within which the population mean is likely to fall for the observed data. The confidence interval is helpful in indicating the precision of the estimated mean effect size. If the confidence interval includes zero, the mean effect size is likely not statistically significant. The confidence interval can be calculated using mean effect size, its standard error, and a *z*-critical value associated with the level of statistical significance of interest to the meta-analyst (i.e., $p \leq .05$ or $p \leq .01$) (Lipsey & Wilson, 2001).

A variety of issues with regard to calculation and analysis of the mean effect size statistic must be addressed. For example, effect sizes from selected studies should be statistically independent. If a single study presents findings using subsets of a sample, then individual effect sizes should be calculated for each subsample, and not from the entire sample. In addition, an analysis of the homogeneity of the effect size distribution

should be done to determine if the analysis will proceed under a fixed effects model or a random effects model. The Q statistic is calculated when assessing the homogeneity of the mean effect size. If the Q statistic reveals a homogeneous distribution, then the analysis of the mean effect size will proceed using a fixed effects model. This model assumes that any effect size observed in a study estimates the population effect with random error that stems from chance, subject-level sampling error. Under this model, no further adjustment or weighting of the mean effect size is necessary. If the Q statistic is statistically significant and reveals a heterogeneous distribution, then the analyst has some choices as to how to proceed. One direction the analyst may take is to proceed using a random effects model. However, the statistical significance of the Q statistic should not drive the decision on which effect size calculation model to utilize. If the researcher determines that the studies included in the meta-analysis do not share a common effect size, then the random effects model should be used (Borenstein, Hedges, Higgins, & Rothstein, 2009). This model assumes that in order to represent the variation among study effect sizes, another random error component must be included. In terms of calculation of the mean effect size, another variance component, one that sums subject-level sampling error and random variability, must be used in the calculation of the mean effect size. Another choice is to proceed under a fixed effects model as mentioned previously (Lipsey & Wilson, 2001). The meta-analyst should report which model is being utilized in the analysis so that the reader has a context within which to understand any findings that are presented. Another statistic that measures the heterogeneity of effect sizes is the I^2 statistic (Higgins & Thompson, 2002). If the Q statistic and I^2 statistic reveal a heterogeneous distribution and it is deemed that the effect sizes from all

the studies do not represent a single effect size, moderator analyses will be conducted using the independent measures criteria described previously. These moderator analyses will involve performing an analysis of variance (ANOVA) on the effect size statistics to determine if they are statistically significantly different from each other. All effect size calculation and statistical analysis will be performed using Comprehensive Meta-Analysis® software.

Other analyses will be conducted on the effect size data. First, "publication bias" or "the file drawer effect," (Iyenger & Greenhouse, 1988) will be analyzed. This phenomenon states that the results of a meta-analysis may be skewed in the positive direction due to the inability to locate all studies related to a particular treatment, since most studies included in meta-analyses come from the published literature. Funnel plots will be created as part of this analysis. The degree of publication bias can be inferred from funnel plots. Another analysis of this phenomenon called the "trim and fill" method (Duval & Tweedie, 2000), which estimates and adjusts the overall mean effect size for the numbers and outcomes of potential missing studies which are imputed in the analysis. For this analysis, funnel plots, which show the distribution of effect sizes around the mean and group them based on their sample size, will be created using CMA. Additionally, the "fail-safe N" statistic (Cooper & Hedges, 1994) will be calculated, providing deeper analysis of publication bias. Second, the "one study removed" analysis (Comprehensive Meta-analysis, 2007) will be conducted. This analysis calculates the overall mean effect size when each of the individual effect sizes is removed from the original calculation to determine if any one particular study has a greater effect on the

overall mean effect size. Results of this analysis will be discussed after it is conducted on CMA.

Once all of the calculations of individual effect sizes and the overall mean effect size are completed (including their confidence intervals), these values will be presented in tabular format. All effect sizes and the data used to calculate them will be presented together. This type of presentation is common in meta-analysis. Any graphs that were generated during the data analysis will be displayed in APA approved format.

CHAPTER 4: RESULTS

Introduction

The purpose of this study was to determine the effects of cooperative learning instructional strategies on student achievement in science. The methodology that was employed to accomplish this goal was meta-analysis, in which a systematic review of available literature was conducted. The statistical measure used in meta-analysis is the effect size, which calculates the overall effect of an intervention by comparing outcome data between a control group and a treatment group. The final set of studies from which data were extracted was compiled through a three-level series of screening in which the criteria got more specific after each level. Data from each of these studies was extracted, with individual effect sizes calculated for each set of data. An overall mean effect size was calculated that represented the effect that cooperative learning instructional strategies had on student achievement in science in comparison to traditional instruction or instruction as usual, both of which can also be defined as individual instruction. Moreover, moderator analyses were performed to determine if differential effects were apparent based on particular characteristics of the subjects (ethnicity or gender), intervention, study methodology, country of origin of a study, or reliability testing of outcome measures.

Database Searches

The electronic database searches described in Chapter 3 yielded a total of 2,506 references, after duplicates were assessed and deleted. The references were downloaded into EndNote® (version 9.0) for sorting and title and abstract screening. Additionally, a similar search completed by the Manager/Information Specialist for the Centre for the Study of Learning and Performance at Concordia University and Information Retrieval Specialist for the Campbell Collaboration, yielded 62 references. The titles and abstracts of these references were coded, and 24 were excluded for various reasons, the most common reasons being that the date of publication of the reference was prior to 1995 or the subjects in the study were not secondary or early post-secondary students. The remaining 38 references were cross-referenced with the existing database. Thirty-six of these studies were found to be duplicates, and two that were new were obtained through Interlibrary Loan at Colorado State University. These remaining references were subsequently excluded because they were only available in Korean, and not English. All of the titles and abstracts that comprised the final set were screened using the Title and Abstract Coding Form (see Appendix A).

Title and Abstract Coding

In order to assess the reliability of the Title and Abstract Coding form, each coder reviewed the same random sample of 50 abstracts taken from the database search. Initial coding yielded 92% agreement (46 out of 50 studies were coded the same) between the coders. Ideally, 100% agreement between coders should be attained, and since that did not occur, the coders discussed any discrepancies on particular studies and came to consensus on which studies to include and exclude. The final title and abstract coding

was completed by the author and yielded 718 studies for the next phase of coding. This resulted in a 71.3% reduction in the number of references from the original pool. Reasons for exclusion of abstracts were varied and included the following: 1) the abstract did not describe an empirical research study, 2) the outcome was not science achievement, 3) the intervention was not cooperative learning, 4) the study was qualitative in design, 5) the study participants were not secondary or post-secondary students, and 6) the academic discipline in which the study was performed was not science. It is important to note that titles and abstracts often do not provide enough information to fully code for inclusion or exclusion. If the specific characteristics of a study that were coded in the title and abstract coding could not be ascertained, the study was included. Full-text versions of the 718 studies that resulted from the title and abstract screening were obtained through the Colorado State University Library. Most of the studies were obtained as electronic copies while others were obtained as hardcopy or microfilm through interlibrary loan.

Full-text Intermediate Coding

Full-text versions of the 718 studies that resulted from the title and abstract screening were obtained through the Colorado State University Library. These studies were screened by the author using the Full-Text Intermediate Screening Protocol (see Appendix B). This step in the screening process yielded 81 studies for the final full-text coding, which represented 3.2% of the original 2,506 abstracts from the database search. Reasons for exclusion were varied and included the following: 1) the article was not an empirical research study, 2) the research methodology focused on a single group of

participants and did not include a control group, 3) cooperative learning was not the intervention, and 4) science achievement was not the outcome in the study.

Final Full-text Coding

The final full-text coding was completed using the Full-text Study Coding form (see Appendix C). This step in the coding was completed by two coders (the primary researcher and an advanced doctoral graduate assistant). Once all studies were coded, the coders communicated which studies each decided should be included in the data extraction for the meta-analysis. The percent agreement value that was calculated was 79%. The coders discussed any discrepancies regarding studies to be included or excluded and came to consensus on those where conflict was evident. The most frequent discrepancy involved whether a study identified one or multiple interventions being utilized. For example, some studies employed the use of computers in addition to the CL intervention in the treatment group, while the control group was not exposed to either the use of computers or the CL intervention. During the consensus discussions, the coders reviewed the studies in question together in order to specifically identify if a study employed multiple interventions or not. Another discrepancy that occurred was whether the intervention in a study was truly CL. In these studies, the CL intervention was typically not identified specifically. In order to determine inclusion or exclusion, the coders reviewed the study together and discussed whether the intervention qualified as CL or not. In both situations, once a decision was made, the study was either included or excluded.

Characteristics of Included Studies

A total of 32 studies, which yielded 52 individual effect sizes, were included in the meta-analysis after the final full-text coding was completed. Table 1 lists the authors, the country in which the study took place, and the participant characteristics. As can be seen in this table, 17 of the studies took place in countries outside the United States, while 15 studies were conducted in the United States. The grade levels of the participants ranged from 6-14. Only 4 studies disaggregated achievement data based on the gender of the participants. Studies conducted in biology and chemistry disciplines were the most numerous (12 each) with earth science (4 studies), physics (3 studies) and general science (1 study) showing many fewer studies. Table 2 describes the methodological characteristics of each of the included studies, the CL intervention, and the outcome measure used to assess student achievement. Eleven studies utilized a cluster randomized methodology, six studies utilized a quasi-experimental with subject matching methodology, and fifteen utilized a quasi-experimental without subject matching methodology. Seventeen studies did not identify a specific type of CL intervention as identified in the meta-analysis reported by Johnson, et al. (2000). These studies were scrutinized extensively to ensure that the intervention was not specifically identified. They were included in the final analysis as cooperative learning because they described the intervention in enough detail that it could be defined as such, with students working in organized group toward a common educational goal. The remaining 15 studies identified a specific CL intervention according to Johnson, et al. (2000) (11 studies), or as another type of cooperative learning structure (4 studies). The outcome measures utilized in each study varied, with 9 studies employing a standardized instrument, and 22 studies

employing a teacher constructed instrument. Only one study utilized a course grade as the only outcome measure, while one study each utilized both course grades and a teacher constructed instrument or a standardized instrument, and one study utilized both teacher constructed and standardized instruments. Data were extracted from the studies and entered into Comprehensive Meta-Analysis™ (CMA) software for effect size calculation.

Table 1

Sample and Participant Characteristics of All Studies

Study	Country	Participant Characteristics		
		Grade Level	% Male	Science Content Area
Abobaker, N.M. (1995)	Yemen	11	48%	Chemistry
Acar, B. & Tarhan, L. (2006)	Turkey	11	NR	Chemistry
Balfakih, N.M.A. (2003)	UAE	10	51%	Chemistry
Banerjee, A. (1997)	India	13	NR	Chemistry
Bilgin, I. (2006)	Turkey	13	NR	Chemistry
Bradley, A.Z., Ulrich, S.M., Jones, Jr., M., & Jones, S.M. (2002)	United States	13-14	NR	Chemistry
Chang, C. & Mao, S. (1999) #1	Taiwan	9	NR	Earth Science
Chang, C. & Mao, S. (1999) #2	Taiwan	9	NR	Earth Science
Chung-Schickler, G.C. (1998)	United States	13-14	45.7%	Biology
De Baz, T. (2001)	Jordan	7	0%	Biology
Dori, Y.J., Yeroslavski, O., & Lazarowitz, R. (1995)	Israel	8	NR	Biology

Table 1 (cont'd)

Sample and Participant Characteristics of All Studies

Study	Country	Participant Characteristics		
		Grade Level	% Male	Science Content Area
Faro, S. & Swan, K. (2006)	United States	9-11	50%	Earth Science
Foley. K. & O'Donnell, A. (2002)	United States	9-11	55.2%	Chemistry
Fontenot, D. W. (1995)	United States	7	NR	Biology
Gayford, C. (1995)	United Kingdom	11	NR	Biology
Hanze, M. & Berger, R.(2007)	Germany	12	NR	Physics
Harskamp, E. & Ding, N. (2006)	China	11	45.5%	Physics
Jensen, M.S. (1996)	United States	13	NR	Biology
Jeon, K., Huffman, D., & Noh, T. (2005)	Taiwan	11	100%	Chemistry
Lazarowitz, R. (1996)	Israel	10	30.7%[t]; 52.5%[c]	Biology

Table 1 (cont'd)

Sample and Participant Characteristics of All Studies

Study	Country	Participant Characteristics		
		Grade Level	% Male	Science Content Area
Lumpe, A. (1995)	United States	HS	48%	Biology
Qualter, A. & Abu-hola, I.R.A. (2000)	Jordan	6, 9, 10	48%	General Science
Roy, H. (2003)	United States	14	NR	Biology
Sadler, K.C. (2002)	United States	13-14	51%	Biology
Schroeder, P.G. (1996)	United States	13-14	NR	Chemistry
Shachar, H. & Fischer, S. (2004)	Israel	11	NR	Chemistry
Snyder, T. & Sullivan, H. (1995)	United States	7	NR	Biology
Starr, E.M. (1995)	United States	13-14	11.1%[t]; 0%[c]	Earth Science
Tao, P. (1999)	Hong Kong	12	NR	Physics

Table 1 (cont'd)

Sample and Participant Characteristics of All Studies

Study	Country	Participant Characteristics		
		Grade Level	% Male	Science Content Area
Trautwein, S.N., Racke, A., & Hillman, B. (1997)	United States	13-14	NR	Biology
Verdel, E.F.O. (1996)	United States	College	40%	Chemistry
Werner, J.L. & Klein, J.D. (1999)	United States	9	NR	Chemistry

NR = Not reported
t = treatment group
c = control group

Table 2

Research Designs, Interventions, and Outcome Measures for All Studies

Study	Research Design	Cooperative Learning Intervention	Outcome Measure(s)
Abobaker, N.M. (1995)	Cluster Randomized	STAD	Standardized/Commercial Instrument
Acar, B. & Tarhan, L. (2006)	Quasi-experimental without subject matching	NS	Teacher Constructed Instrument
Balfakih, N.M.A. (2003)	Cluster Randomized	STAD	Teacher Constructed Instrument
Banerjee, A. (1997)	Quasi-experimental without subject matching	NS	Teacher Constructed Instrument (2)
Bilgin, I. (2006)	Cluster Randomized	NS	Standardized/Commercial Instrument

Table 2 (cont'd)

Research Designs, Interventions, and Outcome Measures for All Studies

Study	Research Design	Cooperative Learning Intervention	Outcome Measure(s)
Bradley, A.Z., Ulrich, S.M., Jones, Jr., M., & Jones, S.M. (2002)	Quasi-experimental with Subject Matching	NS	Course Grade (2); Teacher Constructed Instrument
Chang, C. & Mao, S. (1999) #1	Quasi-experimental without subject matching	Inquiry Groups	Teacher Constructed Instrument
Chang, C. & Mao, S. (1999) #2	Quasi-experimental without subject matching	GI	Standardized/Commercial Instrument
Chung-Schickler, G.C. (1998)	Quasi-experimental with Subject Matching	Learning Together and Alone/STAD	Teacher Constructed Instrument

Table 2 (cont'd)

Research Designs, Interventions, and Outcome Measures for All Studies

Study	Research Design	Cooperative Learning Intervention	Outcome Measure(s)
De Baz, T. (2001)	Quasi-experimental with Subject Matching	Jigsaw	Teacher Constructed Instrument
Dori, Y.J., Yeroslavski, O., & Lazarowitz, R. (1995)	Quasi-experimental without subject matching	Jigsaw	Teacher Constructed Instrument
Faro, S. & Swan, K. (2006)	Quasi-experimental without subject matching	STAD	Standardized/Commercial Instrument
Foley. K. & O'Donnell, A. (2002)	Cluster Randomized	NS	Standardized/Commercial Instrument (2)
Fontenot, D.W. (1995)	Quasi-experimental without subject matching	NS	Teacher Constructed Instrument

Table 2 (cont'd)

Research Designs, Interventions, and Outcome Measures for All Studies

Study	Research Design	Cooperative Learning Intervention	Outcome Measure(s)
Gayford, C. (1995)	Quasi-experimental without subject matching	NS	Teacher Constructed Instrument
Hanze, M. & Berger, R.(2007)	Cluster Randomized	Jigsaw	Teacher Constructed Instrument
Harskamp, E. & Ding, N. (2006)	Cluster Randomized	NS	Teacher Constructed Instrument
Jensen, M.S. (1996)	Quasi-experimental without subject matching	NS	Teacher Constructed Instrument
Jeon, K., Huffman, D., & Noh, T. (2005)	Cluster Randomized	TAPPS*	Teacher Constructed Instrument

Table 2 (cont'd)

Research Designs, Interventions, and Outcome Measures for All Studies

Study	Research Design	Cooperative Learning Intervention	Outcome Measure(s)
Lazarowitz, R. (1996)	Quasi-experimental without subject matching	Jigsaw	Teacher Constructed Instrument
Lumpe, A. (1995)	Cluster Randomized	NS	Standardized/Commercial Instrument; Teacher Constructed Instrument
Qualter, A. & Abu-hola, I.R.A. (2000)	Cluster Randomized	NS	Teacher Constructed Instrument
Roy, H. (2003)	Quasi-experimental without subject matching	NS	Teacher Constructed Instrument
Sadler, K.C. (2002)	Quasi-experimental without subject matching	NS	Standardized/Commercial Instrument; Course grade

Table 2 (cont'd)

Research Designs, Interventions, and Outcome Measures for All Studies

Study	Research Design	Cooperative Learning Intervention	Outcome Measure(s)
Schroeder, P.G. (1996)	Quasi-experimental with Subject Matching	Dyads	Standardized/Commercial Instrument (2)
Shachar, H. & Fischer, S. (2004)	Quasi-experimental without subject matching	GI	Teacher Constructed Instrument
Snyder, T. & Sullivan, H. (1995)	Quasi-experimental with Subject Matching	NS	Teacher Constructed Instrument
Starr, E.M. (1995)	Quasi-experimental without subject matching	Learning Together and Alone	Teacher Constructed Instrument (4)
Tao, P. (1999)	Quasi-experimental with Subject Matching	Dyads	Teacher Constructed Instrument

Table 2 (cont'd)

Research Designs, Interventions, and Outcome Measures for All Studies

Study	Research Design	Cooperative Learning Intervention	Outcome Measure(s)
Trautwein, S.N., Racke, A., & Hillman, B. (1997)	Quasi-experimental without subject matching	NS	Course Grade
Verdel, E.F.O. (1996)	Cluster Randomized	NS	Standardized/Commercial Instrument
Werner, J.L. & Klein, J.D. (1999)	Cluster Randomized	NS	Teacher Constructed Instrument (2)

NS = Not specified
*Thinking aloud pair problem solving

Data Analysis

Heterogeneity of Included Studies

The heterogeneity or homogeneity of studies included in a meta-analysis relates to two types of study variability. Within-study variability refers to the variability due to sampling error since the studies included all have a different sample size of participants, as well as other issues such as an unreliable outcome measure, differences in individual or group accountability assessment of the students by the teacher, or the potential lack of fidelity of the implementation of the treatment by teachers in different classrooms. In addition, differences between students that may not have been accounted for prior to the study such as prior knowledge of the science content or the ability level of the students could skew student achievement data and increase within-study variability. Between-study variability is due to the influence of a variety of characteristics that vary among the included studies, such as characteristics of the samples, differences in the treatment, and design quality (Huedo-Medina, Sánchez-Meca, & Marín-Martínez, 2006). The Q-statistic is a measure of this study variability. If the Q-statistic is not statistically significant, the study sample may be considered to be homogeneous, a fixed-effects model of effect size calculation is used, and it can be assumed that the effect sizes differ only by sampling error and represent a single effect size. If the Q-statistic is statistically significant, the study sample may be considered to be heterogeneous, and a random effects model of effect size calculation is used which includes both within-study and between-study variability (Huedo-Medina, Sánchez-Meca, & Marín-Martínez, 2006). The Q-statistic value was 536.378 for all included studies and was statistically significant ($p<.0001$).

The I^2 statistic measures the extent of the true heterogeneity of studies in a meta-analysis (interpreted as a percentage), not just whether the sample of studies is homogeneous or not. The I^2 statistic value for the studies included in this meta-analysis was 90.492, which was statistically significant ($p<.0001$), indicating that the sample of studies was heterogeneous. This finding further supports the use of a random effects model of analysis. This I^2 value means that 90% of the heterogeneity between the studies is true heterogeneity and not due to sampling error (Huedo-Medina, Sánchez-Meca, & Marín-Martínez, 2006).

Effect Sizes

The results of the heterogeneity analysis in a meta-analysis may impact the particular model by which effect size calculation takes place. When the Q-statistic is statistically significant, as it was in this case, the random effects model of effect size calculation is typically used. However, as mentioned previously, this statistic does not necessarily force the used of this model (Borenstein, Hedges, Higgins, & Rothstein, 2009). Based on the varying characteristics of the studies as seen in Tables 1 and 2, it can be concluded that the pool of studies does not represent a single effect. Consequently, a random effects model of effect size calculation was used to calculate the effect size in each study. Table 3 lists each study, data used for effect size calculation, each effect size with a corresponding confidence interval, and the overall mean effect size. The data extracted from each study varied widely, with 15 sets of pre-test and post-test data, 24 sets of post-test data only, and a wide variety of other types of data including final grades, pre-test to post-test differences, and F, t, and p-values.

Table 3

Meta-Analytic Results of All Studies

Study	Statistic(s) Reported	Effect Size	Confidence Interval	
		Hedges's *g*	Lower	Upper
Abobaker, N.M. (1995)	Pre- and post-test *M*, *SD*, *n*	.584 (males)	.164	1.004
Abobaker, N.M. (1995)	Pre- and post-test *M*, *SD*, *n*	-.193 (females)	-.601	.214
Acar, B. & Tarhan, L. (2006)	Pre- and post-test *M*, *SD*, *n*	2.411	1.616	3.207
Balfakih, N.M.A. (2003)	Post-test, *p*-value	.508 (females)	.253	.764
Balfakih, N.M.A. (2003)	Post-test, *p*-value	.503 (males)	.250	.757
Banerjee, A. (1997)	Post-test *M, SD*, *n*	.015 (females)	-.651	.681
Banerjee, A. (1997)	Post-test *M, SD*, *n*	-.425(males)	-1.080	.231
Bilgin, I. (2006)	Post-test *M, SD*, *n*	1.050 (2 outcomes combined)	.610	1.490
Bradley, A.Z., Ulrich, S.M., Jones, Jr., M., & Jones, S.M. (2002)	*F*-value (MANOVA; difference between 4 instruments)	.149	-.201	.499

Table 3 (cont'd)

Meta-Analytic Results of All Studies

Study	Statistic(s) Reported	Effect Size	Confidence Interval	
		Hedges's *g*	Lower	Upper
Bradley, A.Z., Ulrich, S.M., Jones, Jr., M., & Jones, S.M. (2002)	Final grade *M, SD, n*	-.129* (Year 1 of study)	-.479	.222
Bradley, A.Z., Ulrich, S.M., Jones, Jr., M., & Jones, S.M. (2002)	Final grade *M, SD, n*	-.021* (Year 2 of study)	-.371	.329
Chang, C. & Mao, S. (1999a)	*F*-values for pre- to post-test difference	.181	.022	.339
Chang, C. & Mao, S. (1999b)	Post-test *M, SD, n*	.088	-.751	.928
Chung-Schickler, G.C. (1998)	Pre- to post-test change, *M, SD, n*	-.486	-1.072	.099
De Baz, T. (2001)	Post-test *M, SD, n*	1.483* (high ability students)	.992	1.975
De Baz, T. (2001)	Post-test *M, SD, n*	1.157* (medium ability students)	.687	1.626

Table 3 (cont'd)

Meta-Analytic Results of All Studies

Study	Statistic(s) Reported	Effect Size	Confidence Interval	
		Hedges's *g*	Lower	Upper
De Baz, T. (2001)	Post-test *M, SD, n*	-.335 (low ability students)	-.772	.103
Dori, Y.J., Yeroslavski, O., & Lazarowitz, R. (1995)	Pre- and post-test *M, SD, n*	.766	.374	1.157
Faro, S. & Swan, K. (2006)	Pre- and post-test *M, SD, n*	.309	-.303	.920
Foley. K. & O'Donnell, A. (2002)	Post-test *M, SD, n*	.181 (2 outcomes combined)	-.401	.764
Fontenot, D.W. (1995)	Pre- and post-test *M, SD, n*	2.401	2.150	2.652
Gayford, C. (1995)	Pre- and post-test *M, SD, n*	.236* (high ability students; 2 outcomes combined)	-.324	.796

Table 3 (cont'd)

Meta-Analytic Results of All Studies

Study	Statistic(s) Reported	Effect Size	Confidence Interval	
		Hedges's *g*	Lower	Upper
Gayford, C. (1995)	Pre- and post-test *M*, *SD*, *n*	1.194* (medium ability students; 2 outcomes combined)	.763	1.625
Gayford, C. (1995)	Pre- and post-test *M*, *SD*, *n*	1.290* (low ability students; 2 outcomes combined)	.676	1.904
Hanze, M. & Berger, R. (2007)	Post-test *M, SD, n*	-.236	-.473	.001
Harskamp, E. & Ding, N. (2006)	Pre- and post-test *M*, *SD*, *n*	.475	-.078	1.029
Jensen, M.S. (1996)	Gain score *M, SD, n*; Post-test *M, SD, n*	.225 (3 outcomes combined)	-.066	.516
Jeon, K., Huffman, D., & Noh, T. (2005)	Post-test *M, SD, n*	.496	-.025	1.017

Table 3 (cont'd)

Meta-Analytic Results of All Studies

Study	Statistic(s) Reported	Effect Size	Confidence Interval	
		Hedges's *g*	Lower	Upper
Lazarowitz, R. (1996)	Pre- and post-test *M*, *SD*, *n*	-.058 (2 outcomes combined)	-.428	.309
Lumpe, A. (1995)	Post-test *M, SD, n*; Pre- and post-test, *p*-value	1.593 (2 outcomes combined)	.307	2.878
Qualter, A. & Abu-hola, I.R.A. (2000)	Post-test *M, SD, n*	.351* (females, grade 10, 2 outcomes combined)	-.030	.732
Qualter, A. & Abu-hola, I.R.A. (2000)	Post-test *M, SD, n*	.997* (males, grade 10, 2 outcomes combined)	.568	1.425
Qualter, A. & Abu-hola, I.R.A. (2000)	Post-test *M, SD, n*	.469* (females, grade 9, 2 outcomes combined)	.084	.854

Table 3 (cont'd)

Meta-Analytic Results of All Studies

Study	Statistic(s) Reported	Effect Size	Confidence Interval	
		Hedges's *g*	Lower	Upper
Qualter, A. & Abu-hola, I.R.A. (2000)	Post-test *M, SD, n*	.525* (males, grade 9, 2 outcomes combined)	.117	.932
Qualter, A. & Abu-hola, I.R.A. (2000)	Post-test *M, SD, n*	-.222* (females, grade 6, 2 outcomes combined)	-.604	.161
Qualter, A. & Abu-hola, I.R.A. (2000)	Post-test *M, SD, n*	.512* (males, grade 6, 2 outcomes combined)	.109	.915
Roy, H. (2003)	Pre- to post-test change, *t*-value	.391	.007	.775
Sadler, K.C. (2002)	*F*-values for final grade and pre- to post-test difference	.362* (2 outcomes combined)	.064	.660

Table 3 (cont'd)

Meta-Analytic Results of All Studies

Study	Statistic(s) Reported	Effect Size	Confidence Interval	
		Hedges's *g*	Lower	Upper
Schroeder, P.G. (1996)	Pre- and post-test *M*, *SD*, *n*	-.087	-.626	.452
Shachar, H. & Fischer, S. (2004)	Pre- and post-test *M*, *SD*, *n*	-.233 (high ability students)	-.757	.290
Shachar, H. & Fischer, S. (2004)	Pre- and post-test *M*, *SD*, *n*	.933 (medium ability students)	.388	1.478
Shachar, H. & Fischer, S. (2004)	Pre- and post-test *M*, *SD*, *n*	.663 (low ability students)	.116	1.211
Snyder, T. & Sullivan, H. (1995)	Post-test *M, SD*, *n*	-.345* (high ability students)	-.749	.059
Snyder, T. & Sullivan, H. (1995)	Post-test *M, SD*, *n*	-.377* (medium ability students)	-.657	-.097
Snyder, T. & Sullivan, H. (1995)	Post-test *M, SD*, *n*	-.352* (low ability students)	-.758	.054
Starr, E.M. (1995)	Class grade and Post-test *M, SD*, *n*	.228 (5 outcomes combined)	-.485	.941

Table 3 (cont'd)

Meta-Analytic Results of All Studies

Study	Statistic(s) Reported	Effect Size	Confidence Interval	
		Hedges's *g*	Lower	Upper
Tao, P. (1999)	Post-test, *p*-value	1.553 (School #2)	.400	2.706
Trautwein, S.N., Racke, A., & Hillman, B. (1997)	*n*, *t*-value	.253 (Class #1)	.049	.457
Trautwein, S.N., Racke, A., & Hillman, B. (1997)	*n*, *t*-value	.208 (Class #2)	.010	.406
Verdel, E.F.O. (1996)	Pre- to post-test change, *M, SD, n*	.351	-.088	.791
Werner, J.L. & Klein, J.D. (1999)	Post-test *M, SD, n*	-.533	-1.183	.118
Overall Mean Effect Size		**.400**	**.225**	**.574**

* = *n* for subjects inferred from available data

A forest plot of all effect sizes can be seen in Figure 1. Twenty-six of the calculated effect sizes, which in some cases came from the same study, were statistically significant at the $p < .05$ level. The remaining 25 were not statistically significant at this same level. It is important to note that some of the effect sizes in the forest plot were compiled by study for those that had multiple outcomes or moderator variables. Another observation worth noting from Figure 1 is that the confidence interval for individual effect sizes varied widely between studies. In Figure 1, three columns identify each of the effect sizes from each of the studies. Two columns, "Gender/Ability Level" and "Outcome" identify the effect sizes for studies that disaggregated student achievement data based on the gender or the ability level of the students, and studies that had multiple achievement outcomes. Notice that in both of these columns, some studies have a "Blank" notation. CMA lists this for each study if the "Gender/Ability Level" and "Outcome" are not identified specifically in the data entry. Studies that listed "Blank" in the "Gender/Ability Level" column did not disaggregate data for male and female students. Studies that listed "Blank" in the "Outcome" column reported data for a single outcome measure. Unfortunately, CMA does not allow modification of these columns for the forest plot. The overall mean effect size for the 32 included studies was .400, showing a medium or typical overall effect (Cohen, 1988) of cooperative learning on student achievement in science in comparison to traditional lecture instruction, instruction as usual, or individual versus group work. This effect size was statistically significant at the $p<.05$ level since the confidence interval of .225-.574 does not include zero and the z-value was 4.484 ($p<.0001$).

Study name	Gender/Ability Level	Outcome	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Abobaker (1995)	Females	Blank	-0.193	0.208	0.043	-0.601	0.214	-0.930	0.352
Abobaker (1995)	Males	Blank	0.584	0.215	0.046	0.164	1.004	2.722	0.006
Acar & Tarhan (2007)	Blank	Blank	2.411	0.406	0.165	1.616	3.207	5.940	0.000
Balfakih (2003)	Females	Blank	0.508	0.131	0.017	0.253	0.764	3.895	0.000
Balfakih (2003)	Males	Blank	0.503	0.129	0.017	0.250	0.757	3.894	0.000
Banerjee (1997)	Females	Blank	0.015	0.340	0.115	-0.651	0.681	0.045	0.964
Banerjee (1997)	Males	Blank	-0.425	0.334	0.112	-1.080	0.231	-1.270	0.204
Bilgin (2006)	Blank	Combined	1.050	0.224	0.050	0.610	1.490	4.675	0.000
Bradley, et al. (2002)	Blank	Exams	0.149	0.179	0.032	-0.201	0.499	0.834	0.404
Bradley, et al. (2002)	Year 1	Final grade	-0.129	0.179	0.032	-0.479	0.222	-0.720	0.472
Bradley, et al. (2002)	Year 2	Final grade	-0.021	0.179	0.032	-0.371	0.329	-0.116	0.908
Chang & Mao (1999a)	Blank	Blank	0.181	0.081	0.007	0.022	0.339	2.230	0.026
Chang & Mao (1999b)	Blank	Blank	0.088	0.429	0.184	-0.751	0.928	0.206	0.837
Chung-Schielder (1998)	Blank	Blank	-0.486	0.299	0.089	-1.072	0.099	-1.628	0.103
De Baz (2001)	High ability	Blank	1.483	0.251	0.063	0.992	1.975	5.920	0.000
De Baz (2001)	Low ability	Blank	-0.335	0.223	0.050	-0.772	0.103	-1.500	0.134
De Baz (2001)	Medium ability	Blank	1.157	0.240	0.057	0.687	1.626	4.829	0.000
Dori, et al. (1995)	Blank	Blank	0.766	0.200	0.040	0.374	1.157	3.830	0.000
Faro (2006)	Blank	Blank	0.309	0.312	0.097	-0.303	0.920	0.990	0.322
Foley & O'Donnell (2002)	Blank	Combined	0.181	0.297	0.088	-0.401	0.764	0.610	0.542
Fontenot (1995)	Blank	Blank	2.401	0.128	0.016	2.150	2.652	18.766	0.000
Gayford (1995)	High ability	Combined	0.236	0.286	0.082	-0.324	0.796	0.825	0.409
Gayford (1995)	Low ability	Combined	1.290	0.313	0.098	0.676	1.904	4.117	0.000
Gayford (1995)	Medium ability	Combined	1.194	0.220	0.048	0.763	1.625	5.426	0.000
Harve & Berger (2007)	Blank	Blank	-0.236	0.121	0.015	-0.473	0.001	-1.952	0.051
Harskamp & Ding (2006)	Blank	Blank	0.475	0.282	0.080	-0.078	1.029	1.682	0.093
Jensen (1996)	Blank	Combined	0.225	0.149	0.022	-0.066	0.516	1.516	0.130
Jeon, et al. (2005)	Blank	Blank	0.496	0.266	0.071	-0.025	1.017	1.866	0.062
Lazarowitz (1996)	Blank	Combined	-0.059	0.188	0.035	-0.428	0.309	-0.316	0.752
Lumpe (1995)	Blank	Combined	1.593	0.656	0.430	0.307	2.878	2.428	0.015
Qualter & Abu-hola (2000)	Grade 10 females	Combined	0.351	0.194	0.038	-0.030	0.732	1.807	0.071
Qualter & Abu-hola (2000)	Grade 10 males	Combined	0.997	0.219	0.048	0.568	1.425	4.557	0.000
Qualter & Abu-hola (2000)	Grade 6 females	Combined	-0.222	0.195	0.038	-0.604	0.161	-1.137	0.256
Qualter & Abu-hola (2000)	Grade 6 males	Combined	0.512	0.206	0.042	0.109	0.915	2.487	0.013
Qualter & Abu-hola (2000)	Grade 9 females	Combined	0.469	0.196	0.039	0.084	0.854	2.389	0.017
Qualter & Abu-hola (2000)	Grade 9 males	Combined	0.525	0.208	0.043	0.117	0.932	2.520	0.012
Roy (2003)	Blank	Blank	0.391	0.196	0.038	0.007	0.775	1.998	0.046
Sadler (2002)	Blank	Combined	0.362	0.152	0.023	0.064	0.660	2.383	0.017
Schroeder (1996)	Blank	Combined	-0.087	0.275	0.076	-0.626	0.452	-0.316	0.752
Shachar & Fischer (2004)	High ability	Blank	-0.233	0.267	0.071	-0.757	0.290	-0.874	0.382
Shachar & Fischer (2004)	Low ability	Blank	0.663	0.279	0.078	0.116	1.211	2.375	0.018
Shachar & Fischer (2004)	Medium ability	Blank	0.933	0.278	0.077	0.388	1.478	3.355	0.001
Snyder & Sullivan (1995)	High ability	Blank	-0.345	0.206	0.042	-0.749	0.059	-1.672	0.095
Snyder & Sullivan (1995)	Low ability	Blank	-0.352	0.207	0.043	-0.758	0.054	-1.698	0.090
Snyder & Sullivan (1995)	Medium ability	Blank	-0.377	0.143	0.020	-0.657	-0.097	-2.641	0.008
Starr (1995)	Blank	Combined	0.228	0.364	0.132	-0.485	0.941	0.626	0.531
Tao (1999)	School 1	Blank	1.233	0.465	0.216	0.322	2.145	2.652	0.008
Tao (1999)	School 2	Blank	1.553	0.588	0.346	0.400	2.706	2.639	0.008
Trautwein, et al. (1997)	Class 1	Blank	0.253	0.104	0.011	0.049	0.457	2.431	0.015
Trautwein, et al. (1997)	Class 2	Blank	0.208	0.101	0.010	0.010	0.406	2.055	0.040
Verdel (1996)	Blank	Blank	0.351	0.224	0.050	-0.088	0.791	1.566	0.117
Werner & Klein (1999)	Blank	Blank	-0.533	0.332	0.110	-1.183	0.118	-1.605	0.109
			0.400	0.089	0.008	0.225	0.574	4.484	0.000

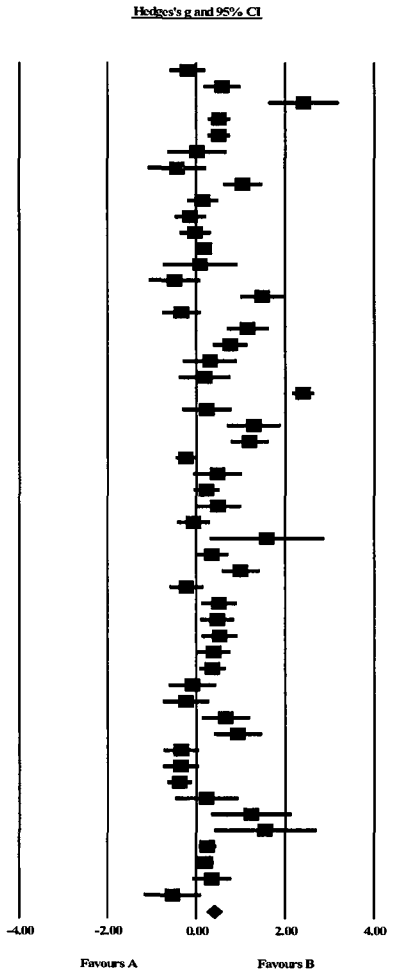


Figure 1: Forest Plot of All Effect Sizes

Heterogeneity Analysis

Based on the heterogeneity results, further examination of the effect sizes for each study revealed two effect size values from different studies that appear to be outliers. The value from the study by Acar and Tarhan (2007) was 2.411 and the value from the study by Fontenot (1995) was 2.401. Both of these values are considerably higher than the next highest effect size and are at least 2 standard deviations from the overall mean effect size value. Lipsey and Wilson (2001) suggest two methods for dealing with outliers. First, they can simply be eliminated or trimmed from the original sample of studies and excluded from the analysis. Second, these values can be "Windsorized", a method that adjusts the values of the outliers to be closer to the next closest effect size value. Another analysis of value for dealing with outliers is the "One Study Removed" (CMA, 2007) analysis, in which each effect size is removed individually from the data set and the resulting overall mean effect size is recalculated.

Trimming and "Windsorizing" of Outliers

In the process of trimming and "Windsorizing" (Lipsey & Wilson, 2001), the effect size values for all the included studies were organized in a spreadsheet and imported into the Statistical Package for the Social Sciences (SPSS®) software for further analysis. A mean for the effect size values was calculated and then further analysis was conducted. The two outliers were "trimmed" from the data set, meaning they were eliminated, with a mean calculated for the remaining effect sizes. In addition, the outlier values (Acar & Tarhan, 2007 and Fontenot, 1995) were "Windsorized," which means that the values were adjusted to equal the next highest effect size value, which in this case was 1.593. Descriptive statistics from these analyses can be seen in Table 4.

Table 4

Trimmed and "Windsorized" Effect Size Values

Effect Size Values	Mean	Standard Deviation	Skewness	Kurtosis
All Studies Included	.429	.680	1.051	1.120
Outliers Removed	.350	.562	.535	-.404
Outliers "Windsorized"	.397	.601	.530	-.572

"One Study Removed"

The "one study removed" analysis (CMA, 2007) revealed that when the effect size value from the Acar and Tarhan (2007) study was removed, the recalculated overall mean effect size value was .368, .032 standard deviations different from the original value. When the effect size value from the Fontenot (1995) study was removed, the overall mean effect size value was .339, .061 standard deviations different from the original value. The removal of the effect sizes from these two studies had the largest impact on the original overall mean effect size in comparison to all the other studies.

Heterogeneity Analysis Conclusions

Based on the results of the heterogeneity analysis, trimming and "Windsorizing," and the "one study removed" analysis, the two extreme outlier effect size values were removed from the data set and any subsequent sensitivity analyses and moderator

analyses. The removal of these values had multiple effects. First, the new overall mean effect size value was .308, and somewhat lower than the original value. The new value was still statistically significant ($z = 4.850, p < .0001, CI = .184\text{-}.433$). Moreover, the overall heterogeneity of the effect size sample decreased somewhat ($Q = 239.981, p < .0001, I^2 = 79.582$), but still necessitated the use of the random effects model for effect size calculation in moderator analyses based on these values and the determination that the studies represented unique effects and not a single one.

Sensitivity Analysis on Included Studies

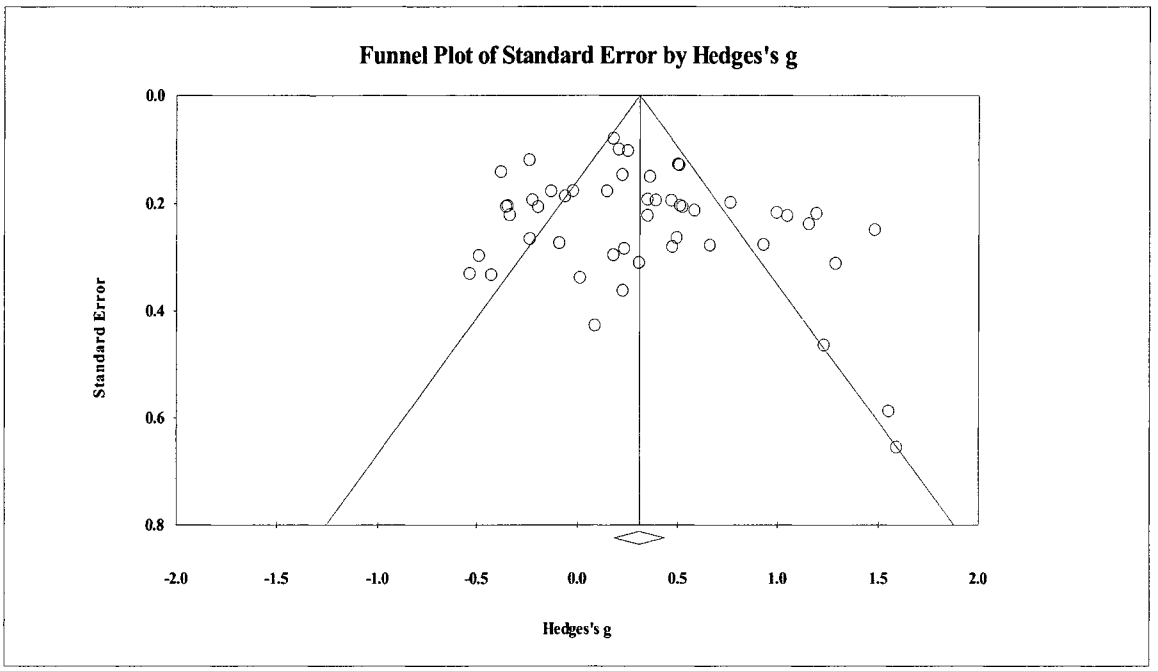
Publication Bias

Assessing the heterogeneity of studies included in a meta-analysis is important in directing the effect size calculation, but other assessments are necessary to address other variables which may affect the conclusions drawn. Publication bias, also known as the "file drawer effect" (Iyenger & Greenhouse, 1988) serves to address the issue that the sample of studies included in a meta-analysis is only representative of those that are published or show statistically significant effects, and not representative of the true sample of studies that could exist. Publication bias can be addressed in multiple ways. In this study, three analyses were performed. First, funnel plots were constructed (Figure 2). These figures plot all the included studies by study size (measured by standard error) versus calculated effect size. In the absence of publication bias, the studies should be symmetrically distributed around the overall mean effect size and clustered near the top of the plot. In the presence of bias, the studies would be distributed asymmetrically around the overall mean effect size and clustered toward the bottom of the plot. This distribution reflects the fact that smaller studies are more likely to be published if they

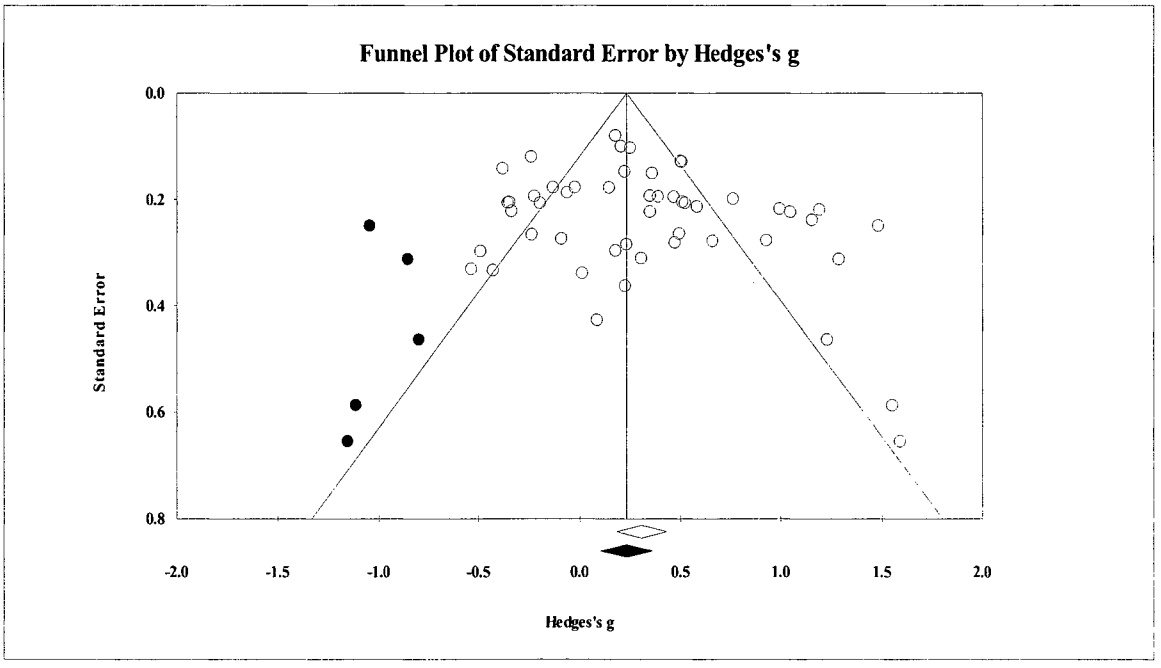
have larger than average effects (Light & Pillemer, 1984). As can be seen in Figure 2, part a plots all of the effect sizes included in the meta-analysis around the overall mean effect size. This symmetry of effect sizes around the mean indicates that publication bias is largely absent from the current study.

Trim and Fill Analysis

The second analysis that was performed was the "trim and fill" analysis (Duval & Tweedie, 2000) which expands on the funnel plot (part b of Figure 2). This particular analysis imputes the studies that may be missing from the study sample, where they are likely to fall on a funnel plot, and then recalculates the combined effect size. This particular analysis revealed an imputed point estimate of the overall mean effect size of .300 (CI = .175-.424), which was statistically significant at the $p < .05$ level and very close to the original overall mean effect size value. This imputed point estimate of the overall mean effect size indicates that any potentially missing studies would have little to no effect on the actual overall mean effect size and adds to the notion that publication bias is largely absent from the current study.



(a)



(b)

Figure 2: Funnel plots of (a) effect sizes from included studies only, and (b) effect sizes from included and imputed studies.

Fail Safe N Analysis

Another method that addresses publication bias is the "fail safe *N*" analysis (Cooper & Hedges, 1994). This particular analysis determines the number of missing studies that would be necessary to nullify the effect of an intervention determined in a meta-analysis. If this number is relatively small, there may be cause for concern regarding any conclusions drawn from the overall mean effect size. However, if this number is large, one can be relatively confident that the treatment effect, while possibly inflated by the missing studies, is not zero. It also reports a *z*-value with an associated *p*-value to determine statistical significance of the calculated effect size. This analysis has two primary limitations. First, it assumes that the effect in the hidden or missing studies is zero, rather than considering that the missing studies could actually have shown an effect in either a positive or negative direction. Secondly, this analysis focuses on statistical significance rather than substantive significance of a treatment. Despite these limitations, this particular analysis does add support to the conclusions taken from a meta-analysis. The "fail-safe *N*" analysis yielded a *z*-value of 9.901 (2-tailed $p <$ .000001) which was statistically significant. The "fail-safe *N*" value was 1,226, which means that this many "null" studies would need to be located and included for the 2-tailed *p*-value to exceed .05 and render the overall mean effect size statistically non-significant. Stated another way, there would need to be 24.5 missing "null" studies for each included study for the observed effect to be nullified. This analysis, in conjunction with the two mentioned above, further support the notion that publication bias is largely absent from the current study.

Moderator Analyses

General Results

Moderator analyses based on certain student characteristics were limited since many of the studies that were included in the meta-analysis did not disaggregate data based on student demographics. Ethnicity was not identified as a moderator variable because no studies disaggregated student achievement data based on this participant characteristic.

In all moderator analyses that were conducted, the Q-statistic and I^2-statistic were statistically significant for heterogeneity of the studies included in each analysis necessitating a random effects model of effect size calculation. Specific values are listed in each section below.

Gender of Students

Gender was identified as a moderator variable, but in only four studies (Table 5). Each of these studies reported outcome data for students in each gender. The data from these studies was entered into CMA separately and identified as "male" or "female," and a separate effect size was calculated. CMA allows the user to perform analyses based on identified moderator variables. Based on the data reported in the table, it appears that the intervention has a null effect on science achievement of female students since the effect size value of .179 was not statistically significant ($z = 1.227$, $p = .220$) and the confidence interval includes zero. This sample of studies exhibited significant heterogeneity as well ($Q = 16.299$, $p = .006$; $I^2 = 69.324$). By contrast, the intervention has a medium effect (Cohen, 1988) on males ($ES = .506$, $z = 3.748$, $p < .0001$), with the sample of studies also exhibiting significant heterogeneity ($Q = 12.822$, $p = .025$; $I^2 = 61.004$). While both of

these effects on each group of students is documented in the effect sizes, a sample of only four studies lacks statistical power (Borenstein, et al., 2009). Additionally, the heterogeneity of the effect sizes is significant and the range of the effect sizes for this moderator analysis is quite broad. These observations make any conclusions regarding the impact of gender to be suspect.

Table 5

Effect Sizes Based on Gender of Participants

Study	Effect Size	Confidence Interval	
Females	Hedges's *g*	Lower	Upper
Abobaker, N.M. (1995)	-.193	-.601	.214
Balfakih, N.M.A. (2003)	.508	.253	.764
Banerjee, A. (1997)	.015	-.651	.681
Qualter, A. & Abu-hola, I.R.A. (2000)	.199	-.219	.617
Overall Mean Effect Size (Females)	**.179**	**-.107**	**.465**
Males			
Abobaker, N.M. (1995)	.584	.164	1.004
Balfakih, N.M.A. (2003)	.503	.250	.757
Banerjee, A. (1997)	-.425	-1.080	.231
Qualter, A. & Abu-hola, I.R.A. (2000)	.671	.364	.977
Overall Mean Effect Size (Males)	**.506**	**.241**	**.771**

Ability Level of Students

Only four studies identified ability level of the students as moderator variable in their results (see Table 6). These studies identified three ability levels, low, medium, and high. These levels correspond to the students' performance on assessment instruments that were completed prior to the intervention. Data from these studies were extracted in a similar manner as with gender. All three effect sizes associated with each ability level were statistically non-significant. The effect size for high ability students was .283 (z = .655, p = .513), with these studies exhibiting significant heterogeneity (Q = 35.798, p < .0001; I^2 = 91.620). The effect size for medium ability students was .714 (z = 1.554, p = .120), with these studies exhibiting significant heterogeneity (Q = 56.254, p < .0001; I^2 = 94.667). The effect size for low ability students was .290 (z = .771, p = .441), with these studies exhibiting significant heterogeneity (Q = 27.043, p < .0001; I^2 = 88.907). The heterogeneity statistics and the broad range effect sizes from each of the four studies are highly varied, and the number of studies is small, which calls into question any major conclusions for this moderator variable.

Table 6

Effect Sizes Based on Ability Level of Participants

Study	Effect Size	Confidence Interval	
High Ability Students	Hedges's *g*	Lower	Upper
De Baz, T. (2000)	1.483	992	1.975
Gayford, C. (1995)	.236	-.324	.796
Shachar, H. & Fischer, S. (2004)	-.233	-.757	.290
Snyder, T. & Sullivan, H. (1995)	-.345	-.749	.059
Overall Mean Effect Size (High Ability Students)	**.283**	**-.563**	**1.129**
Medium Ability Students			
De Baz, T. (2000)	1.157	.687	1.626
Gayford, C. (1995)	1.194	.763	1.625
Shachar, H. & Fischer, S. (2004)	.933	.388	1.478
Snyder, T. & Sullivan, H. (1995)	-.377	-.657	-.097
Overall Mean Effect Size (Medium Ability Students)	**.714**	**-.187**	**1.554**

Table 6 (cont'd)

Effect Sizes Based on Ability Level of Participants

Study	Effect Size	Confidence Interval	
	Hedges's *g*	Lower	Upper
Low Ability Students			
De Baz, T. (2000)	-.335	-.772	.103
Gayford, C. (1995)	1.290	.676	1.904
Shachar, H. & Fischer, S. (2004)	.663	.116	1.211
Snyder, T. & Sullivan, H. (1995)	-.352	-.758	.054
Overall Mean Effect Size (Low Ability Students)	**.290**	**-.448**	**1.027**

Intervention Type

Another moderator variable study suggested in Chapter 3 involved the type of CL being implemented in the study. During the final full-text coding, the intervention in a given study was scrutinized very closely in order to determine what type of cooperative learning was being utilized. Fifteen studies did not specifically identify the CL structure utilized, but described the intervention essentially as students working in groups to achieve a common goal. Specific types of cooperative learning were clearly indicated in 15 studies. Of these studies, four was the largest number that identified a specific CL strategy (STAD and Jigsaw), so the data from these studies was compiled for the

moderator analysis. For all effect size calculations, CL intervention was defined as a moderator variable in CMA as either the specific type of CL, or NS (not specified). Data from any studies that had multiple outcomes or disaggregated data based on subject characteristics were grouped in the effect size calculation (see Table 7). The effect size for the unspecified intervention was .268 ($z = 3.307$, $p = .001$) and was statistically significant. This group of studies exhibited significant heterogeneity. The Q-statistic was 123.734 and was statistically significant ($p < .0001$). The I^2-statistic of 78.18 described significant heterogeneity among the studies as well. The effect size for the structured interventions was also statistically significant ($ES = .367$, $z = 3.480$, $p = .001$). This sample of studies also exhibited significant heterogeneity ($Q = 115.382$, $p < .0001$; $I^2 = 81.8$). Both effect size exhibit medium effect (Cohen, 1988), but the value for the structured interventions was approximately one standard deviation higher than that for the unstructured intervention. For all of the studies included in this meta-analysis, student achievement in science was measured on an individual basis in the treatment and control conditions. Six of the seven meta-analyses mentioned previously which assessed the effectiveness of CL in comparison to traditional lecture or individual learning experiences did not specify the CL type that was employed in the included studies, just that this particular intervention took place. One of the studies (Johnson, et al., 2000) identified ten different CL interventions (described in Chapter 2), calculated overall mean effect sizes for 8 of them, and noted differential effects of each type of CL. In this meta-analysis, the Learning Together format showed the largest effect, with Jigsaw showing the smallest in comparison to individualistic learning efforts. Effect sizes ranged from .13-1.04. All of the seven meta-analyses included studies in which student achievement

was measured individually, with no group accountability. Since most of the previous studies did not identify the intervention type as a moderator, this analysis was completed in the current study.

Table 7

Effect Sizes Based on Intervention Characteristics

Study	Effect Size	Confidence Interval	
Intervention Not Specified	Hedges's g	Lower	Upper
Banerjee, A. (1997)	-.208	-.675	259
Bilgin, I. (2006)	1.050	.610	1.490
Bradley, A.Z., Ulrich, S.M., Jones, Jr., M., & Jones, S.M. (2002)	.000	-.202	.202
Foley. K. & O'Donnell, A. (2002)	.181	-.401	.764
Gayford, C. (1995)	.911	.272	1.550
Harskamp, E. & Ding, N. (2006)	.475	-.078	1.029
Jensen, M.S. (1996)	.225	-.066	.516
Lumpe, A. (1995)	1.593	.307	2.878
Qualter, A. & Abu-hola, I.R.A. (2000)	.432	.121	.743
Roy, H. (2003)	.391	.007	.775

Table 7 (cont'd)

Effect Sizes Based on Intervention Characteristics

Study	Effect Size	Confidence Interval	
Intervention Not Specified	Hedges's *g*	Lower	Upper
Sadler, K.C. (2002)	.362	.064	.660
Snyder, T. & Sullivan, H. (1995)	-.363	-.563	-.163
Trautwein, S.N., Racke, A., & Hillman, B. (1997)	.230	.088	.372
Verdel, E.F.O. (1996)	.351	-.088	.791
Werner, J.L. & Klein, J.D. (1999)	-.533	-1.183	.118
Overall Mean Effect Size (Intervention not Specified)	**.268**	**.109**	**.427**
Intervention Specified			
Abobaker, N.M. (1995)	.193	-.568	.955
Balfakih, N.M.A. (2003)	.506	.326	.686
Chang, C. & Mao, S. (1999a)	.181	.022	.339
Chang, C. & Mao, S. (1999b)	.088	-.751	.928
Chung-Schickler, G.C. (1998)	-.486	-1.072	.099
De Baz, T. (2001)	.764	-.359	1.887

Table 7 (cont'd)

Effect Sizes Based on Intervention Characteristics

Study	Effect Size	Confidence Interval	
Intervention Specified	Hedges's *g*	Lower	Upper
Dori, Y.J., Yeroslavski, O., & Lazarowitz, R. (1995)	.766	.374	1.157
Faro, S. & Swan, K. (2006)	.309	-.303	.920
Hanze, M. & Berger, R.(2007)	-.236	-.473	.001
Jeon, K., Huffman, D., & Noh, T. (2005)	.496	-.025	1.017
Lazarowitz, R. (1996)	-.059	-.428	.309
Schroeder, P.G. (1996)	-.087	-.626	.452
Shachar, H. & Fischer, S. (2004)	.451	-.249	1.150
Starr, E.M. (1995)	.228	-.485	.941
Tao, P. (1999)	1.356	.641	2.071
Overall Mean Effect Size (Intervention Specified)	**.367**	**.160**	**.573**

Experimental Design

Another moderator analysis that was performed relates to experimental design in included studies. No studies identified a random controlled trial (RCT) methodology. Eleven studies utilized a cluster randomized methodology, six utilized a quasi-

experimental with subject matching methodology, and thirteen utilized a quasi-experimental without subject matching methodology (see Table 8). Data from these studies was compiled in a similar fashion as was done with the intervention moderator analysis. These data indicate that the effect size in the included studies differs based on the methodology employed in the study. The studies identified as "cluster randomized" had a statistically significant effect size ($ES = .372$, $z = 3.651$, $p < .0001$) and exhibited significant heterogeneity ($Q = 74.426$, $p < .0001$; $I^2 = 77.158$). The studies identified as "quasi-experimental without subject matching" also had a statistically significant effect size ($ES = .349$, $z = 4.374$, $p < .0001$) and exhibited significant heterogeneity ($Q = 55.949$, $p < .0001$; $I^2 = 67.828$). The "quasi-experimental with subject matching" methodology had a null effect ($ES = .187$, $z = 1.080$, $p = .280$), with the studies also exhibiting significant heterogeneity ($Q = 91.114$, $p < .0001$; $I^2 = 86.830$).

Table 8

Effect Sizes Based on Research Methodology

Study		Effect Size	Confidence Interval	
	Methodology	Hedges's *g*	Lower	Upper
Abobaker, N.M. (1995)	Cluster Randomized	.193	-.568	.955
Balfakih, N.M.A. (2003)	Cluster Randomized	.506	.326	.686
Bilgin, I. (2006)	Cluster Randomized	1.050	.610	1.490

Table 8 (cont'd)

Effect Sizes Based on Research Methodology

Study		Effect Size	Confidence Interval	
	Methodology	Hedges's *g*	Lower	Upper
Foley. K. & O'Donnell, A. (2002)	Cluster Randomized	.181	-.401	.764
Hanze, M. & Berger, R.(2007)	Cluster Randomized	-.236	-.473	.001
Harskamp, E. & Ding, N. (2006)	Cluster Randomized	.475	-.078	1.029
Jeon, K., Huffman, D., & Noh, T. (2005)	Cluster Randomized	.496	-.025	1.017
Lumpe, A. (1995)	Cluster Randomized	1.593	.307	2.878
Qualter, A. & Abu-hola, I.R.A. (2000)	Cluster Randomized	.432	.121	.743
Verdel, E.F.O. (1996)	Cluster Randomized	.351	-.088	.791
Werner, J.L. & Klein, J.D. (1999)	Cluster Randomized	-.533	-1.183	.118
Overall Mean Effect Size	**Cluster Randomized**	**.372**	**.172**	**.572**

Table 8 (cont'd)

Effect Sizes Based on Research Methodology

Study		Effect Size	Confidence Interval	
	Methodology	Hedges's *g*	Lower	Upper
Bradley, A.Z., Ulrich, S.M., Jones, Jr., M., & Jones, S.M. (2002)	Quasi-experimental with Subject Matching	.000	-.202	.202
Chung-Schickler, G.C. (1998)	Quasi-experimental with Subject Matching	-.486	-1.072	.099
De Baz, T. (2001)	Quasi-experimental with Subject Matching	.764	-.359	1.887
Schroeder, P.G. (1996)	Quasi-experimental with Subject Matching	-.087	-.626	.452

Table 8 (cont'd)

Effect Sizes Based on Research Methodology

Study		Effect Size	Confidence Interval	
	Methodology	Hedges's *g*	Lower	Upper
Snyder, T. & Sullivan, H. (1995)	Quasi-experimental with Subject Matching	-.363	-.563	-.163
Tao, P. (1999)	Quasi-experimental with Subject Matching	1.356	.641	2.071
Overall Mean Effect Size	**Quasi-experimental with Subject Matching**	**.187**	**-.153**	**.527**

Table 8 (cont'd)

Effect Sizes Based on Research Methodology

Study		Effect Size	Confidence Interval	
	Methodology	Hedges's *g*	Lower	Upper
Banerjee, A. (1997)	Quasi-experimental without subject matching	-.208	-.675	.259
Chang, C. & Mao, S. (1999a)	Quasi-experimental without subject matching	.181	.022	.339
Chang, C. & Mao, S. (1999b)	Quasi-experimental without subject matching	.088	-.751	.928
Dori, Y.J., Yeroslavski, O., & Lazarowitz, R. (1995)	Quasi-experimental without subject matching	.766	.374	1.157

Table 8 (cont'd)

Effect Sizes Based on Research Methodology

Study		Effect Size	Confidence Interval	
	Methodology	Hedge's *g*	Lower	Upper
Faro, S. & Swan, K. (2006)	Quasi-experimental without subject matching	.309	-.303	.920
Gayford, C. (1995)	Quasi-experimental without subject matching	.911	.272	1.550
Jensen, M.S. (1996)	Quasi-experimental without subject matching	.225	-.066	.516
Lazarowitz, R. (1996)	Quasi-experimental without subject matching	-.059	-.428	.309

Table 8 (cont'd)

Effect Sizes Based on Research Methodology

Study		Effect Size	Confidence Interval	
	Methodology	Hedge's *g*	Lower	Upper
Roy, H. (2003)	Quasi-experimental without subject matching	.391	.007	.775
Sadler, K.C. (2002)	Quasi-experimental without subject matching	.362	.064	.660
Shachar, H. & Fischer, S. (2004)	Quasi-experimental without subject matching	.451	-.249	1.150
Starr, E.M. (1995)	Quasi-experimental without subject matching	.228	-.485	.941

Table 8 (cont'd)

Effect Sizes Based on Research Methodology

Study		Effect Size	Confidence Interval	
	Methodology	Hedge's *g*	Lower	Upper
Trautwein, S.N., Racke, A., & Hillman, B. (1997)	Quasi-experimental without subject matching	.230	.088	.372
Overall Mean Effect Size	**Quasi-experimental without Subject Matching**	**.349**	**.192**	**.505**

Science Discipline

Another moderator analysis that was performed was based on the science discipline that was being taught in each study. Table 9 displays the effect size data based on science discipline. Eleven studies were conducted in biology and chemistry classrooms, four in earth science classrooms, and three in physics classrooms. In biology classrooms, the effect size of .345 was statistically significant ($z = 2.781, p = .005$) and the sample of studies exhibited significant heterogeneity ($Q = 131.765, p < .0001; I^2 = 87.098$). The same could be said for chemistry classrooms ($ES = .235, z = 2.422, p <$

.015; $Q = 58.380$, $p < .0001$; $I^2 = 70.881$). In earth science classrooms, the effect was also statistically significant ($ES = .187$, $z = 2.483$, $p < .013$), but the heterogeneity of this sample of studies was not statistically significant ($Q = .224$, $p = .974$; $I^2 = 0$). The effect size value was the same whether the random effects model of fixed effects model for calculation was used. In physics classrooms, the effect was not statistically significant ($ES = .632$, $z = 1.547$, $p = .122$) and the sample of studies exhibited significant heterogeneity ($Q = 20.636$, $p < .0001$; $I^2 = 85.462$). The effect of the intervention in general science classrooms was not analyzed further since only one study was identified in this science discipline.

Table 9

Effect Sizes Based on Science Discipline

Study			Confidence Interval	
	Science Discipline	Effect Size	Lower	Upper
Chung-Schickler, G.C. (1998)	Biology	-.486	-1.072	.099
De Baz, T. (2001)	Biology	.764	-.359	1.887
Dori, Y.J., Yeroslavski, O., & Lazarowitz, R. (1995)	Biology	.766	.374	1.157
Gayford, C. (1995)	Biology	.911	.272	1.550
Jensen, M.S. (1996)	Biology	.225	-.066	.516

Table 9 (cont'd)

Effect Sizes Based on Science Discipline

Study			Confidence Interval	
	Science Discipline	Effect Size	Lower	Upper
Lazarowitz, R. (1996)	Biology	-.059	-.428	.309
Lumpe, A. (1995)	Biology	1.593	.307	2.878
Roy, H. (2003)	Biology	.391	.007	.775
Sadler, K.C. (2002)	Biology	.362	.064	.660
Snyder, T. & Sullivan, H. (1995)	Biology	-.363	-.563	-.163
Trautwein, S.N., Racke, A., & Hillman, B. (1997)	Biology	.230	.088	.372
Overall Mean Effect Size	**Biology**	**.345**	**.102**	**.588**
Abobaker, N.M. (1995)	Chemistry	.193	-.568	.955
Balfakih, N.M.A. (2003)	Chemistry	.506	.326	.686
Banerjee, A. (1997)	Chemistry	-.208	-.675	.259
Bilgin, I. (2006)	Chemistry	1.050	.610	1.490

Table 9 (cont'd)

Effect Sizes Based on Science Discipline

Study			Confidence Interval	
	Science Discipline	Effect Size	Lower	Upper
Bradley, A.Z., Ulrich, S.M., Jones, Jr., M., & Jones, S.M. (2002)	Chemistry	.000	-.202	.202
Foley. K. & O'Donnell, A. (2002)	Chemistry	.181	-.401	.764
Jeon, K., Huffman, D., & Noh, T. (2005)	Chemistry	.496	-.025	1.017
Schroeder, P.G. (1996)	Chemistry	-.087	-.626	.452
Shachar, H. & Fischer, S. (2004)	Chemistry	.451	-.249	1.150
Verdel, E.F.O. (1996)	Chemistry	.351	-.088	.791
Werner, J.L. & Klein, J.D. (1999)	Chemistry	-.533	-1.183	.118
Overall Mean Effect Size	**Chemistry**	**.235**	**.045**	**.424**

Table 9 (cont'd)

Effect Sizes Based on Science Discipline

Study			Confidence Interval	
	Science Discipline	Effect Size	Lower	Upper
Chang, C. & Mao, S. (1999a)	Earth Science	.181	.022	.339
Chang, C. & Mao, S. (1999b)	Earth Science	.088	-.751	.928
Faro, S. & Swan, K. (2006)	Earth Science	.309	-.303	.920
Starr, E.M. (1995)	Earth Science	.228	-.485	.941
Overall Mean Effect Size	**Earth Science**	**.187**	**.039**	**.335**
Hanze, M. & Berger, R.(2007)	Physics	-.236	-.473	.001
Harskamp, E. & Ding, N. (2006)	Physics	.475	-.078	1.029

Table 9 (cont'd)

Effect Sizes Based on Science Discipline

Study			Confidence Interval	
	Science Discipline	Effect Size	Lower	Upper
Tao, P. (1999)	Physics	1.356	.641	2.071
Overall Mean Effect Size	**Physics**	**.632**	**-.169**	**1.434**

Reliability Testing of Assessment Instrument

The final moderator analysis that was performed was based on whether the researchers conducted reliability testing on the assessment instrument used in the study to measure science achievement outcome. If a study reported final grade as an outcome measure, this value was excluded from the effect size calculation. The other data were compiled as in the moderator analysis completed on country of origin. Table 10 displays the effect size data from this analysis. Nineteen studies reported reliability testing, with an overall mean effect size of .328 ($z = 4.017, p < .0001$), which was statistically significant. This sample of studies exhibited significant heterogeneity ($Q = 176.881, p < .0001; I^2 = 80.778$). Eleven studies did not report reliability testing, with an overall mean effect size of .332 ($z = 2.846, p = .0004$), which was also statistically significant. This sample of studies exhibited significant heterogeneity as well ($Q = 57.346, p < .0001; I^2 = 79.074$).

Table 10

Effect Sizes Based on Reliability Testing of Assessment Instrument

Study	Effect Size	Confidence Interval	
Reliability Testing Reported	Hedges's *g*	Lower	Upper
Abobaker, N.M. (1995)	.193	-.568	.955
Balfakih, N.M.A. (2003)	.506	.326	.686
Banerjee, A. (1997)	-.208	-.675	.259
Bilgin, I. (2006)	1.050	.610	1.490
Chang, C. & Mao, S. (1999a)	.181	.022	.339
Chang, C. & Mao, S. (1999b)	.088	-.751	.928
De Baz, T. (2001)	.764	-.359	1.887
Dori, Y.J., Yeroslavski, O., & Lazarowitz, R. (1995)	.766	.374	1.157
Faro, S. & Swan, K. (2006)	.309	-.303	.920
Jeon, K., Huffman, D., & Noh, T. (2005)	.496	-.025	1.017
Lazarowitz, R. (1996)	-.059	-.428	.309
Qualter, A. & Abu-hola, I.R.A. (2000)	.432	.121	.743
Sadler, K.C. (2002)	.362	.064	.660
Schroeder, P.G. (1996)	-.087	-.626	.452

Table 10 (cont'd)

Effect Sizes Based on Reliability Testing of Assessment Instrument

Study	Effect Size	Confidence Interval	
Reliability Testing Reported	Hedges's *g*	Lower	Upper
Shachar, H. & Fischer, S. (2004)	.451	-.249	1.150
Snyder, T. & Sullivan, H. (1995)	-.363	-.563	-.163
Tao, P. (1999)	1.356	.641	2.071
Verdel, E.F.O. (1996)	.351	-.088	.791
Werner, J.L. & Klein, J.D. (1999)	-.533	-1.183	.118
Overall Mean Effect Size (Reliability Testing Reported)	**.328**	**.168**	**.489**
Reliability Testing Not Reported			
Bradley, A.Z., Ulrich, S.M., Jones, Jr., M., & Jones, S.M. (2002)	.000	-.202	.202
Chung-Schickler, G.C. (1998)	-.486	-1.072	.099

Table 10 (cont'd)

Effect Sizes Based on Reliability Testing of Assessment Instrument

Study	Effect Size	Confidence Interval	
Reliability Testing Not Reported	Hedges's *g*	Lower	Upper
Foley. K. & O'Donnell, A. (2002)	.181	-.401	.764
Gayford, C. (1995)	.911	.272	1.550
Hanze, M. & Berger, R.(2007)	-.236	-.473	.001
Harskamp, E. & Ding, N. (2006)	.475	-.078	1.029
Jensen, M.S. (1996)	.225	-.066	.516
Lumpe, A. (1995)	1.593	.307	2.878
Roy, H. (2003)	.391	.007	.775
Starr, E.M. (1995)	.228	-.485	.941
Trautwein, S.N., Racke, A., & Hillman, B. (1997)	.230	.088	.372
Overall Mean Effect Size (Reliability Testing Not Reported)	**.332**	**.103**	**.561**

CHAPTER 5: DISCUSSION

Introduction

The findings of the study reported here will be addressed in multiple ways in this chapter. First, comparisons with prior meta-analyses addressing the effect of cooperative learning (CL) instruction on student achievement will be made with regard to intervention characteristics, the nature of the sample and settings in each study, methodological characteristics, outcome variable characteristics, and moderator analyses. A justification for the inclusion of the comparison studies will precede the comparisons. Next, recommendations for future research and the use of the results of the current study will be addressed. Lastly, the limitations of the current study and any changes that could have been made to the study will be discussed.

Meta-Analysis as a Research Methodology

The meta-analytic research synthesis technique was developed in the 1970s by Glass (1976) and has undergone a considerable amount of development and evolution since (Lipsey & Wilson, 2001). As mentioned previously, the merits and validity of the meta-analytic technique have been widely reported. The value of this methodology is in the ability to synthesize the results of primary research on a particular intervention into an overall effect. While this technique continues to have its detractors, it has become broadly accepted by medical and social science research circles as an effective means by

which to determine the efficacy of drug treatments or educational interventions for example, and is often used to drive policy decisions in education.

Comparisons with Previous Meta-Analyses

A number of meta-analyses examining the impact of cooperative learning instruction on student achievement have been conducted in the past 20-30 years. Five of these studies will be assessed and compared to the current study reported here.

Rationale for Inclusion of Comparison Studies

Qin, et al. (1995), as mentioned previously, performed a meta-analysis to assess the difference in student achievement between cooperative and competitive efforts. This study was chosen for comparison based on a number of characteristics. First, Johnson and Johnson are two educators who can be considered leading experts on cooperative learning. They are considered pioneers in research regarding this particular intervention due to the fact that they did some of the first analyses of the impact of cooperative learning on student achievement in comparison to individual learning. Second, this particular review examined the impact of cooperation on student achievement in comparison to competition. Although the current study reported in this dissertation reports a comparison between traditional, lecture-based instruction and cooperative learning, competition is typically an inherent characteristic of this type of instruction so a comparison is valid. Lastly, the Qin, et al. (1995) study included a total of 46 individual studies in the meta-analysis, which constitutes a large sample for this type of research.

Johnson, et al. (2000), performed a meta-analysis in which they identified ten specific types of cooperative learning and reported the effects for eight of them. This review was chosen for comparison for a variety of reasons as well. As in the Qin, et al.

(1995) study, the participation of the Johnson brothers was one criterion for inclusion. In addition, Johnson, et al. (2000) identified specific types of cooperative learning, which was chosen *a priori* as a moderator analysis for the current study. Also, the authors identified student achievement as the outcome variable. Lastly, the authors identified 164 studies for inclusion, representing a large sample size.

Springer, et al. (1999) performed a meta-analysis on the impact of cooperative learning on the achievement of students in science, mathematics, engineering, and technology at the undergraduate level. This review was chosen for inclusion because science achievement was one of the outcome variables, as in the current study. Additionally, the authors measured the achievement in science of undergraduate students, both from 2-year colleges and 4-year institutions, similar to the current study. Lastly, they examined ethnicity as a moderator variable, a decision that was also made *a priori* in the current study.

Bowen (2000) performed a meta-analysis on the impact of cooperative learning on the achievement of students in chemistry classes. This review was chosen for inclusion for many of the same reasons as the Springer, et al. (1999) review. The review population included high school and undergraduate students and examined chemistry achievement as the outcome.

The final review that will be used for comparison purposes is the Scott, et al. (2005) review. This meta-analysis examined the impact of a wide variety of instructional interventions on student achievement in science when compared to traditional, lecture-based instruction. The reasons for its inclusion in the comparison are that it was published relatively recently in comparison to the other studies, and that it was

comprehensive in the types of interventions it examined, which included cooperative learning.

All of the included reviews identified student achievement as the outcome variable. The review reported here also identified student achievement as the outcome variable, making comparisons with previous reviews valid with regard to the outcome measure.

Overall Effect Comparisons

The mean overall effect size calculated in this review indicates that cooperative learning has a positive effect on student achievement in science. The value of .308, with a confidence interval of .184-.433 and $p < .0001$, indicates that the effect size of the intervention is statistically significant. The calculated z-value confirms this conclusion. The number of studies included in the analysis (30) represents a large enough sample to justify this conclusion. The sample of studies exhibited a high level of heterogeneity based on both the Q and I^2 statistics, and the range of calculated effect sizes. The results reported here support the notion that cooperative learning as an instructional intervention increases student achievement in science in comparison to traditional instruction or individual learning techniques. In addition, the publication bias analysis further supports the findings that cooperative learning has a positive effect on student achievement in science.

The Qin, et al. (1995) review reported an overall mean effect size of .55. This study reported a high level of heterogeneity in the sample as well ($Q = 805.48$, $p < .05$; no I^2 value was reported). While they did not report a confidence interval or a z-value with associated p-value representing statistical significance, it is safe to assume that their

results were significant based on the value of the overall mean effect size and the size of the sample of studies (46). The overall mean effect size value is higher than the one reported in the current study, and it also supports the notion that cooperative learning increased student achievement in comparison to traditional instruction. With regard to the overall effect of cooperative learning on student achievement, the primary difference between the current study and the Qin, et al. (1995) review is that they did not identify the academic disciplines in which the original studies were conducted. Despite this fact, it can be concluded that the results of the overall effect reported in this review support those reported in the current study, that cooperative learning has a positive effect on student achievement. This review reported three moderator analyses that deserve mention, none of which have any impact on the current study. The first moderator analysis examined the type of problem that students were solving in their cooperative or competitive task. The authors found that students performed better on nonlinguistic problems, those that involve the use of symbols in the response, such as in mathematics, than on linguistic problems, those that involve the use of narrative in the response. The second moderator analysis assessed the impact of the clarity of the problem that students completed, either well-defined (with expectations clearly communicated), or ill-defined (with expectations being unclear). No difference in effect sizes was seen between these two types of problems. The third moderator analysis assessed differences in achievement based on the age of the study participants. The authors compared two groups of students, pre-school/elementary and secondary/adults, but found no difference in the effect sizes between these two groups. None of these moderators were identified in any of the studies

included in the current review, so the relevance of the Qin, et al. (1995) findings is limited.

The Johnson, et al. (2000) review did not report an overall mean effect size. However, they did report effect sizes for each type of cooperative learning intervention. These specific values have been listed previously in Chapter 2. The authors calculated effect sizes from 164 included studies, a considerably larger sample than the current study. The sample of studies displayed considerable heterogeneity, as evidenced by Q statistics and the range of effect size values (no I^2 value was reported). When CL was compared with competitive learning, effect sizes ranged from .18 to .67, indicating that CL had a positive impact on student achievement. When CL was compared to individualistic learning, effect sizes ranged from .13 to .91, indicating that CL again had a positive impact on student achievement. The effect size for the current review falls in the middle of the range of effect sizes in the Johnson, et al. (2000) review and therefore compares favorably with their findings. The Johnson, et al. (2000) review did not identify the academic discipline in which the original studies were conducted. As with the Qin, et al. (1995) review, this review supports the conclusions of the current review that cooperative learning has a positive effect on student achievement.

The Springer, et al. (1999) review reported an overall mean effect size of .42 from 9 included studies that assessed the impact of cooperative learning on student achievement in science disciplines. The authors reported a range of effect sizes of -.25-1.5, but did not report a z-value with associated p-value representing statistical significance. In addition, the sample of studies displayed significant heterogeneity (Q = 23.59, $p < .05$; no I^2 value was reported). The overall effect size reported is similar in

value to that reported in the current review and is more comparable than the two reviews mentioned above since Springer, et al. (1999) examined the effect of cooperative learning on science achievement. Additionally, they only included studies in which the participants were undergraduate students, a group that was included in the current review. The other 2 reviews included studies in which the grade level of the participants ranged from pre-school to post-secondary. The major criticism of their review is that the sample size of included studies was only 9. The statistical power of a meta-analysis in which fewer than 10 studies are included is suspect (Borenstein, et al., 2009), thus calling the findings and comparison with the current review into question. Colliver, Feltovitch, and Verhulst (2003) published a critique of the Springer, et al. (1999) review, calling into question the results that were reported based upon the notion that only four of the nine studies analyzed were truly small-group learning, with only three providing reliable data for effect size calculation. They go on to state that the remaining five studies show primarily a null effect. Based on this critique as well, the results of the Springer, et al. (1999) review, could be called into question.

Bowen (2000) conducted a meta-analysis on the effect of cooperative learning on student achievement in chemistry. In his review, he reported an overall mean effect size of .37 (SD = .39) based on 30 effect sizes calculated from 15 studies. The range of effect sizes was less than -.40 to 1.0 based on a frequency distribution reported in the review. No heterogeneity statistics were reported, nor was a z-value with associated p-value or confidence interval reported. Despite these shortcomings, it is safe to conclude that cooperative learning had a positive effect on student achievement in chemistry. The

overall mean effect size value reported in the Bowen (2000) review is similar in value to that reported in the current review, thus supporting the conclusions of the current review.

Scott, et al. (2005) performed a meta-analysis assessing the impact on student achievement in science of a variety of instructional interventions in comparison to traditional instruction. One of the interventions they analyzed was collaborative learning strategies, or cooperative learning. They reported an overall mean effect size of .958 (CI: .777-1.14; $t = 10.41, p < .05$) for 3 included studies. The 3 studies exhibited significant heterogeneity ($Q = 20.865$; no I^2 value reported). Although the effect size reported in this review supports the results of the current review, the sample size of 3 is quite small and calls into question any conclusions regarding the impact of the intervention on student achievement in science.

Conclusions Related to Overall Effect

Based on comparisons with other meta-analyses conducted over the past 15 years, the current review compares favorably with regard to the overall effect of cooperative learning on student achievement. Although sample sizes in two of the reviews were somewhat small and two of the larger meta-analyses did not identify the academic discipline in which the included reviews were conducted, all five reviews reported that CL increased student achievement, thus supporting similar findings in the current review.

Study Inclusion Methodology Comparisons

The initial pool of studies was acquired through electronic database searching using detailed search criteria, reviews of annotated bibliographies, and reviews of reference lists from previous meta-analyses. As described previously, the current review employed a tri-level screening protocol using methods similar to those found in Lipsey

and Wilson (2001), with each level increasing in the specificity of the inclusion criteria. First, titles and abstracts were screened using the rubric in Appendix A. The reliability of the rubric as a screening tool was analyzed, as mentioned in Chapter 3 and 4. Once all titles and abstracts were screened, full-text copies of all remaining studies were acquired from various sources. The full-text of all remaining studies was screened for inclusion using the protocol in Appendix B. All remaining studies were then screened for final inclusion in the meta-analysis using the protocol in Appendix C. Once the final pool of studies was identified, statistics necessary for effect size calculation were extracted from each study.

Four of the five meta-analyses identified for comparison did not describe database searching in any great detail except the Scott, et al. (2005) review, but they never used the phrases "cooperative learning" or "collaborative learning" in any of their searches. Moreover, none of the reviews except Scott, et al. (2005) provided any specific information regarding the screening process to determine inclusion or exclusion of studies, and based on the small sample of studies in Scott, et al. (2005), any conclusions from their results are suspect anyway.

Conclusions Related to Study Inclusion Methodology

A review of the five meta-analyses used for comparison shows either an omission or lack of detail in study inclusion methodology. The current review developed the inclusion protocols and criteria *a priori* in order to ensure that the any results obtained and conclusions reported could be substantiated and supported by sound meta-analytic methodology. Based on these characteristics, the current review exceeds these other five studies in terms of study acquisition and inclusion methodology and criteria. This high

level of internal consistency in methodology further supports the findings and conclusions of the current meta-analysis that have been reported previously.

Intervention Comparisons

In the current review, the intervention of interest, cooperative learning, was initially defined as any intervention in which students worked together toward a common goal on academic assignments. This is a definition that is widespread in the literature. Further research on cooperative learning revealed that many different types exist. As a result, ten specific types of cooperative learning taken from Johnson, et al. (2000) were included in the study screening protocol. Each of these types of CL was further researched so that a complete understanding of the structure and delivery of each could be understood. In all of the studies, the control group experienced either traditional/lecture instruction or instruction as usual, both of which could be defined as individual instruction for the purpose of comparison.

A moderator analysis was done to examine if there were differential effects based on the specific characteristics and implementation of the intervention on student science achievement in comparison to instruction in the control condition. Fifteen studies did not specify a particular CL intervention, and 15 did specify the intervention. In the 15 that did not identify a specific CL intervention, the authors reported that the students worked in groups on the specific assignments, but did not provide any further details regarding the structure of the activities or intervention. The moderator analysis showed that the structured intervention had a somewhat larger effect on student achievement in science than the unstructured intervention. Both effect sizes were statistically significant. The difference between the two overall mean effect sizes was .099, which is approximately .1

standard deviations difference. Based on this value it can be concluded that the structure or specificity and delivery of the intervention had a differential effect on student achievement in science. The structured intervention had a greater impact on student achievement than did the unstructured intervention when compared to instruction in the control condition. It is important to note that the effect sizes from studies that identified a specific CL intervention were pooled since no specific CL type was identified more than four times, with some being identified only once, making comparisons between specific types of CL difficult based on a lack of statistical power (Borenstein, et al., 2009).

The Qin, et al. (1995) review did not identify specific types of cooperative learning or provide any detailed information regarding how the intervention was delivered in the studies they included in their meta-analysis. They distinguished between cooperative and competitive academic efforts (rather than individual instruction) and examined differences in student achievement based on these variables. Based on the characteristics of their review, comparisons regarding the impact of the structure, specificity, or delivery of the intervention on student achievement cannot be made with the current review.

The Johnson, et al. (2000) review identified 10 specific types of CL interventions and reported overall mean effect sizes for eight of them. In this review, they compared the achievement of students who experienced the intervention with those that either experienced competitive or individual learning environments. Studies included in the current meta-analysis did not identify competitive learning environments as the comparison factor, but identified individual learning as the comparison factor. Based on this characteristic, the current review will only be compared to the results of individual

versus cooperative instruction in the Johnson, et al. (2000) review. As mentioned previously, when pooled, studies that used a specific CL format reported a higher effect on student achievement in the current review. When disaggregated, the groups of studies that identified the CL format show varying effects, some statistically significant and others not statistically significant. Since a meta-analysis pools data from multiple studies, this observation has no impact on the conclusions regarding the efficacy of the structured intervention (Borenstein, et al., 2009). Moreover, since the largest number of studies that identified the CL intervention in the current meta-analysis is four, conclusions based on specific forms of CL are suspect due to a lack of statistical power. In the Johnson, et al. (2000) review, only three specific CL interventions included more than eight studies in the analysis. These were "Learning Together" (57 studies; *ES* = 1.04), "Jigsaw" (14 studies; *ES* = .29), and "Constructive Controversy" (11 studies; *ES* = .91). The pooled effect size in the current review falls into the range of values for the three identified interventions in the Johnson, et al. (2000) review. These values lend support to the conclusion in the current review that structured CL interventions have a greater impact on student achievement than unstructured CL interventions.

None of the other three meta-analyses (Springer, et al., 1999; Bowen, 2000; & Scott, et al., 2005) identified specific types of CL interventions. Consequently, it is not possible to make any comparisons related to type of CL intervention and affect on student achievement with the current review.

Conclusions Related to Intervention Characteristics

Only one of the previous meta-analyses (Johnson, et al., 2000) identified specific types of CL in their analysis. This review included a large sample of studies in its

analysis. Despite the fact that the other meta-analyses did not identify the CL intervention, the results of this particular review support the conclusions of the current study that specific, structured types of CL interventions have a greater impact on student achievement than unstructured CL interventions.

Nature of the Sample and Setting Comparisons

The grade levels of students who represented the experimental and control samples in the individual studies that were included in the current review were secondary (middle, junior high, and high school students, with grade 6 included only if it was not part of an elementary school) and early post-secondary (community college students or university freshmen and sophomore students). All of these grade levels were represented in differing frequencies in the sample of studies included in the meta-analysis. The science discipline in which the intervention was implemented defined the setting of each study. In the current review, five different science disciplines were represented in the sample of studies (Biology, Chemistry, Earth Science, General Science, and Physics).

A moderator analysis was undertaken to examine if there were differential effects based on the science discipline in which the study was conducted. Studies done in biology and chemistry classes were the most numerous. In biology (11 studies), the overall mean effect size of .345 was statistically significant and represented the highest significant effect size for all disciplines. The overall mean effect size for chemistry (11 studies) was .235 and was also statistically significant. In earth science classes (4 studies), the overall mean effect size of .187 was considerably lower than in biology and chemistry, but was statistically significant also. In physics classes (3 studies), the effect size of .632 was much higher than all other disciplines, but was not statistically

significant. Only one study was done in a general science class, with a statistically significant effect size of .432. Based on these results, it could be concluded that CL works better in biology classes than chemistry, but due to the small sample sizes, comparisons with the other science disciplines are questionable. A comparison of CL methods utilized in each study reveals very little to describe why the effect sizes for each discipline are different. In biology, 7 of the 11 studies used unstructured CL, and in chemistry, six of the 11 studies did the same. All four studies done in earth science used a structured CL format, while two of the three studies in physics did the same. Consequently, it is difficult to make any conclusions regarding the differential success of CL in various disciplines in relation to the structure of the CL intervention.

Qin, et al. (1995) disaggregated data based on the age of students. They identified students as either "secondary" students or "adults" and pooled the data. They reported an overall mean effect size for this group of students of .60 from 33 studies. They reported no significant difference between the achievement of elementary students and secondary and adult students. The findings of this review support the findings of the current review that CL interventions have a positive impact on the achievement of secondary and early-post-secondary students. They did not identify the setting of each study by the academic discipline, so no comparison on this factor can be made with the current review.

Johnson, et al. (2000) included studies from elementary through post-secondary grade levels, but did not report specific effect sizes for each range of grade level. Additionally, they did not identify the specific academic discipline in which the included studies occurred. As a result, it is difficult to make any comparisons regarding these study characteristics with the current meta-analysis.

The Springer, et al. (1999) meta-analysis included only studies that were conducted in 2-year or 4-year higher education institutions and reported effect sizes for each educational level. In addition, they reported effect sizes based on three academic disciplines in which a study occurred (Science, Mathematics, and Allied Health). As mentioned previously, nine of the studies included in this meta-analysis were conducted in science disciplines. As mentioned prior, the effect size for science was .42, but only included nine studies. Unfortunately, specific science disciplines were not identified, nor was the educational level disaggregated in each of the academic disciplines. Based on these observations, it is difficult to make any comparisons with the current review based on the educational level of the students or the science discipline in which the study took place.

The Bowen (2000) meta-analysis focused specifically on the achievement of high school and college students in chemistry. He reported an effect size of .37 from 15 studies, which is higher than the effect size for chemistry classes in the current study (.235), but did not disaggregate effect size data based on educational level of the students. His findings support those of the current review that cooperative learning has a positive impact on the achievement of students in chemistry, but conclusions based on educational level cannot be made.

The Scott, et al. (2005) review identified specific science disciplines, but did not disaggregate them for the cooperative learning intervention. As stated previously, they only included three studies in their analysis of CL, making the results of this review and any comparisons to the current review suspect.

Conclusions Related to Sample and Setting Comparisons

All five of the meta-analyses included in this comparison included studies in which the impact of the CL intervention on student achievement was measured in either secondary or early post-secondary settings, or both, and three identified science classes as the setting for the studies. Unfortunately, none of them disaggregated effect size data using both grade level and academic discipline, making comparisons with the current review based on these characteristics inconclusive.

Comparisons Based on Study Participant Characteristics

Gender

A moderator analysis was undertaken in the current review to examine if there were differential effects of the intervention on male and female students. Four of the 30 included studies disaggregated achievement data on the basis of gender. Based on these four studies, a differential effect was observed, with the intervention showing a positive and statistically significant effect on the achievement of male students and a null effect on the achievement of female students, with zero in the confidence interval. While the effect size calculated for males was statistically significant, making any generalizations regarding differential achievement due to gender from the data is suspect due to the small sample of studies included in this particular analysis. Borenstein, et al. (2009) state that under the random effects model, if the number of studies in a meta-analysis or moderator analysis is low (less than 10), the statistical power of the analysis will remain low regardless of the total participant sample size across studies. Based on this fact and the fact that only four studies were included in the current analysis, making a claim that there was a differential effect of CL on males compared to females is somewhat questionable.

Only one of the five meta-analyses considered for comparison did a moderator analysis on gender (Springer, et al., 1999) and they did not report any significant differences in achievement between males and females who experienced cooperative learning instruction. Additionally, the authors did not identify if the studies were performed in science classrooms or in the other disciplines they examined. Based on the above observations, not comparison between the current review and the comparison reviews can be made with regard to a differential effect of the intervention based on the gender of the participants.

Ability Level

A moderator analysis was completed to examine if there were differential effects of the intervention on students of different ability levels, low, medium, and high. Four of the thirty included studies disaggregated achievement data on the basis of student ability level. The overall mean effect size for each ability level showed a null effect since the confidence interval for each included zero and none were statistically significant based on their *z*-values. Despite this finding, any generalization describing the effect of the intervention on students of differing ability levels is suspect due to the small number of studies included in this analysis.

None of the five meta-analyses selected for comparison did moderator analyses on the ability level of the student participants in included studies, so no comparison can be made with the current review.

Comparisons Based on Characteristics of Included Studies

Methodology

A moderator analysis was completed to examine if there were differential effects based on the experimental design and methodology employed in the included studies. Cluster randomized studies and studies that were quasi-experimental without subject matching, both showed statistically significant, positive effects, while quasi-experimental studies with subject matching showed a null effect. While the cluster randomized design and quasi-experimental with subject matching designs are typically viewed as more robust methodologies than quasi-experimental designs (IES, 2006), the fact that the quasi-experimental without subject matching design had the largest effect seems puzzling.

Further analysis of the 13 studies that used the quasi-experimental without subject matching design is necessary to explain the effect size value associated with this methodology. Seven of the 13 studies were performed in biology classrooms, which showed a higher effect size than any other discipline. By comparison, two of these studies were performed in chemistry classrooms and four were performed in earth science classrooms. The disproportionate distribution of studies that were performed in biology classrooms could explain why this particular methodology had a higher effect size than the other two. Intervention type would appear to have little or no effect on the effect size, and six studies that identified the CL type and seven studies that did not. These deeper analyses of the methodology effect sizes provide a statistical explanation for the larger value of the effect size for the quasi-experimental without subject matching design rather than a substantive one. Since this moderator analysis included small numbers of studies

in each category, any conclusions that differentiate the effect based on study methodology could be called into question.

The Springer, et al. (1999) meta-analysis did a methodology moderator analysis as well, but identified one sample, single group, pretest/posttest designs versus two sample, treatment/control group designs. They did report a higher effect size in the treatment/control design, but as mentioned previously, did not disaggregate the study data based on science discipline. None of the other four reviews did a methodology moderator analysis. These observations make comparisons between the current review and previous reviews based on the methodology of included studies questionable.

Reliability Testing of Assessment Instrument

A moderator analysis was completed to determine if there were differential effects based on whether the author(s) of included studies conducted reliability testing on the assessment instrument used to measure the student achievement outcome. Only one review (Scott, et al., 2005) examined the reliability testing of the assessment instrument used in included studies, but they did not disaggregate this characteristic based on the use of the CL intervention. Based on these observations, no comparison regarding reliability testing of the assessment used in included studies can be made with the current review.

All the studies included in the meta-analysis reported treatment (CL) versus control groups, with 19 studies (59%) reporting that control groups experienced traditional, lecture instruction, and 13 studies (41%) reporting that subjects in the control groups worked individually on the educational activities versus working in groups.

Significance of the Reported Results

The meta-analysis reported here adds to the research base by providing current and more methodologically sound inclusionary criteria than prior reviews describing the positive effect of cooperative learning on student achievement in science. The tri-level nature of the review screening provides a higher level of confidence in the results in comparison to the Bowen (2000) and Springer, et al. (1999) reviews, as well as other meta-analyses that have been assessed. Moreover, these reviews reported no publication bias analyses, which, as has been reported for the current review, further support the findings. No other meta-analyses on the impact of cooperative learning on student achievement in science have been performed since the Bowen (2000) study, so the current review provides a much needed synthesis and update to the research base on this intervention.

The results of this review also have a number of other implications which will be addressed here. With regard to teacher education programs, the positive effect that cooperative learning has on student achievement in science provides empirical evidence that this type of intervention should be examined in detail in pre-service teaching methods courses. The history, development, and implementation of this particular intervention should be infused into the curriculum of these types of courses, particularly in science teaching methods courses. Internationally, the students in the United States have continued to underperform in science when compared to other countries (TIMSS, 2003). The results of this review report a positive impact of cooperative learning on student achievement in science. This being the case, it stands to reason that if this intervention is included in the curriculum of science methods courses, then teachers

coming out of teacher preparation programs in which they learn about CL will be more likely to use it in their classrooms. As a result, student achievement in science in the United States could improve in the coming years.

Other implications of the results reported here relate to how journal editorial boards review and select original research for publication. Most research that is published shows positive effects (Lipsey & Wilson, 2001). In order to truly assess the impact of any instructional intervention, all research results should be published, including those that report null or even negative results. Some of the studies included in this review reported both, but the preponderance of published studies report positive results. As reported previously, one of the major criticisms of meta-analysis is that the results do not include all potential studies that may have been performed. Researchers must be willing to submit studies that report negative or null effects for publication. Journal editors should be more explicit in soliciting these types of studies for submission. If researchers submit and subsequently publish null or negative result studies, and journal editors actively solicit these types of studies, this criticism could be addressed.

Another issue for journal editorial boards to consider relates to the quality of the research methodology in published studies. As was reported earlier, 19 of the studies included in the final review were of a quasi-experimental design. While this methodology has its place in social science research involving classroom-based interventions, many confounding variables can affect the results. Journal editorial boards could suggest that classroom-based intervention studies utilize a cluster randomized methodology, or at the very least, a quasi-experimental with subject matching methodology in order to account for confounding variables. Moreover, these boards

could use a screening protocol similar to the one reported here to assess the research methodology in submitted manuscripts. A protocol such as this would allow reviewers and editors to assess study methodology in a more critical and reliable manner. One of the criticisms of meta-analysis has been that if a review includes unsound research, then the results are also unsound (Lipsey & Wilson, 2001). If editorial boards require or at least suggest that more reliable methodologies be used, and assess methodology using some type of screening protocol, this particular criticism of meta-analytic results could be addressed.

One of the major obstacles encountered in this review involved the statistics that were reported in the included studies. Many studies reported means, standard deviations, and numbers of subjects (*N*) in control and treatment samples, but equally as many reported other forms of statistics, without reporting numbers of subjects. Another suggestion to editorial boards would be that studies be required to report means and standard deviations, as well as the numbers of subjects in both treatment and control groups. The effect size statistic can be calculated from other statistical values, but in some cases in this review, the *N* for each sample had to be estimated because these values were not reported directly in the study.

Suggestions for Future Research

The results of this meta-analysis support the notion that cooperative learning instruction improves student achievement in science when compared to traditional instruction. One of the original intents of this research project was to examine and analyze the effect that CL had on the science achievement of female students and nonmainstream (Lee & Luykx, 2006) students who have traditionally underperformed in

science (NAEP, 2006) in comparison to mainstream students. Nonmainstream students are those whose achievement has lagged behind that of other students and typically include ethnic minorities, students from lower socioeconomic classes, and students whose primary language is one other than English. Mainstream students are typically those who are Caucasian, at middle and higher socioeconomic levels, and whose primary language is English (Lee & Luykx, 2006). Unfortunately, none of the studies included in this analysis disaggregated data based on student ethnicity or socioeconomic status in such a way as to elicit any conclusions. Only four included studies disaggregated data based on gender of the students. Based on this information, no valid conclusions can be made regarding the impact of cooperative learning instruction on the achievement of students who have traditionally underperformed in science. The results of the current review call for more research into the effect that cooperative learning instruction has on the achievement of students who have traditionally underperformed in science. Many of the studies reviewed during the title and abstract, and intermediate full-text coding included students in these groups, but the majority of these studies focused on student attitudes toward science and not science achievement. While student attitude toward science can play an important role in a student's achievement in science, any conclusive causal relationship is yet to be reported in the literature. The studies included in this report that examined gender effects showed that males benefit more from CL instruction than, females, but due to the small sample size (four studies), any generalizations to the larger population is suspect. More reliable and valid studies such as those included here examining gender effects need to be conducted before any significant conclusions regarding gender differences in science achievement can be made. Since recent reports

on the performance of students in the United States have shown that underrepresented students tend to underperform in science (NAEP, 2006), studies that address the impact of cooperative learning on the science achievement of these students in comparison to mainstream students is necessary.

With regard to future meta-analytic research, the author strongly suggests that anyone planning on conducting a meta-analysis purchase and use meta-analytic software. This type of software application provides the user with a number of options regarding how to analyze data. This software allows the used to enter any type of statistical data, provided the individual has sample numbers. These data are then converted to effect size statistics and allow the researcher to perform any number of different analyses. All statistical analysis performed in this review were done using CMA®, but many other brands of software are available.

Limitations of the Current Review

The meta-analytic methodology that was developed and followed throughout this review was very sound, providing results that are valid and conclusive. Nonetheless, a variety of limitations to a review such as the one reported here warrant discussion. One limitation of this review that could be reported relates to the initial database searches. The databases that were used to locate and acquire studies for the meta-analysis are widely utilized in academe; however, many of them do not locate studies from countries other than the United States. Searches of international databases may have provided citations that were not otherwise located by the databases that were used which could have had a positive or negative effect on the overall mean effect size that was calculated.

Another limitation relates to the actual impact of cooperative learning on student achievement in science with regard to student characteristics. While it has been concluded that the intervention has a positive impact on student achievement in science across grade levels and ages, the cognitive and emotional development of students across ages can be quite diverse (Piaget, 1950). Older students tend to be more mature from a developmental perspective, so concluding that cooperative learning has an equal effect across grade levels and ages could be questioned.

A third limitation to be considered relates to the fidelity of the implementation of the intervention in the included studies. Approximately half of the studies included in this review identified a specific CL intervention, while the other half did not. The description of the implementation of the intervention in each of the included studies was quite diverse. Consequently, it is difficult to determine if the intervention was implemented in its intended format if it was defined as a specific type of cooperative learning. Moreover, if the intervention was not specifically identified, it is very difficult to determine how it was delivered and implemented in a particular study unless it was specifically reported in the original research study. Based on these observations, there could have been a differential effect of CL based more on the quality and fidelity of the implementation of the intervention rather than the intervention itself.

A similar limitation of the review relates to the fidelity of the assessment instrument and its ability to truly assess student achievement. As mentioned previously, 19 of the included studies performed some sort of reliability testing on the assessment instrument that was used to measure student achievement. Within this group, some studies used statistical methods of analysis, while others used expert reviews to determine

instrument reliability. Eleven studies did no reliability testing of any type on the assessment instrument, possibly calling into question the notion that the instrument truly measured the outcomes that were intended. Based on this observation, outcome results from studies in which no reliability testing was done on the assessment instrument could be biased, and therefore impact the overall mean effect size that was calculated.

Lastly, a final judgment on the impact of cooperative learning on student achievement in science is fully dependent on the integrity of the research methodology in each of the included studies. As described previously, the methodology employed in the included studies was quite diverse, with three different methodologies being employed. Thirteen of the included studies utilized a quasi-experimental without subject matching approach. This particular research methodology is considered to be of lower quality according to the Institute of Education Sciences, which provides a hierarchy entitled the "Levels of Evidence for Assessing Program Effectiveness" (Cobb, personal communication, May 22, 2006). Based on this observation, the results from these types of studies could introduce a sample bias that could skew the overall mean effect size in either a positive or negative direction. Additionally, how the research methodology was described in each of the included studies varied widely, with some studies providing very detailed descriptions and others providing vague descriptions. Consequently, it could be argued that the research methodology utilized in each of the included studies could have had a differential effect on the outcomes of the study, which in turn could have impacted the value of the overall mean effect size.

References

American Psychological Association. (2001). *Publication manual of the American Psychological Association (5th ed.)*. Washington, D. C.: American Psychological Association.

Aronson, E., Blaney, N., Sikes, J., Stephan, C., & Sapp, M. (1978). *The jigsaw classroom*. Beverly Hills, CA: Sage.

Bandura, A. (1977). *Principles of behavioral modification*. New York: Holt, Reinhart, & Winston.

Bilgin, I. (2006). The effects of pair problem solving technique incorporating Polya's problem solving strategy on undergraduate students' performance in chemistry. *Online Submission; Revista de Educacion en Ciencias (Journal of Science Education), 7*(2), 101-106.

Bilgin, I., & Geban, O. (2006). The effect of cooperative learning approach based on conceptual change condition on students' understanding of chemical equilibrium concepts. *Journal of Science Education and Technology, 15(1)*, 31-46.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H. (2009). *Introduction to Meta-Analysis*. Chichester, United Kingdom: Wiley.

Bowen, C. W. (2000). A quantitative literature review of cooperative learning effects on high school and college chemistry achievement. *Journal of Chemical Education, 77*(1), 116-119.

The Campbell Collaboration. (2006). Retrieved January 30, 2007 from http://www.campbellcollaboration.org.

Chang, C.-Y., & Mao, S.-L. (1999b). The effects on students' cognitive achievement when Using the cooperative learning method in earth science classrooms. *School Science and Mathematics, 99*(7), 374-379.

The Cochrane Collaboration. (2006). Retrieved January 30, 2007 from http://www.cochrane.org.

Cohen, E. (1986/1994). *Designing groupwork: strategies for the heterogeneous classroom*. New York: Teachers College Press.

Cohen, J. (1998). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Colliver, J.A., Feltovich, P.J., & Verhulst, S.J. (2003). Small group learning in medical education: a second look at the Springer, Stanne, and Donovan meta-analysis. *Teaching and learning in medicine, 15*(1), 2-5.

Cook, S. (1969). Motives in a conceptual analysis of attitude-aelated behavior. In W. Arnold and D. Levine (Eds.), *Nebraska Symposium on Motivation* (Vol. 17). Lincoln: University of Nebraska Press

Cooper, H. & Hedges, L.V. (1994). *The handbook of research synthesis*. New York: Sage.

Deutsch, M. (1949). An experimental study of the effects of cooperation and competition upon group processes. *Human Relations, 2*, 199-232.

DeVries, D., & Edwards, K. (1974). Student teams and learning games: Their effects on cross-race and cross-sex interaction. *Journal of Educational Psychology, 66*(5), 741-749.

Duval, S. and Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89-98.

Glass, G. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.

Hanze, M., & Berger, R. (2007). Cooperative learning, motivational effects, and student characteristics: an experimental study comparing cooperative learning and direct instruction in 12th grade physics classes. *Learning and Instruction, 17*(1), 29-41.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107-128.

Huedo-Medina, T.B., Sánchez-Meca, J., & Marín-Martínez, F. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods, 11*(2), 193-206.

Higgins, J.P.T., & Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539-1558.

Hopewell, S., McDonald, S., Clarke, M., & Egger, M. (2002). Grey literature in meta-analyses of randomized trials of health care interventions. *The Cochrane Database of Methodology Reviews, 4*, Article Number MR000010.

Howard, B.C. (1996, February). *A meta-analysis of scripted cooperative learning*. Paper presented at the Eastern Educational Research Association annual meeting In Boston, MA.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage Publications.

The Institute of Education Sciences. (2007). Retrieved January 31, 2007 from http://ies.ed.gov/timss.

Johnson, D.W., & Johnson, R. (1974). Instructional goal structure: cooperative, competitive, or individualistic. *Review of Educational Research, 44*, 213-240.

Johnson, D.W., & Johnson, R. (1979). Conflict in the classroom: Controversy and learning. *Review of Educational Research, 49*, 51-70.

Johnson, D.W., & Johnson, R. (1989). *Cooperation and competition: theory and research.* Edina, MN: Interaction Book Company.

Johnson, D. W. & Johnson, R. T. (1999). Making Cooperative Learning Work. *Theory into Practice, 38*(2), 67-73.

Johnson, D.W., & Johnson, R.T. (1999). *Learning together and alone: Cooperative, competitive, and individualistic learning* (5th ed.). Boston: Allyn & Bacon.

Johnson, D.W., & Johnson, R. (2002). Ensuring diversity is positive: cooperative community, constructive conflict, and civic values. In Thousand, J.S., Villa, R.A., & Nevin, A.I. (Eds.; 2nd ed.), *Creativity and collaborative learning: the practical guide to empowering students, teachers, and families* (pp. 197-208). Baltimore, MD: Paul H. Brooks.

Johnson, D.W., Johnson, R. & Holubec, E. (1993). *Cooperation in the classroom* (6th ed.). Edina, MN: Interaction Book Company.

Johnson, D.W., Johnson, R.T., & Stanne, M.B. (2000). Cooperative learning methods: A meta-analysis. *Cooperative Learning Center, University of Minnesota.* Retrieved November 10, 2007 from http://www.co-operation.org/pages/cl-methods.html.

Johnson, D.W., Maruyama, G. Johnson, R., Nelson, D., & Skon, L. (1981). Effect of cooperative, competitive, and individualistic goal structures on achievement: a meta-analysis. *Psychological Bulletin, 89*, 47-62.

Iyenger, S., & Greenhouse, J.B. (1988). Selection models and the file drawer problem (with discussion). *Statistical Science, 3*, 109-135.

Kagan, S. (1992). *Cooperative learning resources for teachers.* San Juan Capistrano, CA: Resources for Teachers.

Kagan, S. (1994). *Cooperative learning.* San Clemente, CA: Resources for Teachers.

Kulik, C-L.C. & Kulik. J.A. (1982). Effects of ability grouping on secondary students: a meta-analysis of evaluation findings. *American Educational Research Journal, 19*(3), 415-428.

Lee, O. & Luykx, A. (2006). *Science Education and Student Diversity: Synthesis and Research Agenda*. New York, NY: Cambridge University Press.

Lewin, K. (1935). *A dynamic theory of personality*. New York: McGraw-Hill.

Light, R. & Pilleme, D. (1984). Summing up: The science of reviewing research. Cambridge, MA: Harvard University Press.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Miles, M.B., & Huberman, A.M. (1994). *Qualitative Data Analysis*. Thousand Oaks, CA: Sage.

Morgan, G.A. , Gliner, J.A., & Harmon, R.J. (2006). *Understanding and Evaluating Research in Applied and Clinical Settings*. Mahwah, NJ: Lawrence Erlbaum Associates.

Murray, F.B. (2002). Why understanding the theoretical basis of cooperative learning enhances teaching success. In Thousand, J.S., Villa, R.A., & Nevin, A.I. (Eds.; 2nd ed.), *Creativity and collaborative learning: the practical guide to empowering students, teachers, and families* (pp. 175-180). Baltimore, MD: Paul H. Brooks.

National Center for Education Statistics. (2003). Trends in international math and science study. Retrieved January 20, 2008 from http://nces.ed.gov/timss/results03.asp.

The Nation's Report Card. (2006). National Assessment of Educational Progress (NAEP), United States Department of Education. Retrieved January 29, 2007 from http://nationsreportcard.gov/science_2005.

Pennisi, E. (2005). How did cooperative behavior evolve? *Science, 309*, 93.

Pepitone, E.A. (1985). Children in cooperation and competition: antecedents and consequences. In Slavin, R., Sharan, S., Kagan, S., Hertz Lazarowitz, R., Webb, C., & Schmuck, R. (Eds.), *Learning to Cooperate, Cooperating to Learn* (pp.17-65). New York: Plenum Press.

Petitti, D. B. (2000). *Meta-analysis, decision analysis, and cost-effectiveness analysis: methods for quantitative synthesis in medicine* (2nd ed.). New York, NY: Oxford University Press.

Petticrew, M. & Roberts, H. (2006). *Systematic reviews in the social sciences: a practical guide.* Malden, MA: Blackwell.

Piaget, J. (1950). *The psychology of intelligence.* New York: Harcourt.

Popham, W.J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership, 56*(6), 8-15.

Programme for International Student Assessment. (2006). Retrieved January 22, 2008 from http://www.pisa.oecd.org/document/2/0,3343,en_32252351_32236191_39718850_1_1_1_1,00.html.

Qin, Z, Johnson, D.W., & Johnson, R. (1995). Cooperative versus competitive efforts and problem solving. *Review of Educational Research, 65*(2), 129-143.

Sapon-Shevin, M., Ayres, B.J., & Duncan, J. (2002). Cooperative learning and inclusion. In Thousand, J.S., Villa, R.A., & Nevin, A.I. (Eds.; 2nd ed.), *Creativity and collaborative learning: the practical guide to empowering students, teachers, and families* (pp. 209-222). Baltimore, MD: Paul H. Brooks.

Scott, T. P., Tolson, H., Schroeder, C., Lee, Y. H., Huang, T. Y., Hu, X., & Bentz, A. (2005). *Meta-analysis of national research regarding science teaching.* Texas A & M University Center for Mathematics and Science Education.

Shachar, H., & Fischer, S. (2004). Cooperative Learning and the Achievement of Motivation and Perceptions of Students in 11th Grade Chemistry Classes. *Learning and Instruction, 14*(1), 69-87.

Sharan, S., & Sharan, Y. (1976). *Small group teaching.* Englewood Cliffs, NJ: Educational Technology Publications.

Sharan, S., & Sharan, Y. (1992). *Group investigation: Expanding cooperative learning.* New York: Teacher's College Press.

Skinner, B.F. (1968). *The technology of teaching.* New York: Appleton-Century-Crofts.

Slavin, R. (1978). Student teams and achievement divisions. *Journal of Research and Development in Education, 12*, 39-49.

Slavin, R. (1980). Cooperative learning. *Review of Educational Research, 50*, 315-342.

Slavin, R. (1983). *Cooperative learning* (1st ed.). New York: Longman.

Slavin, R. (1986). *Using student team learning* (3rd ed.). Baltimore, MD: Johns Hopkins University.

Slavin, R. (1995). *Cooperative learning* (2nd ed.). Needham Heights, MA: Allyn and Bacon.

Slavin, R., Leavey, M., & Madden, N. (1982). *Team-Assisted Individualization: Mathematics Teacher's Manual*. Johns Hopkins University, Center for Social Organization of Schools.

Slavin, R., Leavey, M., & Madden, N. (1986). *Team Accelerated Instruction: Mathematics*. Watertown, MA: Charlesbridge.

Slavin, R., Sharan, S., Kagan, S., Hertz Lazarowitz, R., Webb, C., & Schmuck, R. (1985). *Learning to cooperate, cooperating to learn*. New York: Plenum Press.

Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: a meta-analysis. *Review of Educational Research, 69*(1), 21-51.

Stevens, R., Madden, N., Slavin, R., & Farnish, A. (1987). Cooperative integrated reading and composition: Two field experiments. *Reading Research Quarterly, 22*, 433-454.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wikipedia. (2008). Retrieved September 17, 2008 from http://www.wikipedia.org.

References for Studies Included in Final Meta-Analysis

Abobaker, N. M. (1995). *An investigation of cooperative learning for teaching large high school science classes in Yemen.* Unpublished doctoral dissertation, University of Idaho, Moscow.

*Acar, B., & Tarhan, L. (2007). Effect of Cooperative Learning Strategies on Students' Understanding of Concepts in Electrochemistry. *International Journal of Science and Mathematics Education, 5*(2), 349-373.

Balfakih, N. M. A. (2003). The Effectiveness of Student Team-Achievement Division (STAD) for Teaching High School Chemistry in the United Arab Emirates. *International Journal of Science Education, 25*(5), 605-624.

Banerjee, A. C., & Vidyapati, T. J. (1997). Effect of lecture and cooperative learning strategies on achievement in chemistry in undergraduate classes. *International Journal of Science Education, 19*, 903-910.

Bilgin, I. (2006). The Effects of Pair Problem Solving Technique Incorporating Polya's Problem Solving Strategy on Undergraduate Students' Performance in Chemistry Access ERIC: FullText. *Online Submission; Revista de Educacion en Ciencias (Journal of Science Education), 7*(2), 101-106.

Bradley, A. Z., Ulrich, S. M., Jones, M., Jr., & Jones, S. M. (2002). Teaching the Sophomore Organic Course without a Lecture. Are You Crazy? *Journal of Chemical Education, 79*(4), 514-519.

Chang, C.-Y., & Mao, S.-L. (1999a). Comparisons of Taiwan science students' outcomes with inquiry-group versus traditional instruction. *Journal of Educational Research, 92*(6), 340-346.

Chang, C.-Y., & Mao, S.-L. (1999b). The Effects on Students' Cognitive Achievement When Using the Cooperative Learning Method in Earth Science Classrooms. *School Science and Mathematics, 99*(7), 374-379.

Chung-Schickler, G. C. (1998). *The effect of cooperative learning on the attitudes toward science and the achievement of students in a non-science majors' general biology laboratory course at an urban community college.* Unpublished doctoral dissertation, Florida International University, Miami.

De Paz, T. (2001). The Effectiveness of the Jigsaw Cooperative Learning on Students' Achievement and Attitudes toward Science. *Science Education International, 12*(4), 6-11.

Dori, Y. J., Yeroslavski, O., & Lazarowitz, R. (1995). *The Effect of Teaching the Cell Topic Using the Jigsaw Method on Students' Achievement and Learning Activity.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco, CA.

Faro, S., & Swan, K. (2006). An Investigation into the Efficacy of the Studio Model at the High School Level. *Journal of Educational Computing Research, 35*(1), 45-59.

Foley, K. E. (1995). *Cooperative learning and visual organizers: Effect on students' solving mole problems in chemistry.* Unpublished doctoral dissertation, Rutgers The State University of New Jersey, New Brunswick.

*Fontenot, D. W. (1995). *The effects of cooperative learning methods in conjunction with traditional lectures in seventh-grade earth science classes.* Unpublished doctoral dissertation, The University of Southern Mississippi, Hattiesburg.

Gayford, C. (1995). Science Education and Sustainability: A Case-Study in Discussion-Based Learning. *Research in Science and Technological Education, 13*(2), 135-145.

Hanze, M., & Berger, R. (2007). Cooperative Learning, Motivational Effects, and Student Characteristics: An Experimental Study Comparing Cooperative Learning and Direct Instruction in 12th Grade Physics Classes. *Learning and Instruction, 17*(1), 29-41.

Harskamp, E., & Ding, N. (2006). Structured Collaboration versus Individual Learning in Solving Physics Problems. *International Journal of Science Education, 28*(14), 1669-1688.

Jensen, M. S. (1996). Cooperative Quizzes in the Anatomy and Physiology Laboratory: A Description and Evaluation. *Advances in Physiology Education, 16*(1), S48-S54.

Jeon, K., Huffman, D., & Noh, T. (2005). The Effects of Thinking Aloud Pair Problem Solving on High School Students' Chemistry Problem-Solving Performance and Verbal Interactions. *Journal of Chemical Education, 82*(10), 1558-1564.

Lazarowitz, R., Baird, F. H., & Bowlden, V. (1996). Teaching biology in a group mastery learning mode: high school students' academic achievement and affective outcomes. *International Journal of Science Education, 18*, 447-462.

Lumpe, A. T., & Staver, J. R. (1995). Peer Collaboration and Concept Development: Learning about Photosynthesis. *Journal of Research in Science Teaching, 32*(1), 71-98.

Qualter, A., & Abu-Hola, I. R. A. (2000). Approaches To Teaching Science in the Jordanian Primary School. *Research in Science and Technological Education, 18*(2), 227-239.

Roy, H. (2003). Studio vs. Interactive Lecture Demonstration--Effects on Student Learning. *Bioscene, 29*(1), 3-6.

Sadler, K. C. (2002). *The effectiveness of cooperative learning as an instructional strategy to increase biological literacy and academic achievement in a large, nonmajors college biology class.* Unpublished doctoral dissertation, Tennessee State University, Nashville.

Schroeder, P. G. (1996). *Science as argument: A context using peer dyads to promote conceptual change among community college chemistry students.* Unpublished doctoral dissertation, Kansas State University, Manhattan.

Shachar, H., & Fischer, S. (2004). Cooperative Learning and the Achievement of Motivation and Perceptions of Students in 11th Grade Chemistry Classes. *Learning and Instruction, 14*(1), 69-87.

Snyder, T., & Sullivan, H. (1995). Cooperative and Individual Learning and Student Misconceptions in Science. *Contemporary Educational Psychology, 20*(2), 230-235.

Starr, E. M. (1995). Cooperative learning and its effects on geology achievement and science attitudes of preservice elementary-school teachers. *Journal of Geological Education, 43*, 391-394.

Tao, P.-K. (1999). Peer Collaboration in Solving Qualitative Physics Problems: The Role of Collaborative Talk. *Research in Science Education, 29*(3), 365-383.

Trautwein, S. N., Racke, A., & Hillman, B. (1996). Cooperative Learning in the Anatomy Laboratory. *Journal of College Science Teaching, 26*(3), 183-188.

Verdel, E. F. O. (1996). *Collaborative learning and computer-based instruction in introductory chemistry.* Unpublished doctoral dissertation, The University of Texas, Austin.

Werner, J. L., & Klein, J. D. (1999) *Effects of Learning Structure and Summarization during Computer-Based Instruction.* Paper presented at the national convention of the Association for Educational Communications and Technology, Houston, TX.

*These two studies were subsequently excluded from the final analysis.

APPENDIXES

Appendix A: Title and Abstract Screening Rubric

Appendix B: Full-Text Intermediate Screening Protocol

Appendix C: Full-Text Study Coding Form

APPENDIX A:

Title and Abstract Screening Rubric

APPENDIX A

Title and Abstract Screening Rubric

Question	Options	Definition
Is the year of publication between 1995-2007?	Yes = Y No = N Not sure = Y	The date of publication of the study falls into the range of 1995-2007.
Is the study published in English?	Yes = Y No = N Not sure = Y	The study must be published in English.
Is the study empirical in design?	Yes = Y No = N Not sure = Y	The study must report that quantitative data was collected.
Does the study identify cooperative learning as the intervention of interest?	Yes = Y No = N Not sure = Y	The study must report that some form of cooperative learning as the intervention that was being tested.
Is the educational level of the participants secondary or early post-secondary?	Yes = Y No = N Not sure = Y	The study participants must be considered secondary (not grade 6 elementary) or early post-secondary.

APPENDIX A (cont'd)

Title and Abstract Screening Rubric

Question	Options	Definition
Was the study performed in a science discipline?	Yes = Y No = N Not sure = Y	The study must have been done in a science discipline (biology, chemistry, earth and space science, physics).
Was student achievement the outcome of the study?	Yes = Y No = N Not sure = Y	The study must report that some form of student science achievement is the outcome.

APPENDIX B:

Full-text Intermediate Screening Protocol

APPENDIX B

Full-text Intermediate Screening Protocol

Section 1: Coder Information

1.1 Name of Coder _____

1.2 Date Study Coded _____

1.3 Study Identification Number _____

Section 2: Intervention(s)

The study must identify cooperative learning in name as the intervention of interest or must describe the intervention in such a way as to be considered cooperative learning according to the meta-analytic interest of the investigators.

2.1 Study identifies cooperative learning as the intervention of interest _____

If yes, indicate "1". If cooperative learning is not specified but the intervention aligns with the meta-analytic interest, indicate "2". If the intervention is not identified as cooperative learning and does not align with the meta-analytic interest, indicate "3".

2.2 Length of intervention _____

If intervention is 10 weeks or longer, indicate "1".

If the intervention is less than 10 weeks, indicate "2".

If not listed, indicate "3".

Section 3: Settings and Subjects

The study must identify middle/junior high schools, high schools, community/junior colleges, or universities as the setting for the intervention.

The study must identify the grade level of the subjects participating in the study.

3.1 Study identifies middle/junior high schools, high schools, community/junior colleges, or universities as the setting. _____

If yes, indicate "1", if no, indicate "2".

3.2 Study identifies the grade level of the subjects as appropriate for the analysis (secondary, CC/JC or first 2 years or university). _____

If yes, indicate "1", if no, indicate "2".

Section 4: Outcome(s)

The study must identify the outcome(s) of the study as student achievement in science. The study must identify the type of instrument used to measure the outcome (e.g., standardized/commercial assessment instrument or teacher constructed assessment instrument).

4.1 Study identifies student achievement in science as the outcome. _____

If yes, indicate "1", if no, indicate "2".

4.2 Study identifies the type of instrument used to measure the outcome._____

If the instrument is standardized/commercial, indicate "1". If the instrument is teacher constructed, indicate "2". If the instrument is not identified specifically, but aligns with the meta-analytic interest, indicate "3". If the instrument is not identified, indicate "4".

Section 5: Research Design

The study must identify randomized controlled trial or quasi-experimental as the research design.

5.1 Research design of the study _____

If the study identifies randomized controlled trial as the design, indicate "1". If the study identifies a quasi-experimental design, indicate "2". If the study identifies any other type of design, indicate "3".

Section 6: Date of Publication

The study must identify a date of publication between 1995-2007.

6.1 Date of publication of study _____

If the study identifies a date of publication between 1995-2007, indicate "1". If the study identifies any other date of publication, indicate "2".

Section 7: Effect Size Data

The study must report quantitative data which will allow effect size statistics to be calculated. This may include means, standard deviations, *F*-values, *t*-values, or *p*-values.

7.1 Data reported in study _____

If the study includes means, standard deviations, *F*-values, *t*-values, or *p*-values, indicate "1". If the study does not include these values, indicate "2".

Section 8: Final Code of Study

The study must be coded according to the following criteria in order to qualify for complete full-text screening.

Code 2.1 = "1" or "2"

Code 2.2 = "1"

Code 3.1 = "1"

Code 3.2 = "1"

Code 4.1 = "1"

Code 4.2 = "1", "2", or "3"

Code 5.1 = "1"

Code 6.1 = "1"

Code 7.1 = "1"

8.1 Study qualifies for complete full-text screening _____

If study meets all above criteria, indicate "Y" for yes, study qualifies. If study does not meet at least one of the above criteria, indicate "N" for no, study does not qualify.

8.2 Justification for exclusion of study

If 8.1 = "N", provide a brief explanation justifying its exclusion from complete full-text screening.

APPENDIX C:

Full-Text Study Coding Form

APPENDIX C

Full-Text Study Coding Form

Section 1: Source of the Report

1.1 Study Identification Number (ID; whole number) _____

1.2 Name of Coder _____

1.3 Type of Report _____

Key: 1 = Book

2 = Journal article or book chapter

3 = Dissertation (Master's or doctoral)

4 = Conference proceedings

5 = Unknown (source is unidentifiable)

1.4 Source of the Report _____

Key: 1 = Electronic search

2 = Personal archives

3 = Review of journals

4 = Review of reference lists (i.e., bibliography)

5 = Colleagues

1.5 Publication Year of Report (full four-digit year) _____

Year not listed = N/A

1.6 Publication Status _____

Key: 1 = Published, peer reviewed

2 = Unpublished

1.7 Coding of Manuscript and Inclusion of Study (return to this section after decision has been made on inclusion of study; if any of codes listed in 5.1 = "N", exclude study; if get "N" code before completing coding, indicate specific code that excluded). _____

Key: Y = If study is to be included, continue coding

N = Study is excluded (list reasons)

Section 2: Study Description

2.1 Country of Origin of the Study _____

Key: 1 = United States

2 = Other (specify country)

2.2 Research Design _____

Refers to the empirical design of the study and whether the interventions are active, inactive, or non-existent, as described in the study.

Key: 1 = Randomized controlled trial (participants are assigned to treatment and control groups randomly)

2 = Cluster-randomized with random assignment to treatment and control groups

3 = Quasi-experimental with subject matching

4 = Quasi-experimental without subject matching

5 = Single Group, Pretest/Posttest Design (Study employs a quantitative methodology using a single treatment group with single pretest and posttest)

6 = Within Subjects Design (time-series, counterbalance, crossover)

7 = Other (correlational, descriptive, no intervention)

2.3 Types of Cooperative Learning Interventions _____

Code the type of cooperative learning intervention. If author specifies type of CL, code as such. If not listed here, summarize the intervention based on the author's description. If more than one type of CL is listed, list all types in code.

Key: 1 = Learning together and alone

2 = Teams-Games-Tournaments

3 = Group Investigation

4 = Constructive Controversy

5 = Jigsaw (I or II)

6 = Student Teams Achievement Divisions

7 = Complex Instruction

8 = Team Accelerated Instruction

9 = Co-op co-op

10 = Cooperative Integrated Reading & Comprehension

11 = Type of cooperative learning not specified

12 = If CL named by author is not accurate

13 = Other (if cannot specifically identify CL type)

14 = Intervention not CL

15 = Multiple interventions

2.3.1 Length of intervention _____

Key: 1 = Intervention is 4 weeks or less in length

2 = Intervention is 4-10 weeks in length

3 = Intervention is longer than 10 weeks in length

4 = Length not specified

2.4 Outcomes

Code the outcome(s) in the study.

2.4.1 Outcome 1 _____

Key: 1 = Student achievement based on assessment instrument.

2 = Student achievement based on course grade or grade point average.

2.4.1.1 Source of assessment for Outcome 1 _____

Code the source of the assessment.

Key: 1 = Standardized/commercial instrument.

2 = Teacher constructed instrument.

3 = School archives (for course grades or grade point average)

2.4.1.2 Reliability/validity of assessment instrument _____

Code whether reliability/validity testing was performed on assessment instrument.

Key: 1 = Reliability statistics reported.

2 = Reliability statistics not reported.

3 = If standardized/commercial instrument and no reliability statistics reported.

4 = Reliability statistics not reported but instrument was reviewed by experts.

2.4.2 Outcome 2 _____

Key: 1 = Student achievement based on assessment instrument.

2 = Student achievement based on course grade or grade point average.

2.4.2.1 Source of assessment for Outcome 1 _____

Code the source of the assessment.

Key: 1 = Standardized/commercial instrument.

2 = Teacher constructed instrument.

3 = School archives (for course grades or grade point average)

2.4.2.2 Reliability/validity of assessment instrument _____

Code whether reliability/validity testing was performed on assessment instrument.

Key: 1 = Reliability statistics reported.

2 = Reliability statistics not reported.

3 = If standardized/commercial instrument and no reliability statistics reported.

4 = Reliability statistics not reported but instrument was reviewed by experts.

2.4.3 Outcome 3 _____

Key: 1 = Student achievement based on assessment instrument.

2 = Student achievement based on course grade or grade point average.

2.4.3.1 Source of assessment for Outcome 1 _____

Code the source of the assessment.

Key: 1 = Standardized/commercial instrument.

2 = Teacher constructed instrument.

3 = School archives (for course grades or grade point average)

2.4.3.2 Reliability/validity of assessment instrument _____

Code whether reliability/validity testing was performed on assessment instrument.

Key: 1 = Reliability statistics reported.

2 = Reliability statistics not reported.

3 = If standardized/commercial instrument and no reliability statistics reported.

4 = Reliability statistics not reported but instrument was reviewed by experts.

2.5 Educational Setting _____

What was the educational setting of the study as described by the authors?

Key: 1 = Middle/Junior High School (Any public or private institution including grades 6-8 or 9. Grade 6 must be associated with a middle school curriculum or part of a secondary school, not an elementary school)

2 = High School (Any public or private school including grades 9 or 10-12; includes freshmen/women, sophomores, juniors, seniors. If study reports grade 9, this defaults as high school.)

3 = Community/Junior College (2 year) (A public or private institution that awards associate degrees, certificates, and/or provides a transfer bridge to four-year institutions.)

4 = Four-year College (Any public or private college or university that awards baccalaureate degrees.)

5 = Boarding School (A school of any grade level where the students live on the premises)

6 = Grade 6 but part of elementary curriculum.

7 = Other (Educational setting cannot be placed into any of the above categories.)

8 = Indeterminate or non-educational (Cannot adequately describe the educational setting.)

Section 3: Study Population

3.1 Grade of Learners _____

Refers to the educational grade level of the participants as reported by the authors.

Key: Specify single grade or range of grades.

U = Unknown

3.2 Science Content Area _____

Key: 1 = General Science

2 = Biology

3 = Chemistry

4 = Earth Science (Includes Astronomy, Geology, Meterology)

5 = Physics

6 = Physical science

7 = Engineering

8 = Other (not science discipline)

3.3 Gender of Learners _____

Gender statistics reported by the authors. If code "1", go to 3.3.1

Key: 1 = Yes

2 = No

3.3.1 Ratio of genders

Key: 1 = Ratio of females to males > 2:1

2 = Ratio of males to females > 2:1

3 = Gender ratio is less than 2:1 or greater than or equal to 1:1

4 = Genders segregated

3.4 Race/Ethnicity of Learners _____

Race/ethnicity statistics reported by the authors. If code "1", go to 3.4.1.

Key: 1 = Race/ethnicity is reported

2 = Race/ethnicity is not reported

3.4.1 Percentages of Race/Ethnicity of Learners

If the answer to 3.4 was "1", specify the percentage in the table of race/ethnicity of learners in the study as reported by the authors.

_____ Latino/Latina/Hispanic

_____ African-American

_____ Native American Indian

_____ Asian-Pacific Islander

_____ Multi-racial

_____ Caucasian/White/Anglo

_____ Other

3.5 Other "At-Risk" Group Specified _____

Refers to any group of students deemed "at-risk" that are not listed in 3.3 and/or 3.4 above. Other "at-risk" statistics reported by the authors. If code "1", go to 3.5.1.

Key: 1 = Other "at-risk" groups reported.

2 = Other "at-risk" groups not reported.

3.5.1 Percentages of other "at-risk" groups

If the answer to 3.5 was "1", specify the percentage in the table of other "at-risk" groups of learners in the study as reported by the authors.

_____ Lower socioeconomic status/free or reduced lunch recipients

_____ English language learners

_____ Special education/disability

_____ Other (specify)

Section 4: Statistics

4.1 What statistics are reported in the study? _____

Key: 1 = Means with SD, N

2 = *t*-test values with *p* value only, N

3 = *t*-test values with *p* value with means and SD, N

4 = *F* values with *p* value only, N

5 = *F* values with *p* value with means and SD, N

6 = Correlation coefficients with *p* value

7 = Chi-square values with *p* value

8 = Other statistics reported but not listed above (specify)

9 = Statistics reported, but no N values for treatment/control groups

10 = No statistics reported

Section 5: Other _____

5.1 Does the report meet the inclusionary criteria?

Key: 1 = Yes

2 = No (Check "No" only if 2.2 is coded as "5-7", OR 2.3 is coded "14" OR "15", OR 2.5 is coded as "6-8", OR 3.1 is coded "U", OR 3.2 is coded as "8", OR 4.1 is coded as "9" OR "10".)

NOTE TO CODERS: If a code cannot be assigned, then leave blank.