2013

A Meta-Analysis of the Effectiveness of Stem-Programs in the United States

Dayoung An
Cleveland State University

Follow this and additional works at: https://engagedscholarship.csuohio.edu/etdarchive

Part of the Urban Studies and Planning Commons

How does access to this work benefit you? Let us know!

Recommended Citation
https://engagedscholarship.csuohio.edu/etdarchive/75

This Dissertation is brought to you for free and open access by EngagedScholarship@CSU. It has been accepted for inclusion in ETD Archive by an authorized administrator of EngagedScholarship@CSU. For more information, please contact library.es@csuohio.edu.
A META-ANALYSIS OF THE EFFECTIVENESS OF STEM-PROGRAMS IN THE UNITED STATES

DAYOUNG AN

Bachelor of Science in Public Administration
Sangmyung University
March 2001

Master of Science in Public Administration
Hankuk University of Foreign Studies
March 2003

Submitted in partial fulfillment of requirement for the degree
DOCTOR OF PHILOSOPHY IN URBAN STUDIES AND PUBLIC AFFAIRS
at
CLEVELAND STATE UNIVERSITY
JULY 10
This dissertation has been approved for

The Department of URBAN STUDIES and the College of Graduate Studies by

Dissertation Chairperson, Chieh-Chen Bowen, Ph.D.

Department & Date

William M Bowen, Ph.D.

Department & Date

Jeffry L. Brudney, Ph.D.

Department & Date

Michael Schwartz, Ph.D.

Department & Date
ACKNOWLEDGEMENTS

I would never have been able to finish my dissertation without the guidance of my committee members, help from friends, and support from my family.

I would like to express my deepest appreciation to my advisor, Dr. Chieh-Chen Bowen, for her excellent guidance, patience, and providing me with critical points for my writing of this study and patiently correcting my writing. Without her guidance and persistent help, this dissertation would not have been possible. I would like to thank Dr. Jeffry Brudney, who let me have many opportunities to work with him and present our work in South Korea. It was honor to work with him. His policy perspectives on my dissertation stimulated and extended my academic curiosity in policy areas and volunteerism. I would also like to thank Dr. Bill Bowen for remaining committed to the success of my dissertation and guiding my doctoral coursework for the past several years. His care for Korean students, including me, led me to successfully complete this study. Special thanks go to Dr. Michael Schwartz, who was willing to participate in my final defense committee to the very end. His last question in the defense presentation challenged me in a great way and helped me figure out an important issue.

I would like to thank Misun, Sunjoo, Changshik, Taekyoung and all of my other Korean friends, who as good friends were always willing to help and give their best suggestions. It would have been a lonely cubicle without them. I wish to thank Dr. Sugie Lee, for guiding me and sharing his doctoral life which encouraged me to patiently work hard. Thanks to Jim and Sue, Marian, Ruth and Sandy, and my other ABF classmates in
Parma Height Baptists Church, and pastor Yong Kim in South Korea, for encouraging and praying for my successful American life.

I would also like to thank my wonderful parents, lovely sister, two brothers, and two sisters in law in South Korea. All of them were always supporting me and encouraging me with their best wishes. Their endless financial and spiritual support led me complete my research and move on next stage. This support was a renewable energy source for overcoming any difficult times in my life. Also, thanks to Ashley, Michael, and my American Mom for accepting and loving me. Finally, I would like to thank my husband, Eddie Woodworth. He was always there to cheer me up, stood by me through the good times and bad, and continues to support me to achieve my dreams. I am your Trouble forever. Love you all!!
ABSTRACT

The federal government has been spending a large amount of funds on STEM programs. It is important to examine the effectiveness of such spending. Much research has been conducted in the past 30 years for this particular purpose; however, the results of such evaluations have not painted a clear picture of the effectiveness of STEM programs. The goal of this meta-analysis is to investigate whether STEM programs are successful in the outcomes they claim to achieve.

Such a meta-analysis must integrate all of the empirical studies which reported their effort in evaluating the effectiveness of STEM programs, based on measures of (1) engagement, (2) capability, and (3) continuity (Jolly, Campbell, and Perlman (2004). Previous studies on the STEM program effectiveness used either a within-subject design or a between-subject design. First, each of these research designs was investigated independently to examine whether the particular design has an effect on estimates of STEM program effectiveness. In addition, other moderators were investigated to determine factors that could influence estimates of STEM program effectiveness: pedagogical types of programs, program funders, program creators, grade level of sample groups, regional locations, instrument reliabilities, publication types, time, etc. The meta-analysis covers literature published between 1980 and 2010. The total number of studies included in this meta-analysis is 91.

This study finds that all three outcome variables have positive effect sizes at the moderate level with the weighted mean effect sizes of .346 for engagement, .456 for capability, and .369 for continuity measurements. Additionally, the between-subject design versus within-subject design has an effect on estimates of the engagement and
capability outcomes of STEM programs. Finally, some moderator variables were statistically significant on the outcome variables with the mean effect sizes: program strategy, program creator, regional location, and educational level of sample group. These findings suggest the best pedagogical types of STEM programs, program creators, and the most effective target groups in order to achieve maximum effects of STEM programs.
# Table of Contents

ABSTRACT.................................................................................................................................................. v

List of Tables .............................................................................................................................................. x

List of Figure........................................................................................................................................... xii

CHAPTER I ................................................................................................................................................ 1

INTRODUCTION ........................................................................................................................................ 1

1.1 Statement of Problem .......................................................................................................................... 1

1.1.1 STEM Education Programs in the United States ........................................................................ 1

1.1.2 Inconsistent Results in the Effectiveness of STEM Programs ...................................................... 3

1.2 Objective of the Study ......................................................................................................................... 6

1.2.1 Measures of Effectiveness ............................................................................................................. 7

1.2.2 Independent Variables .................................................................................................................... 8

1.3 Justification of the Meta-analysis ....................................................................................................... 9

1.4 Definition of Terms ............................................................................................................................. 10

1.5 Organization of the Study ................................................................................................................... 11

CHAPTER II .............................................................................................................................................. 13

BACKGROUND OF STEM PROGRAMS.......................................................................................... 13

2.1 Concepts of STEM programs ............................................................................................................. 13

2.2 Multiple Goals and Target Groups for Federal Programs ................................................................. 15

2.3 Federal and State Research and Development for STEM Programs ............................................. 17

2.4 Federal Investment for STEM Education Programs ........................................................................ 20

CHAPTER III .......................................................................................................................................... 23

LITERATURE REVIEW ..................................................................................................................... 23
5.1.2 Heterogeneity for Three Effectiveness Measurements .................................. 76
5.2 Moderator Variables for the Engagement Outcomes ........................................ 77
5.3 Moderator Variables for the Capability Outcome ............................................ 92
5.4 Moderator Variables for the Continuity Outcome .......................................... 109
CHAPTER VI ........................................................................................................ 123
DISCUSSION ......................................................................................................... 123
6.1 Implication for Research .................................................................................. 123
6.2 Implication for STEM Education Policy ......................................................... 127
   6.2.1 Policy Implication related to the program strategy ................................. 127
   6.2.2 Policy Implication related to the program creator ................................. 129
   6.2.3 Policy Implication related to the grade level ......................................... 130
6.3 Limitations of This Study ................................................................................ 132
6.4 Conclusion ........................................................................................................ 134
6.5 Recommendation for Future Research ......................................................... 137
Reference ............................................................................................................. 139
Appendix ............................................................................................................... 158
List of Tables

Table 2-1 Goals and Number of STEM Programs Targeted to Multiple Groups .......... 17
Table 2-2 Total R&D and Federal Investment for STEM R&D ............................. 19
Table 2-3 Federal STEM Education Programs Funding and Number of Programs .... 21
Table 3-1 Three Effectiveness Measures of STEM programs .............................. 25
Table 4-1 Search Results by Research Database including a Manual Search ............. 51
Table 4-2 The Study Characteristics by Categories ............................................ 56
Table 4-3 Kappa Coefficients of Dependent and Independent variables ................. 65
Table 5-1 The Effect Sizes and Homogeneity Test Results of Three Outcomes ......... 75
Table 5-2 Effect of Moderator Variables on Weighted Mean Effect Sizes in the
Engagement of Students and Teachers ............................................................ 78
Table 5-3 The Effect Sizes of STEM Programs on Engagement by Research Design .... 80
Table 5-4 The Effect Sizes of STEM Programs on Engagement by Program Strategy ... 82
Table 5-5 The Effect Sizes of STEM Programs on Engagement by Founding Source ... 83
Table 5-6 The Effect Sizes of STEM Programs on Engagement by Program Creator..... 85
Table 5-7 The Effect Sizes of STEM Programs on Engagement by Regional Location.. 87
Table 5-8 The Effect Sizes of STEM Programs on Engagement by Educational Level .. 89
Table 5-9 The Effect Sizes of STEM Programs on Engagement by Publication Year .... 90
Table 5-10 The Effect Sizes of STEM Programs on Engagement by Reliability .......... 92
Table 5-11 Effect of Moderator Variables on Weighted Mean Effect Sizes in the
Capability of Students and Teachers ................................................................. 94
Table 5-12 The Effect Sizes of STEM Programs on Capability by Research Design..... 96
Table 5-13 The Effect Sizes of STEM Programs on Capability by Program Strategy..... 98
Table 5-14 The Effect Sizes of STEM Programs on Capability by Founding Source ..... 99
Table 5-15 The Effect Sizes of STEM Programs on Capability by Program Creator .... 102
Table 5-16 The Effect Sizes of STEM Programs on the Capability by Education Level ................................................................. 104
Table 5-17 The Effect Sizes of STEM Programs on Capability by Regional Location . 105
Table 5-18 The Effect Sizes of STEM Programs on Capability by Publication Year .... 107
Table 5-19 The Effect Sizes of STEM Programs on Capability by Reliability .......... 108
Table 5-20 Effect of Moderator Variables on Weighted Mean Effect Sizes in Student Continuity ................................................................................................................................. 110
Table 5-21 The Effect Sizes of STEM Programs on Continuity by Research Design ... 111
Table 5-22 The Effect Sizes of STEM Programs on Continuity by Program Strategy .. 113
Table 5-23 The Effect Sizes of STEM Programs on Continuity by Founding Source ... 114
Table 5-24 The Effect Sizes of STEM Programs on Continuity by Program Creator ... 116
Table 5-25 The Effect Sizes of STEM Programs on Continuity by Education Level .... 117
Table 5-26 The Effect Sizes of STEM Programs on Continuity by Regional Location. 119
Table 5-27 The Effect Sizes of STEM Programs on Continuity by Reliability .......... 120
Table 5-28 The Effect Sizes of STEM Programs on Continuity by Published Year...... 122
List of Figure

Figure 3 - 1 Procedure of Meta-analysis on effectiveness of STEM programs............... 45
CHAPTER I

INTRODUCTION

1.1 Statement of Problem

1.1.1 STEM Education Programs in the United States

The Trends in International Mathematics and Science Study (TIMSS) data (National Center Education Statistics, 2007) shows that the levels of mathematics and science (M&S) achievements of U.S. students are below those of students in other countries. According to the TIMSS fourth-grade and eighth-grade assessments in M&S achievement in 2007, U.S. scores in mathematics are 11th and 9th place out of 35 and 46 countries, respectively. U.S. fourth- and eighth-graders’ performance in science achievement are 8th and 11th place out of 35 and 46 countries, respectively.

There are some concerns about students’ apparent lack of interest in science in K-12 (Simpson & Oliver, 1990; Greenfield, 1996; Jovanovich & King, 1998) and a
continued disinterest in STEM fields in college (Simpson & Oliver, 1990; Bazler, Spokane, Ballare, & Fugate, 1993). Improving recruitment and retention rates of students in STEM-associated programs is an important tool to increase the number of future potential STEM workers.

Along with the emerging importance of a STEM workforce, concerns about the quality of STEM education have spurred reform in educational policies and training programs. The lack of an effective “pipeline” or “pathway” from early science and math education to successful science- and math- related careers is recognized as a problem in the American education system (American Electronics Association, 2005; Teacher College, 2005). The National Research Council (2001) suggested that the role of education in increasing the supply of qualified labor is to encourage students to acquire both technical and foundation skills.

To help students achieve in science and mathematics, we need well-qualified teachers (National Research Council, 2001, p. 232). To improve students’ achievement in science, it is essential to have teachers who are knowledgeable in science content, confident of their ability to guide and conduct science investigations, and well-versed in pedagogical skills (Radford, 1998). The changing and expanding demands of teaching jobs have prompted increased attention to the importance of professional development in providing teachers with opportunities to increase their professional knowledge (Elmore, 2002; Little, 1993). The National Research Council pointed out that most teachers lack the professional development and support needed to incorporate technology into daily instruction and, as a result, significant numbers of these teachers ignore the pedagogical uses of technology.
The STEM fields must not only attract more students, but also ensure that students obtain better knowledge about STEM through the educational system as a whole. STEM background and training must start early in education systems in preparation for college-level study.

### 1.1.2 Inconsistent Results in the Effectiveness of STEM Programs

There are numerous studies on the effectiveness of individual STEM education programs. However, individual studies have reported various estimates of STEM program effectiveness. Some have found that STEM programs create positive influence on students’ and teachers’ interest, commitment, and test scores in various STEM disciplines. Others have found that STEM programs create little or even negative effects in students’ or teachers’ interest, commitment or test scores.

#### 1.1.2.1 Positive Results for STEM Programs

Conventional wisdom might have believed that after a large amount of funds are spent on STEM programs, there are supposed to be positive outcomes from these programs. Indeed, many studies on STEM programs’ effectiveness have shown that STEM programs have had some success for students and teachers in enhancing engagement in STEM fields (Shymansky, Yore, & Anderson, 2004; MaDevitt & Troyer, 1995; Strawitz & Malone, 1986; Sorge, Newsom, & Hagerty, 2000; Kyle, Bonnstetter, & Gadsden, 1988; Ucovic, Morries, Dickman, Postlethwait, & Wetherwax, 2002). Some studies have reported that STEM programs lead to a gain of knowledge and skill (Ruberg, Chen, & Martin, 2007; Lott, 2003; Frantz, DeHaan, Demetrikopoulos & Carruth, 2006;
Vargas-Gomez & Yager, 1987; Smith & Erd, 1986; Hamrick & Harty, 1987; Radford, 1997; McGinis, 2003). Some studies (Randy, Steve, & Tad, 1998; Carney, Chawla, Wiley, & Young, 2006) have concluded that participation in STEM programs leads to students’ continuity in STEM-related fields.

1.1.2.2 Negative Results for STEM Programs

Besides the positive results from STEM programs mentioned above, many studies have shown that STEM programs have failed to improve engagement, capability, or continuity for students and teachers (Ruberg, Chen, & Martin, 2007; Sorge, Newsom, & Hagerty, 2000; Stake & Mares, 2005; McGinis, 2003; Lott, 2003; Ault, 2006; Barlow & Villarejo, 2004; Shymansky, Yore & Anderson, 2004; Shapley & Luttrell, 1992). For example, Shapley and Luttrell (1992) concentrated on the effectiveness of a Mentor In-Service Training program that aimed to increase the enjoyment of teaching science through targeting teachers’ beliefs about teaching science and their understanding of the nature of science. Conducted in the elementary schools of a large metropolitan district during 1988-1989, this study compared pretest with posttest scores, showing that participating teachers were less likely to enjoy teaching science.

1.1.2.3 Inconclusive Results for STEM Programs

In addition, a great number of studies show inconclusive results of STEM programs (Sorge, Newsom, & Hagerty, 2000; Nagda, Gregerman, Jonides, Hippel, & Lerner, 1998; Moseley, Reinke, & Bookout, 2003; Martin, 2003; Minger & Simpson, 2006; Robardey, Allard, & Brown, 1994; Upadhyay & DeFranco; 2008). For example, Robardey, Allard, and Brown (1994) evaluated a Full Option Science System (FOSS) to
measure students’ attitude toward science and teachers’ attitude toward teaching science. This evaluation was conducted in Little River and Miller counties in Arkansas, as well as Bowie and Cass Counties in Texas with third- through sixth-grade students and teachers. The inference from this research is that the participants’ attitude toward science as measured by pretest/posttest design was unaffected by the program (did not show a statistically significant difference between pre-test and post-test).

A traditional narrative review of all these evaluations could not produce an accurate and systematic reflection of the effectiveness of STEM programs. In any such narrative, the synthesis of the results of individual research studies on STEM programs’ effectiveness would be highly susceptible to bias by the reviewer’s subjectivity in selecting previous studies and aggregating across their results. Additionally, it becomes more and more difficult to comprehensively synthesize the existing studies when the number of studies is large and the results are diverse.
1.2 Objective of the Study

The total U.S Research and Development (R&D) expenditure by the federal government, private industry, and other funding sources was estimated at $323 billion in 2007. Of this, federal agencies have funded approximately $28.4 billion to academic institutions for STEM R&D programs in FY2007 (National Science Foundation, 2006). The examination of whether STEM programs funded by the government are effective allows public officials and practitioners to think strategically about implementing future STEM programs.

A meta-analysis is the most useful method to integrate inconsistent results from large numbers of previous studies (Bowen, 2008). Individual evaluations of STEM programs have been performed with empirical studies under a variety of definitions, units of analysis, and contexts of study; thus, the results of the impact of STEM programs are inconsistent. The purpose of this research is to integrate the inconsistent results concerning the effectiveness of STEM education programs by conducting a meta-analysis. A meta-analysis allows the inconsistent results from the individual studies to be statistically and comprehensibly aggregated (Cooper, 1989; Cooper & Hedge, 1994). This study provides an aggregate level assessment of STEM programs over the last three decades for the use of policy decision makers and practitioners.
1.2.1 Measures of Effectiveness

Jolly, Campbell, and Perlman (2004) have identified three measures for students’ successful pursuit and entry into STEM careers: engagement, capability, and continuity. These reflect the relevant conceptual definitions that individual researchers have used to define and evaluate the outcomes from STEM programs. This meta-analysis will be conducted based on these three measures of effectiveness.

‘Engagement’ is defined as students’ and teachers’ interest and initial involvement toward STEM-related disciplines. Applications of the concept of engagement have varied somewhat between individual studies: it has referred for instance to interest in STEM fields, confidence in the STEM fields, enjoyment of STEM areas, post-STEM program belief change, and students’ motivation. The common reference through all applications is students’ and teachers’ affective and behavioral experience. Scholars and others who have interest in evaluating STEM programs remain interested in students’ affective and behavioral experience resulting from their participation in these programs. Thus, even though the concept has been applied in slightly different ways, it remains relevant for purposes at hand.

‘Capability’ refers to skills or knowledge improvement from STEM programs. The skills and knowledge to which the concept of ‘capability’ has been applied have ranged from hard skills, such as GPA improvement, to soft skills, such as the ability to solve problems or to reason. Research studies have identified the target groups including both students and teachers, and these groups ranged from K-12 through doctoral levels. The range of skill sets across all levels of participants is subsumed under the category of “capability”. The defining feature of improvements in capability is not the level of skill
or knowledge with which the individual starts prior to the program, but rather the degree
to which this starting level is augmented or improved as a result of the program.

‘Continuity’ is defined as the opportunity for students and teachers to move ahead
to the next level of the educational and work systems. The concept of continuity is
important to consider because it measures the direct linkage between a STEM programs’
processes and activities today, and the effectiveness of these processes and activities in
terms of tomorrow’s workforce. The concept describes how STEM programs retain
students within STEM fields. Even though the studies on the continuity derived from
STEM programs are less numerous than those of the two above variables, many studies
have been conducted over the last three decades: Barlow and Villarejo, 2004; Kyle,
Bonnstetter, and Gadsden, 1988; Clewell, Cohen, and Tsui 2005; Carney, Chawla, Wiley,
and Young, 2006; Maton, Hrabowski III, and Schmitt, 2000; Nagda, Gregerman, Jonides,
von Hippel and Lerner, 1998; Sorge, Newsom, and Hagerty, 2000. These studies do not
present a clear picture of whether STEM programs produce a competent workforce in the
STEM areas.

1.2.2 Independent Variables

Previous studies conducted for the purpose of evaluating the effectiveness of
STEM programs were classified as either within-subject design or between-subject
design. Within-subject design is usually illustrated by comparing outcomes from the
same group of participants measured twice. The first measure is taken before they
participated in STEM programs and the second measure is taken after the participation.
Between-subject design is usually illustrated by comparing outcomes from two groups of
participants. One group of participants who did not go through any STEM program is labeled as the control group and the other group who went through STEM programs is labeled as the treatment group. The effectiveness of the STEM program is estimated by measuring the differences between the treatment group and the control group.

Each of the two designs has its own strengths and weaknesses. A within-subject design inherently eliminates individual differences, which is a main concern of a between-subject design. On the other hand, the between-subject designs use random assignment to control for most of the threats to internal validity such as history, maturation, and testing effect. As a result of random assignment, any threats would be expected to be equally manifested in both the experimental and control groups. However, the majority of studies on STEM program evaluation have used the nonrandom assignment.

### 1.3 Justification of the Meta-analysis

There are numerous studies documenting the impact of STEM programs on engagement, capability, and continuity of students and teachers in STEM fields; however, the results of these studies are inconsistent in terms of reporting the impact of STEM programs as a function of the great investment of government resources. This study explores the extent to which the overall effectiveness of STEM programs is moderated by various study characteristics. The independent variables comprise types of research designs for evaluating STEM programs and the dependent variables include engagement, capability, and continuity. Bowen (2008) stated that moderators may be expected to
change the direction or magnitude of the relationship between the independent variable and the dependent variable. In this study, individual study characteristics are anticipated to have an impact on the outcomes of STEM programs.

1.4 Definition of Terms

*Program*: The Academic Competitiveness Council (2007) defines a program as “the largest of identifiable set of projects or activities that have generally similar objectives, strategies, and target audiences” (U.S. Department of Education, 2007, p. 11).

*Effectiveness of STEM program*: the capability to produce, generate, propagate, or transmit effects from projects or activities in the subjects of science, technology, engineering, and/or mathematics; effects specifically observed in enhanced students’ and teachers’ engagement, capability, and/or continuity.

*Engagement*: this dependent variable in this study includes attitudes, interest, motivation, and initial involvement toward STEM fields for teachers and students.

*Capability*: this dependent variable includes the students’ and teachers’ knowledge and skills necessary to do work and succeed in STEM fields.

*Continuity*: this dependent variable includes the opportunity to move ahead to the next level of education or further participation in work activities.

*Control group*: a group of students or teachers who were not selected to participate in STEM programs.

*Experimental group*: a group of students or teachers who were randomly or non-randomly selected to participate in STEM programs.
**Pre-test:** a test of the effectiveness of a STEM program prior to receiving this STEM program.

**Post-test:** a test of the effectiveness of a STEM program after such as STEM program is administered.

### 1.5 Organization of the Study

Following this introductory chapter, Chapter 2 describes the background of STEM programs. Chapter 2 includes the concept of STEM programs as commonly defined by many government organizations and scholars, a variety of objectives and multiple target groups of federal STEM programs, different types of STEM programs, and federal and state research and development (R&D) for STEM programs.

Chapter 3 provides a review of the literature. The review presents an overview of the inconsistent results of the existing studies on the effectiveness of STEM programs, the definitions of three variables (engagement, capability, and continuity), and methodologies used in individual studies. Furthermore, after comparing the methods of narrative review and meta-analysis, Chapter 3 describes the overall procedure of meta-analysis for integrating the existing studies to determine the effectiveness of STEM programs.

Chapter 4 describes the methodology of a meta-analysis as a way to comprehensively and statistically integrate the inconsistent results. Stage 1 involves addressing the research question along with the hypothesis to be examined. In Stage 2, databases are searched for relevant studies. Stage 3 involves setting out the search
criteria for selection. Stage 4 involves the results of the search. Stage 5 includes the coding process of variables analyzed in the meta-analysis. In Stage 6, the statistical calculations for effect sizes are addressed.

Chapter 5 presents the analysis of results utilized for the study. These results include the effect sizes of the three dependent variables by study designs and other moderator variables: pedagogical types of STEM programs, target groups, field types, program funder, programs creator, and participants’ ability level, educational level of sample group, regional location, school type, treatment duration, subject assignments, testing instruments, publication type, and time.

Chapter 6 provides some practical guidelines for researchers as well as important policy implications based on the results of STEM program effectiveness (as evaluated under the engagement, capability, and continuity outcomes). Furthermore, after recommending some policy implications for developing successful STEM programs, Chapter 6 includes conclusions, limitation of this study, and future research.
CHAPTER II

BACKGROUND OF STEM PROGRAMS

2.1 Concepts of STEM programs

The term “STEM” is defined in various ways by scholars and organizations. Some studies (Lantz, 2009; Morrison, 2006; National High School Alliance, 2008) are concerned with the integrative characteristics of STEM education. On the other hand, the STEM Education Caucus, the STEM Education Coalition, the Business-Higher Education Forum and the National High School Alliance for Education concentrate on the products of STEM educational programs. Recently, focusing on the outcomes of STEM programs has become more important due to a demand for an innovative technological workforce in the 21st century.

Science, technology, engineering, and mathematics education often has been called a meta-discipline; Hays, Blaine, and Lantz (2009) defined a meta-discipline as a “creation of a discipline based on the integration of other disciplines knowledge into a
new ‘whole’” (p.1). This interdisciplinary bridging among discrete disciplines is now treated as an entity, known as STEM (Morrison, 2006). The National High School Alliance (NHSA), a partnership of nearly fifty organizations representing a diverse cross-section of perspectives and approaches, defines STEM as an “integrative approach to teaching and learning that draws on the foundation of each individual field to form a cohesive course of instruction” (NHSA, 2008, p. 1).

The Business-Higher Education Forum (BHEF), an independent non-profit organization of leaders from American businesses, colleges and universities, and foundations, has a more specific definition for STEM: “an initiative for securing American’s leadership in science, technology, engineering, and mathematics fields and identifying promising strategies for strengthening the educational pipeline that leads to STEM careers” (BHEF, 2010 , p. 8). The STEM Education Coalition outlines the value of STEM education within their mission objectives, proposing that STEM education has a vital role in “enabling the U.S. to remain the economic and technological leader of the global marketplace of the 21st century” (STEM Education Coalition, 2008).

The STEM Education Caucus focuses national attention on STEM education by asserting that:

Science, Technology, Engineering and Mathematics (STEM) education is responsible for providing our country with three kinds of intellectual capital: scientists and engineers who will continue the research and development that is central to the economic growth of our country; technologically proficient workers who are capable of dealing with the demands of a science based, high technology workforce; and
scientifically literate voters and citizens who make intelligent
decisions about public policy and who understand the world around
them (STEM Education Caucus, 2008).

The differing emphases of the various STEM definitions allow for different
interpretations of the success of STEM education. At the same time, the large number of
disparate definitions as well as applications may lead to evaluations with non-comparable
results. Thus, this research is predicated upon the assumption that STEM programs
should be assessed in relation to their influence toward their proposed goal, and revised
as needed to assure quality and productivity.

2.2 Multiple Goals and Target Groups for Federal Programs

Increasing the number of students in STEM fields and improving the quality of
these programs are key goals for the 207 federally sponsored STEM programs that, as of
2005, were being conducted by 13 agencies (Government Accountability Office, 2005).
Another goal of STEM programs is to improve teacher quality from kindergarten through
the 12th grade levels. Specific goals outlined in the GAO survey (2005) included the
following: 1) attracting and preparing students at any education level to pursue
coursework in STEM areas; 2) attracting students to pursue collegiate degrees (from
Associate’s degrees to doctoral and postdoctoral work); 3) providing growth and research
opportunities for college and graduate students in STEM fields; 4) attracting graduates to
pursue careers in STEM fields; 5) improving teacher education in STEM areas; and 6) improving or expanding the capacity of institutions to promote STEM fields.

According to the Academic Competitiveness Council (ACC) (2007), STEM programs support activities in a wide variety of areas, including the following: STEM curriculum development; teacher professional development, recruitment and retention; institutional support (including programs to strengthen the educational capability of minority-serving institutions); student financial assistance; outreach and recognition to motivate interest in or continued work in STEM fields; and research aimed at improving STEM education.

Most of the programs are not targeted at specific groups but aim to serve a wide range of students, educators, and institutions. Of the 207 federal programs extant in 2005, 54 were targeted to one group of students or teachers and 151 had multiple target groups of students and faculty. The target groups for the remaining two programs were not identified (GAO, 2005). Few programs were targeted to elementary and secondary teachers as well as to kindergarten through 12th grade students. Table 2-1 summarizes the goals and number of STEM programs targeted to multiple goals and groups.
Table 2 - Goals and Number of STEM Programs Targeted to Multiple Groups

<table>
<thead>
<tr>
<th>STEM Program</th>
<th>Contents</th>
</tr>
</thead>
</table>
| Goal         | - Attract and prepare students at all educational levels to pursue coursework in STEM areas (114)  
               - Attract students to pursue STEM postsecondary degrees (two-years through a Ph. D. program) and postdoctoral appointments (137)  
               - Provide growth and research opportunities for college and graduate students in STEM fields (103)  
               - Attract graduates to pursue careers in STEM fields (131)  
               - Improve teacher education in STEM areas (73)  
               - Improve or expand the capacity of institutions to promote STEM fields (90) |
| Target Group | - Elementary school students (28)  
               - High school students (53)  
               - Four-year college students (96)  
               - Postdoctoral scholars (70)  
               - Secondary school teachers (50)  
               - Institutions (82)  
               - Middle school students (34)  
               - Two-year college students (58)  
               - Graduate students (100)  
               - Elementary school teachers (39)  
               - College faculty or instructional staff (79) |

Notice: The number of programs in parentheses is total number of programs targeted to each group.
Source: U.S. GAO (2005, p. 15)

Following Jolly, Campbell, & Perlman (2004), this study condenses the wide variety of goals for STEM programs into three goals for both students and teachers: engagement, capability, and continuity. The target groups are based on the combining goals described by the GAO and ACC and condensed from 11 classifications into three groups: students K-12, undergraduate students through postdoctoral scholars, and teachers K-12.

2.3 Federal and State Research and Development for STEM Programs

Historically, governments have made attempts to nurture intellectual human capital. Through legislation, the scope of STEM fields has expanded from science and
mathematics to information technology. In addition, the enactment of relevant laws resulted in more comprehensive education to support human capital through STEM programs. Moreover, the level of government investment in STEM has increased.


In addition, the National Science Foundation Act of 1950 required that the NSF initiate and maintain a program for the determination of the total amount of money needed for scientific and engineering research. Through changes in federal legislation, the explosion of computer/information technology, and the changes in workforce needs in the 20th century, the federal government has invested more in R&D and STEM education fields.
Table 2 - 2 Total R&D and Federal Investment for STEM R&D

<table>
<thead>
<tr>
<th>Year</th>
<th>Total R&amp;D</th>
<th>Total R&amp;D/GDP*</th>
<th>Federal R&amp;D</th>
<th>Federal R&amp;D/GDP*</th>
<th>STEM R&amp;D*</th>
<th>STEM/Federal R&amp;D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>63,224</td>
<td>2.27%</td>
<td>29,986</td>
<td>1.07%</td>
<td>4,791</td>
<td>15.97%</td>
</tr>
<tr>
<td>1990</td>
<td>151,990</td>
<td>2.62%</td>
<td>61,607</td>
<td>1.06%</td>
<td>10,471</td>
<td>16.99%</td>
</tr>
<tr>
<td>2000</td>
<td>267,562</td>
<td>2.73%</td>
<td>66,406</td>
<td>0.68%</td>
<td>19,877</td>
<td>29.93%</td>
</tr>
<tr>
<td>2005</td>
<td>323,005</td>
<td>2.60%</td>
<td>93,734</td>
<td>0.75%</td>
<td>28,381</td>
<td>30.27%</td>
</tr>
</tbody>
</table>

Source: National Science Foundation, Division of Science Resources Statistics. The complete data are available at http://www.nsf.gov/statistics in *National Patterns of R&D Resources*.

Notes: * Gross Domestic Product. These data are not adjusted for inflation. These data are based on performer surveys of expenditures for calendar years, and thus differ from data presented elsewhere from American Association for the Advancement of Science by fiscal year. These data also exclude R&D facilities.

The federal government provides the cornerstone of STEM education programs by mandating increased funding through legislation. Table 2-2 shows that the proportion of federal R&D spending targeted specifically toward STEM fields doubled from 1980 to 2005. According to the statistics from the National Science Foundation Survey of Federal Science Foundation and Engineering Support to Universities, Colleges, and Nonprofit Institutions (1996), federal agencies funded approximately $28.4 billion to academic institutions for STEM research and development programs in FY2007. The total U.S R&D, which has been funded by federal, industrial, and other sources, was estimated at $323 billion in 2007. The share of the federal R&D relative to GDP has increased slightly from 2.27% in 1980 to 2.60% in 2005.

Nationally, state support represented the largest source of revenue for elementary and secondary education. For example, 49% of funding for STEM programs came from states in 2002-03; local sources made up 43%; and the remaining 8% came from the federal government. In the school year 2002-03, the expenditure on STEM programs
from all sources totaled approximately $388 billion. The 2003 national average for spending on elementary and secondary education was 3.55% of GDP, a slight increase from 3.37% in 1994. It ranged from 2.23% to 5.09% of gross state product (GSP) among individual states. Spending for elementary and secondary expenditures as a share of GSP decreased in 17 states during the 1994-2003 periods (NSF, 2006).

The expansive amount of money from federal, state, and local governments allocated to STEM programs has been invested in the hopes that STEM education programs would enhance the essential competencies of the STEM workforce. Innovative capacity is highly desired: a recent National Governors Association Center (NGAC, 2010) statement announced that state funding of STEM programs was specifically tied to that goal. Consequently, there is a need to understand which STEM programs effectively provide desirable results for students and teachers. An overall assessment of STEM programs’ success over the past three decades can aid attempts to effectively allocate money and time to improve STEM programs. A meta-analytic approach to such an assessment may be used to consolidate and systematically analyze the various research studies performed to date, to do so on the basis of deep, long, logical chains of statistical reasoning. In addition, a synthesis of the studies on the effectiveness of STEM programs can contribute to better outcomes throughout STEM programs.

2.4 Federal Investment for STEM Education Programs

STEM education programs play a key role in educating the next generation of scientists and engineers. Federal agencies promote STEM fields with large investments
because these programs clearly contribute to new knowledge that ultimately drives the innovation process in their agencies. Table 2-3 displays the amount of federal agency funding and the number of the programs for STEM education nationally by agency.

Table 2 - 3 Federal STEM Education Programs Funding and Number of Programs by Agency (FY 2005-2007)

<table>
<thead>
<tr>
<th>Agency Name</th>
<th>FY2005 Funding</th>
<th>FY2006 Funding</th>
<th>FY2007 Funding</th>
<th>N</th>
<th>#</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Department of Agriculture</td>
<td>38,429,000</td>
<td>39,595,350</td>
<td>40,362,000</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Department of Commerce</td>
<td>36,028,049</td>
<td>38,717,250</td>
<td>19,974,250</td>
<td>9</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>Department of Defense</td>
<td>179,046,312</td>
<td>178,116,672</td>
<td>170,153,068</td>
<td>8</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Department of Education</td>
<td>461,157,189</td>
<td>705,523,110</td>
<td>1,065,028,820</td>
<td>9</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>Department of Energy</td>
<td>12,004,000</td>
<td>12,097,000</td>
<td>14,290,000</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Department of Health and Human Services</td>
<td>850,112,378</td>
<td>855,496,464</td>
<td>851,314,808</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Department of Homeland Security</td>
<td>10,600,000</td>
<td>13,300,000</td>
<td>12,500,000</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Department of the Interior</td>
<td>23,318,491</td>
<td>22,763,222</td>
<td>16,923,202</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Department of Transportation</td>
<td>135,573,000</td>
<td>152,216,500</td>
<td>151,990,500</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Environmental Protection Agency</td>
<td>11,100,000</td>
<td>11,055,000</td>
<td>7,400,000</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>National Aeronautics and Space Administration</td>
<td>178,800,000</td>
<td>162,400,000</td>
<td>153,300,000</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>National Science Foundation</td>
<td>878,250,000</td>
<td>923,760,000</td>
<td>970,650,000</td>
<td>31</td>
<td>35</td>
<td>34</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2,814,418,419</td>
<td>3,115,040,568</td>
<td>3,473,886,648</td>
<td>93</td>
<td>105</td>
<td>97</td>
</tr>
</tbody>
</table>

Source: U.S. Department of Education (2007, p.50)

According to the Academic Competitiveness Council (ACC) report (2007), 12 federal agencies and departments had 97 STEM education programs in fiscal year 2007. These 12 federal agencies reported spending about $3,473 billion for their STEM
education programs in that year. Most of the federal government’s R&D for STEM education programs has been allocated based on the programs’ mission-oriented budget. Accordingly, the R&D is intended to serve the goals and objectives of each agency with a broad variety of programs.

The vast majority of funding for STEM programs comes from three departments (National Science Foundation (NSF), U. S. Department of Education (ED), and U. S. Department of Health and Human Services (HHSD)) with much smaller shares contributed by the nine other federal agencies. These three agencies sponsored nearly half of the programs and spent about 83 percent of the funds in the fiscal year 2007.
CHAPTER III

LITERATURE REVIEW

3.1 Introduction

There are a large number of studies evaluating the effectiveness of the various STEM programs. Individual studies of the impact of STEM programs have been conducted in a variety of contexts, different times, and different sample sizes. Individual studies conducted in different conditions show inconsistent results. A closer look at these studies reveals that that a more coherent and consistent approach is needed for measuring the effectiveness of STEM programs.

This chapter is divided into four sections. Section 3.2 focuses on providing the various definitions of the three outcomes of STEM programs as described by Jolly, Campbell, and Perlman (2004) (engagement, capability, and continuity) examined in individual studies. The main objective for this section is to describe and enumerate the range of possible definitions related to outcomes of STEM programs. Section 3.3 focuses
on different methods used in individual studies on STEM program effectiveness. The main goal for this section is to overview which methods previous studies have used when evaluating STEM programs. Section 3.4 reviews the inconsistent results regarding the effectiveness of STEM programs. Section 3.5 focuses on the limitations of a narrative literature review. The final section of this chapter provides an overview of meta-analysis of STEM programs.

3.2 Three Measures for the Effectiveness of STEM Programs

Jolly, Campbell, and Perlman (2004) identified three measures for the successful pursuit and entry into a STEM career: engagement, capacity, and continuity. These three concepts distill the variety of terms used by individual researchers. In other words, each term encompasses the various concepts that are operationally defined in individual research. These three most comprehensively measure the range of operational concepts of STEM program effectiveness. This section applies the definition of engagement, capability, and continuity and classifies each individual study into a cohesive list (Table 3-1). Classification of research outcomes into three definitions allows for meta-analysis to integrate the range of individual studies by the various definitions of the successes of STEM programs, to meaningfully and coherently compare the studies and reduce the information contained in their conclusions into summary form.
Table 3 - 1 Three Effectiveness Measures of STEM programs for Students and Teachers

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Students</th>
<th>Teachers</th>
</tr>
</thead>
</table>
| Engagement | - Attitude toward STEM fields  
- Interest in STEM fields  
- Comfortable with STEM fields  
- Enjoyment of in STEM areas  
- Post-STEM program belief change  
- Students’ motivation, and confidence | - Teachers’ belief in the STEM fields  
- Attitude toward teaching in STEM fields  
- Attitude toward inquiry and teaching STEM content  
- Comfortable with STEM teaching |
| Capability (Knowledge) | - Achievement in STEM  
- Gaining skills to communicate  
- Process skills  
- Reasoning  
- Problem solving  
- Number relations  
- Improving the conceptual understanding of STEM  
- Algebraic proficiency | - Teachers’ knowledge of the subject matter and pedagogy  
- Conceptual knowledge about STEM  
- Teaching ability of STEM topics  
- Implementation  
- Propositional knowledge  
- Procedural knowledge  
- Communication interaction |
| Continuity | - Completing the classes  
- Continuing to study in STEM areas  
- Getting jobs in STEM fields  
- Persistence in science classes  
- Enrolling in STEM-related areas  
- Retention in STEM  
- Pursue further STEM fields courses | - |

3.2.1 Engagement

The studies that evaluate engagement vary in the way that they define engagement toward STEM programs. Fredericks (2004), Papanastasiou (2000), and Kim (2006) defined engagement as positive or negative attitudes, including interest, motivation, and initial involvement toward STEM fields for teachers and students. Many past studies have used participant ‘attitude’ as the performance criterion against which to evaluate the outcomes of STEM programs. Following Jolly, Campbell, and Perlman (2004), here these studies are subsumed under the concept of engagement (Lott, 2003; Stake & Mares., 2005; Shymansky, Yore & Anderson, 2004; Radford, 1998; Kyle, Bonnstetter, & Gadsden, 1988; Ucovic, Morries, Dickman, Postlethwait, & Wetherwax, 2002; Sorge,
Allport (1935) defined attitude as “a mental and neural state of readiness, organized through experience, exerting a directive or dynamic influence upon the individual’s response to all objects and situations with which it related” (p. 810). Fishbein and Ajzen (1975) suggested a similar interpretation of the notion by stating that attitude was “a learned predisposition to respond in a consistently favorable or unfavorable manner with respect to a given object” (p. 6). Thurstone and Chave (1929) explained the concept of attitude to represent the “sum-total of a man’s inclinations and feelings, prejudice or bias, preconceived notions, ideas, fears, threats, and convictions about any specific topic” (p. 74).

Other studies (Stake & Mares, 2005, 2001; Frantz, DeHaan, Demetrikopoulos & Carruth, 2006) used confidence and motivation toward science to measure attitude. The eight-item Science Self-Concept Scale (Campbell, 1991) measured students’ confidence in their science ability, including five positive items (e.g., “I have a lot of confidence in my abilities in science”), and three negative items (e.g., “Science is hard for me”). Finally, Shymansky, Yore, & Anderson (2004) assessed students’ affective stance toward science. It is assumed herein that even though various authors used different definitions of engagement in their studies they are still similar enough to represent them all meaningfully using the single concept of engagement.
3.2.2 Capability

Capability is defined as the students’ and teachers’ knowledge and skills necessary to conduct work and succeed in STEM fields. Some studies have examined the effectiveness of STEM programs by using the term “achievement” in regard to students or teachers (Ault, 2006; Rubba, McGuyer, & Wahlund, 1991; S. Lewis & Lewis, 2005; Adamson, Banks, & et al., 2003; Waugh, 1985; Garnes, Lindbeck, & Griffin, 1987, Choi & Gennaro; 1987, Lawrenz, Huffman, & Gravely, 2007; Shymansky, Yore, & Anderson, 2004; Adamson, Banks, Burtch, Cox III, Judson, Turley, Benford, & Lawson, 2003; Maton, Hrabowski III, Schmitt, 2000). Ault (2006) examined sixth grade students’ mathematical achievement in terms of algebraic knowledge, conceptual understanding, problem-solving, and number relations. Achievement can be a pervasive indicator of capability in STEM related disciplines.

Some studies have investigated the impact of STEM programs on improving knowledge and skills of students or teachers (Leonard, 1992; Hall & McCurdy, 1990; Grenawalt, et al., 1998; Ruberg, Chen, & Martin, 2007; Falconer, Joshua, Wychoff, & Sawads, 2001; Radford, 1997; Sawada, Piburn, & Judson, 2002; Brunkhorst, 1992; Carney, Chawla, Wiley, & Young, 2006; Radford, 1998). Researchers expected that positive performance-based outcomes would be expected for all groups of students and teachers after participating in the STEM programs. Lawrenz, Huffman, and Gravely (2007) measured how often teachers used various standards-based instructional methods to improve their capacity in teaching. Jeanpierre, Oberhauser, and Freeman (2005) investigated knowledge level and higher order thinking skills as the outcomes of the Professional Development Opportunity. The program aimed to enhance the ability of
teachers to successfully translate the full range of their inquiry skills into their science teaching practice. The definition of “capability” used by Radford (1998) also contained improvement in the science content knowledge and process skills. Thus, there are two broad areas that comprise the concept of capability as used herein: achievement and improvement of knowledge/skills. In this study, the term “capability” will refer to achievement and improvement of knowledge/skills of students and teachers that can be revealed quantitatively by scores, survey results, and interview data.

3.2.3 Continuity

Continuity is defined as the opportunity to move forward to the next level of education or to further participation in work activities. Continuity of students and teachers refers to the likelihood that students and teachers will gain employment in a field directly related to STEM. In addition, continuity refers to students who continue to study and seek higher education in STEM fields. Researchers can identify the notion of continuity in the formal systems of education and in the informal guidance that helps students and teachers navigate the educational system.

The formal system is concerned with whether students are studying and seeking advanced education in the STEM fields. Barlow and Villarejo (2004) found that a group with the Biology Undergraduate Scholars Program (BUSP) experience was more likely to successfully complete the three-quarter sequence in general chemistry and calculus than a comparison group. Maton, Hrabowskil III, and Schmitt (2000) explored the graduation rates in science and engineering and admittance to graduate school between the Meyerhoff Scholars Programs (MSP) students and comparison students at the University
of Maryland, Baltimore County in 1998. The MSP was designed to increase the number of under-represented minorities who pursue graduate and professional degrees in science and engineering. The results demonstrated that Meyerhoff students had greater success at retaining and graduating in the STEM area than non-Meyerhoff students who declined the program offer. These two researchers used the terms “graduate”, “complete” and “retain” to indicate the outcomes refereed to herein as continuity in STEM programs.

While the first part of the definition of continuity refers to STEM fields, the other part of the definition refers to employment in a job related to the STEM fields. There are few studies on the effectiveness of STEM programs in terms of continuity regarding the workforce in the STEM area. Carney, Chawla, Wiley, & Young (2006) investigated the opportunities for the Integrative Graduate Education and Research Traineeship (IGERT) students to engage in internships or work off campus. The purpose of the IGERT programs was to facilitate racial diversity in student participation, thus contributing to the development of a diverse, globally-engaged, science and engineering workforce. The result of the survey with 306 graduate students participating in the IGERT programs showed that the students had more opportunities to conduct internships or work off campus than 566 graduate students without the IGERT program.

In this meta-analysis, continuity refers to completion of STEM programs for students at all levels (K-12, undergraduate, and graduate). Empirical data used in previous studies of the measure of continuity included graduate rate, enrollment in STEM course, internships and interests in employment in STEM fields.
3.3 Designs of Previous Studies Examined for Effectiveness of STEM Programs

The methods used in all of the previous research analyzed herein have been based upon either (a) within-subject design or (b) between-subject design. These research designs can both help to control for potentially confounding variation and to ensure the isolation of student and/or teacher knowledge and skill levels in examining the attitude, attraction, and retention of students and/or teachers toward STEM fields. Nonetheless, each type of design brings its own logical limitations. As a result, the research on STEM education program effectiveness is subject to potential research biases attributable to these limitations, especially non-randomization and lack of designed control for confounding variables. Even though some studies have used controlled trials, others have had limitations of time-, sample- and context-specificity that can undermine their applicability, relevance and usefulness in other contexts (Davies, 2003). These design problems limit the validity of results of STEM program evaluation.

Recognition of the design problems in previous studies of the effectiveness of STEM programs implies the usefulness if not necessity of a meta-analysis of STEM program effectiveness. Many previous studies examined the effectiveness of STEM programs but they generated inconsistent results. Only a meta-analysis can statistically integrate the findings of previous studies to present a comprehensive trend of the effectiveness of STEM programs. Therefore, this section reviews methods used in the individual studies on the effectiveness of STEM programs.
3.3.1 Within-Subject Design (Pretest-Posttest Design)

In a within-subject design, an outcome variable is measured for the same subject on multiple occasions. The most common within-subject design is conducted when individual participants are measured once before a treatment and once after a treatment then the researcher wishes to determine if there is a significant difference between these measurements. The first measurement is called the pretest, or a baseline measurement, and the second measurement is called the posttest measurement (Bonate, 2000).

A within-subject design has limitations attributable to changes associated with the passage of time between tests conducted during the study. For example, maturation and/or any outside events that occur during the time when a within-subject experiment is conducted may potentially have an influence on the participants’ scores. In addition, any physiological or psychological changes that occur during the time the within-subject experiment is conducted can influence participants’ scores. Finally, the within-subject design threatens internal validity vis-à-vis the pretest effects on the results of posttest in a treatment group. The familiarity of items due to the exposure from the pretest might increase the scores on the posttest. There has been argument about the effect of pretesting on posttest scores (Cambell & Stanly; 1963, Bracht & Glass; 1968, and Welch & Walberg; 1970).

Although a within-subject design is disadvantageous in that it threatens internal validity vis-a-vis history, maturation, and pretest effects, there are two distinct advantages of using this design. The design essentially eliminates all of the problems based on individual differences that are the primary concern of a between-subjects design. In other words, the variability among participants is intrinsically limited by comparing subjects to
themselves, albeit at a different time. Another obvious advantage of using this design is that fewer participants are needed in a study. This saves time in recruiting participants, especially if participants have certain characteristics that are not common to most populations (Gliner & Morgan, 2000).

With the purpose of measuring student/teacher behavior change resulting from STEM programs, quite a few studies use a within-subject design without a comparison group (Jeanpierre, Oberhauser, & Freeman, 2004; Stake & Mares (2001, 2004); Yore & Anderson, 2004). Jeanpierre, Oberhauser, and Freeman (2005) compared teachers’ participation before and after a Professional Development program that provided a science teaching experience intended to demonstrate the teachers’ understanding of research techniques without a control group. After the program, the participants were significantly more likely to understand science content relative to before the program. Stake and Mares’ (2001, 2004) studies implemented pre-test and post-test designs regarding students’ attitudes toward science without a comparison group. Their pre-post comparisons did not indicate the presence of a positive impact on the participants’ attitudes toward science. Shymansky, Yore, & Anderson (2004) demonstrated the insignificant effect of the Science, Parents, Activities and Literature program (PALs) on students’ achievement and attitude toward science on multiple-year comparisons.

3.3.2 Between-Subject Design (Treatment versus Control Group)

Between-subject designs are widely used in STEM program evaluation, primarily for the purpose of comparing groups and/or measuring change resulting from participation in such programs. It is desirable for the experimental group and control
group to be as similar as possible except that the experimental group went through the
treatment while the control group did not. The classical method of insuring such
similarity is through the use of randomization of subject assignment to groups. Although
most researchers would like to incorporate randomization of group assignment in their
design, various practical, ethical, and political realities have often prevented them from
doing so. Therefore, many studies have had to settle and have used naturally occurring
groups when conducting a study based upon a between-subject design.

Without randomization of group assignment, a between-subject study risks the
possibility that systematic but unobserved differences between individuals within the
experimental group and control group will cause variation in the outcome variable. A
nonequivalent control group in the quasi-experimental method is the predominant design
for research in STEM program evaluations. A “nonequivalent control group design”
means that an intervention and a control group was used, but assignment to groups was
not randomized and there was no matching of subjects between groups (Cook &
Campbell, 1979; Singleton & Straits, 1999). The nonequivalent method can suffer from
assignment bias and individual differences which threaten internal validity (Shadish,
Cook, & Campbell, 2002). An experimental-versus-control group design is exposed to a
potential internal validity threat whenever the nonrandom assignment of various
individuals to groups leads to bias.

There are numerous STEM evaluation studies that have employed quasi-
experimental design with a nonequivalent control group (Barlow & Villarejo 2004; Lott,
2003; Adamson, Banks, Burtch, et al., 2003; Maton, & Schmitt, 2000). The studies with a
control group showed inconsistent results of the STEM programs on engagement,
capacity, and continuity of students or teachers. Some studies are significantly positive (Barlow & Villarejo; 2004, Lawrenz, Huffman, & Gravely; 2007), negative (Shymansky, Yore, & Anderson; 2004, Lott; 2003) or insignificant (Harty, Samuel, & Andersen; 1991). For example, Barlow and Villarejo (2004) compared a treatment group to a control group to demonstrate whether the BUSP program had a positive effect on undergraduate students who were pursuing graduate study. As a result of the program, participants were found to have had a greater likelihood of going on to graduate study. In contrast, Lott (2003) demonstrated that students who participated in the Alabama Science in Motion (ASIM) program did not have better attitudes towards science than students who did not participate.

### 3.3.3 Conclusion

For the purpose of comparing groups and/or measuring change resulting from experimental treatments, pre-test/post-test designs are widely used in assessing outcomes of STEM programs. Meta-analysis explicitly recognizes that variation in study designs might itself yield variation in regard to estimates of programs on the engagement, capability, and continuity of students and teachers in STEM fields. In this study, in order to control for the design of each individual study contained in the analysis, the two types of designs for the meta-analysis are broadly categorized as 1) within-subject and 2) between-subject designs in randomized experimental, quasi-experimental, and non-experimental methods.
3.4 Research Findings on the Effectiveness of STEM Programs

Some studies on the STEM programs’ effectiveness have shown interesting successes both in teachers’ progress and in students’ demonstrated mastery. Shymansky, Yore, and Anderson (2004) for instance discovered that the program “Science, Parents, Activities and Literature Program” (PALs) had a positive impact on third- and fourth-grade student attitudes toward science. Evidently the most successful program was the Louis Stokes Alliances for Minority Participation Program (LSAMP) (2005) conducted in 50 states with a sample of 596 teachers and 2,220 students during the three-year period of 2003-2006. The LSAMP program was aimed at increasing the quality and quantity of students successfully completing STEM related baccalaureate degree programs. The participants were underrepresented minorities, and they were found to have been significantly more likely to pursue Master’s and doctoral programs in STEM fields relative to those minority group members who did not participate in the national program. The LSAMP graduates also exceeded the national rate of graduate degree completion for underrepresented minorities (and non-underrepresented minorities) in national samples.

Unlike the success of some STEM programs, however, other programs lack conclusive evidence of their effectiveness. Minger and Simpson (2006) studied the impact of a standards-based science course, Science 226, on student attitudes toward science. The Science 226 course was taught by instructors from the departments of biology, earth science, and chemistry in the college of Science and Engineering at St. Cloud State University. Their results did not show any positive or negative effect on the attitudes of the 120 students who were enrolled in the class. Moseley, Reinke, and
Bookout (2003) studied the effects of a training- and-teaching program for chemistry education on teachers’ attitudes towards teaching chemistry. Comparison of teachers’ beliefs about their own teaching ability between the treatment and comparison groups revealed no significant difference in teachers’ beliefs regarding the development of a teaching strategy in elementary teachers.

Finally, several studies have reported mixed results (Barlow & Villarejo, 2004; Kyle, Bonnstetter, & Godsden, 1988, Carney, Chawla, Wiley, & Young, 2006; Randy, Steve, & Tad, 1998). The Biology Undergraduate Scholars Programs (BUSP), located at the University of California – Davis, aimed to address university-wide racial/ethnic disparities in graduation in the biological sciences. Barlow and Villarejo (2004) found that the BUSP participants were more likely to successfully complete basic science and math courses. In addition, the mean grade point average (GPA) of BUSP students who persisted in general biology was substantially higher than that of the non-BUSP students. However, in calculus classes, BUSP students had essentially the same mean GPA as did the non-BUSP students.

A review of the narrative literature reports on the results with the specific information about individual characteristics of individual studies on STEM program effectiveness. For instance, the review above describes which outcome variables, what types of STEM programs, which courses, what sample size, and what research designs have been used in studies. However, as studies accumulate, it becomes difficult to systematically and logically process the information from all of the results from the many constituent studies and integrate their results into meaningful summary statements. Yet this is exactly the purpose of formal meta-analysis. The limitations of narrative review of
the relevant studies warrant the use of meta-analysis to comprehensively synthesize all relevant studies.

3.5 Limitations of Primary Literature

A review of the existing studies on STEM programs with three measures for the effectiveness of the STEM programs revealed inconsistent results. According to King and He (2005), narrative reviews present verbal descriptions of past studies focusing on theories and frameworks, elementary factors and their roles (predictor, moderator), and/or research outcomes regarding a hypothesized relationship. Bowen (2008) stated that narrative literature reviews are subjective, as the review results are heavily influenced by the reviewers’ idiosyncratic ideas, beliefs, and values. One researcher may selectively include a limited number of studies to support his or her own research agenda while another researcher may simply select a different set of previous studies and come to a completely different conclusion reviewing the same type of studies. Bowen (2008) pointed out that reviewer subjectivity intervenes in the process of selecting the previous studies and then manifests itself in the various results of different reviewers.

While a narrative review can be somewhat informative for policy makers, the process is embedded with several problems that make it susceptible to potentially high levels of unrecognized bias. Some narrative reviewers often do a summary on a group of studies without clearly describing their methodology. This lack of description makes it extremely difficult, if not impossible, for other researchers to replicate their study and/or examine the validity of their findings (Petrosino & Lavenverg, 2007; King & He, 2005;
This inherent bias makes it all too common for researchers working from the same set of studies to arrive at different conclusions.

Reviewers too often have not clearly and explicitly stated their standards for determining what studies were included in a narrative review. Studies were selected at the potentially arbitrary discretion of the reviewers. At its worst, a reviewer advocating a position could selectively include only those studies favoring that viewpoint (Petrosino & Lavenberg, 2007). Moreover, unpublished studies are sometimes excluded from narrative reviews.

As the number of studies on a phenomenon of interest increases, narrative reviewers have an increasingly difficult task combining these findings in the absence of any standardized approach. Houston, Peter, and Sawyer (1983) likened the process to one in which reviewers seek to make sense out of many data points that result from a single research project. Extracting meaning from the many studies in narrative fashion is not unlike an attempt to extract meaning from a raw data matrix prior to reduction.

Narrative reviewers have been further criticized for failing to critically examine the characteristics, design, effect, and conclusions drawn from independent studies and for overlooking the impact of moderating variables. When a narrative review tries to examine relationships among outcomes and potential moderator variables, large and too-often-ignored difficulties arise in efforts to consider every moderator that might influence estimates of the size of the effect of the independent (causal) variable upon the dependent (effect) variable. Shadish, Cook, and Campbell (2002) pointed out that the narrative reviewers must also consider imprecision in the moderator variables.
Some narrative reviewers strive to minimize these biases by clearly indicating search criteria, including unpublished studies and delineating study boundaries. Even when this is done, however, narrative reviews still do not provide a measure of the effect size. In contrast, and to its credit, meta-analysis allows an investigator to estimate effect sizes in the process of integrating inconsistent results from a vast number of studies on a given topic (Glass, 1976; Cook et al, 1992; McMillan & Schumacher, 1984).

### 3.6 Overview of Meta-analysis

It is difficult to generalize the findings of individual studies that have evaluated STEM programs. Numerous individual studies on STEM programs have been conducted in a variety of settings and conditions, and they have considered aspects such as measures of effectiveness, research designs, contexts of study, and so on. Therefore, inconsistent results from previous studies provide a need for a statistical integration of their findings.

Glass (1976) proposed three levels of data analysis: primary-, secondary-, and meta-analysis. He defined primary analysis as “the original analysis of the data in a research study” and secondary analysis as “the re-analysis of the data for the purpose of answering the original research question with better statistical techniques, or answering new questions.” Glass coined the term meta-analysis, “the analysis of analysis,” to refer to “the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings” (p. 3).

Meta-analysis is most useful in integrating inconsistent results from previous studies. Davies (2003) warned: “Single studies, even if they are randomized controlled
trials or other types of experimental inquiry, have limitations of time-, sample- and context-specificity which can undermine their applicability, relevance and usefulness in other contexts” (p. 366). Cook, Cooper, and Cordray (1992) emphasized that it is difficult to generalize the findings of a single study as single studies tend to “illuminate only one part of a large explanatory puzzle” (p.3). Green and Hall (1984) stated, “A single study is never definitive no matter how memorable and newsworthy it may be” (p.38). King and He (2005) added:

Meta-analysis enables the combining of various results, taking into account the relative sample sizes, thereby permitting studies showing insignificant results to be analyzed with studies showing significant effects. The overall result may be either significant or insignificant, but it is undoubtedly more accurate and more credible because of the overarching span of such analysis (p. 671).

Andrews and Harlen (2006) have concluded that it is impossible for any individual to be able to read and formulate a coherent opinion from a vast body of literature on a particular topic. However, a meta-analysis allows a researcher to make robust conclusions from many previous studies through the application of sophisticated methods. Meta-analytical studies generally tend to produce more reliable results than traditional reviews because the unobserved and unrecognized bias of the investigator is minimized (Bowen, 2008).

A meta-analysis generally posts a focused, empirical question, whereas a narrative review usually posts a general and broad question (Cook, Mulrow, & Haynes, 1997).
Most meta-analyses in the behavioral sciences have been exploratory in nature (Durlak, & Lipsey, 1991). Durlak and Lipsey (2008) suggested that a meta-analysis offers either a specific hypothesis *a priori* or at least identifies the important theoretical, conceptual, or procedural questions to be addressed. As Richardson, Wilson, Nishikawa, and Hayward (1995) stated, the question in meta-analysis can be formulated explicitly according to four variables: a specific population and setting, the condition of interest, an exposure to a test or treatment, and one or more specific outcomes.

The next feature, sources and search, has been identified by Bowen (2008) to be subjective in the context of traditional review because of the selection process and the limited number of selected studies. In addition, the connection between clinical recommendations and evidence in narrative review is often “tenuous, incomplete, or worse still—based on a biased citation of studies” (Ravnskow, 1992; Neihouse & Priske, 1989). However, the sources for meta-analysis must be as comprehensive as possible. The comprehensive search procedure of a meta-analysis account for this shortcoming in a traditional review study, and consequently, a meta-analysis reduces the bias by including numerous relevant studies.

In addition, the search for the related studies on the STEM effectiveness should be as exhaustive as possible. Durlak and Lipsey (2008) and Bowen (2008) suggested an exhaustive search of all databases, and formal and informal channels: manual journal search, examination of references list from reviews and identified studies, computer and manual searches of abstracting and indexing databases, and contact with persons or organizations likely to have produced or to know of studies.
In regard to selection standards, those specific standards prescribed for meta-analysis are totally different from those for narrative review. It is unlikely that narrative review has certain, definite, and explicit standards for selection of the studies to be summarized. However, an investigator conducting a meta-analysis must carefully specify the inclusion criteria used for retaining previous studies (Bowen, 2008). A reviewer must provide inclusion and exclusion criteria and keyword mapping (minimizing data entry errors) (Durlak, Lampman, & Wells, 1991). The explicit selection criteria in meta-analysis help to prevent the researcher’s subjectivity from entering into the results of the study of a given topic.

The quantitative coding of study characteristics permits researchers to keep track of a large amount of potentially important information and then conduct a more detailed breakdown of this data. In other words, Durlak and Lipsey (1991) stated that meta-analysis can “easily handle scores of variables from hundreds of studies” (p. 293). In addition, Wolf (1986) cautioned researchers to code the quality of the design of studies, and to determine whether “the results differ for poorly designed studies and well designed studies” (p.39). Categorizing studies into those that are published and those that are unpublished is an example of the coding of a study characteristic.

In a traditional narrative review, synthesis is a qualitative summary; in a meta-analysis, synthesis is a quantitative summary. A narrative review summarizes different primary studies into a holistic interpretation shaped by the reviewers’ own experience, existing theories and models (Campbell Collaboration, 2001; Kirkevold, 1997). A meta-analysis is a statistical technique offering the possibility of classifying and measuring the conditions and results of many individual studies in a more precise and rigorous way than
traditional reviews (Cook, Mulrow, & Haynes; 1997). As a result, by quantitatively combining the results of numerous studies, meta-analyses can create more precise, powerful, convincing, and unbiased conclusions.

The final advantage of meta-analyses is the ability to assimilate many different effect sizes. An effect size is a way to standardize the outcomes from previous studies and to make comparisons of outcomes under different moderators. One underlying assumption of meta-analyses is that every effect size is independent from others (Smith & Glass, 1977). For instance, a single study reports multiple outcomes associated with effects of STEM programs (e.g. both engagement and continuity). A meta-analysis permits calculating multiple measures of effects related to each separate outcome measure.

Figure 3-1 illustrates a comparison between narrative review and meta-analysis. The application of a meta-analysis to STEM programs provides a more concise understanding of the research and synthesis of individual studies. To conduct a meta-analysis, first of all, a researcher begins with a unique research question. In the next phase, the researcher has to identify and locate the relevant studies, as many as possible, based on the specific criteria. Such criteria-based selection is one of the distinguishing characteristics of meta-analysis compared with the narrative review. The coding process in phase 5 allows the researcher to take account of the quality of the constituent studies in calculating the effect sizes, which make meta-analysis possible. Some of the ways to consider the quality of studies, for instance, are whether or not studies are published in peer reviewed journals, or in trade publications, as well as whether they are based upon random samples or non-equivalent samples. Finally, meta-analysis allows the calculation
of effect sizes. Each effect size in the statistical calculation can be examined across studies while independent events (e.g. one observation) can be examined as well. The detailed procedure for the meta-analysis of the effectiveness of STEM programs will be more precisely addressed in the methodology section.
Figure 3 - 1 Procedure of Meta-analysis on effectiveness of STEM programs

Phrase I: research question
- research question
  (Effectiveness of STEM programs)
  engagement
  capability
  continuity

Phrase II: source of research studies
- comprehensive source
  computer
  printed journal
  reference

Phrase III: searching for studies
- exhausted search
  key words

Phrase IV: selecting studies
- criteria-based selection
  evaluating STEM program
  with three outcomes
  in United States
  about students/teachers
  empirical studies

Phrase V: coding process
- systematic coding
  dependent variables
  independent variables
  moderator variables
  study characteristics

Phrase VI: statistical calculation
- effect sizes
  mean/standard deviation
  F-value from ANOVA
  t-value
  $\chi$-square

synthesis
quantitative analysis
CHAPTER IV

METHODOLOGY

4.1 Research Questions and Hypotheses

This chapter describes the methodology applied to test the following hypotheses:

H1:

STEM programs in the United States show positive effects in engagement outcomes.

If the overall effect-size of engagement outcomes is positive, this hypothesis is supported by the data. One would expect to see positive improvement in students’ and teachers’ interests and involvement after large amount of federal funds have been invested in the STEM programs. If the overall effect size of engagement outcomes is insignificant or even negative, this hypothesis is not supported by the data. The less than
positive outcomes may serve as a warning sign on the effectiveness of current STEM programs and may call for examination, investigation and further accountability of the current STEM programs. However, whether or not there is significant difference of the effect sizes resulting from within-subject design versus between-subject design remains an empirical question.

H2:

STEM programs in the United States show positive effects in capability outcomes.

If the overall effect-size of capability outcomes is positive, this hypothesis is supported by the data. One would expect to see positive improvement in students’ and teachers’ skills and knowledge after experiencing STEM programs funded by the federal governments. If the overall effect size of capability outcomes is insignificant or even negative, this hypothesis is not supported by the data. However, whether there is significant difference of the effect sizes resulting from within-subject design versus between-subject design remains an empirical question.

H3:

STEM programs in the United States show positive effects in continuity outcomes.

If the overall effect-size of continuity outcomes is positive, this hypothesis is supported by the data. One would expect to see positive impact on students’ continuous involvement in STEM fields after experiencing STEM programs funded by the federal
government. If the overall effect size of continuity outcomes is insignificant or even negative, this hypothesis is not supported by the data. However, whether there is significant difference of the effect sizes resulting from within-subject design versus between-subject design remains an empirical question.

4.2 Research Database and Search Procedure

Computer searches

Research studies were located by querying research databases for articles. Databases included in this meta-analysis were Education Research Complete (ERC) from 1998 to 2009 (1998 being the earliest year content electronically available), Academic Search Complete (ASC) from 1998 to 2009 (1998 being the earliest year available), Computer’s Applied Science Complete (CASC) from 1985 to 2009, and Education Resource Information Center (ERIC) from 1981 to 2010.

Manual Searches

Electronic searches are automatically more convenient than manual searches. However, they are not more accurate than manual searches (Guzzo, Jackson, & Katzell, 1987). To more accurately and completely search, the Journal of STEM Education from 2001 (first volume) to 2010, the Journal of Elementary Science Education from 1989 (first volume) to 2010, the Journal of Engineering Education from 2005 to 2010 (2005 being the earliest year content available), and the Journal of Research in Science Teaching from 1981 to 2010 were manually searched. The author reviewed the reference list of each relevant study to assess whether certain journals are a main publication outlet for research in STEM-related education areas. The author identified these five journals as good outlets for research in STEM-related fields because the total number of relevant studies for meta-analysis was predominantly found in these five journals. The total number of studies included in this meta-analysis is 91 which are listed by journals in Appendix 1.

Search Procedure

Stage 1: The electronic search was conducted using the following combination of keywords: ‘stem program,’ ‘pre-test,’ and ‘post-test’ or ‘stem program,’ ‘experimental group,’ and ‘control group.’ When the databases did not allow searching for combinations of keywords, only the ‘stem program’ was used in the search. The resulting articles are listed in Table 4.1.
Stage 2: After all of the relevant studies in the universe were identified in Stage 1, the author examined the titles. The author selected studies for possible inclusion; when the titles contained the following words that measure program outcomes: assess, evaluate, improve, promote, enhance, effect, efficacy, results, impact, develop, learning or words that measure effectiveness of STEM program (i.e., attitude, interest, belief, engagement, skill, knowledge, capacity or continuity). The total number of these articles was listed in Stage 2 of Table 4-1.

Stage 3: The author then reviewed the abstracts of these articles. The research excluded studies when their abstracts identified a target of other countries, had a target group other than student/teacher, were not concerned with any one of the three outcomes (engagement, capacity, and continuity), or did not use empirical methods. The number of articles that remained after this stage is identified in Stage 3 of Table 4-1.

Stage 4: The methodology and results sections in each study were reviewed. All studies to be considered had to be quantitative in nature with quasi-experimental designs. Appropriate empirical studies on the effectiveness of the relevant STEM program had to have either a pretest/posttest design or a control-treatment group comparison. In addition, each study had to provide sufficient statistical information to calculate an effect size for a meta-analysis. The number of articles that remained after this stage is identified in Stage 4 of Table 4-1.
<table>
<thead>
<tr>
<th>Journal Title</th>
<th>Year</th>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Stage 3</th>
<th>Stage 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educational Evaluation and Policy Analysis</td>
<td>1981-2010</td>
<td>134</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Journal of Science Education and Technology</td>
<td>1992*-2010</td>
<td>107</td>
<td>48</td>
<td>35</td>
<td>11</td>
</tr>
<tr>
<td>Science Education</td>
<td>1996-2010</td>
<td>60(1)</td>
<td>14(0)</td>
<td>8(0)</td>
<td>2(0)</td>
</tr>
<tr>
<td>American Journal of Evaluation</td>
<td>1996-2010</td>
<td>151(75)</td>
<td>0(2)</td>
<td>0(0)</td>
<td>0(0)</td>
</tr>
<tr>
<td>Journal of Negro Education</td>
<td>1981-2010</td>
<td>46</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Journal of STEM education</td>
<td>2001*-2010</td>
<td>manual</td>
<td>17</td>
<td>16</td>
<td>7</td>
</tr>
<tr>
<td>Journal of Elementary Science Education</td>
<td>1989*-2010</td>
<td>manual</td>
<td>51</td>
<td>38</td>
<td>21</td>
</tr>
<tr>
<td>Education Resource Information Center</td>
<td>1981-2010</td>
<td>113</td>
<td>34</td>
<td>29</td>
<td>9</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>239</strong></td>
<td><strong>164</strong></td>
<td><strong>91</strong></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Stage 1: Total number of Titles from search  
Stage 2: Number retained after reading title  
Stage 3: Number retained after reading abstract  
Stage 4: Number retained after methodology or context  
* refers to a starting year of the journal website 1981-2010  
( ) Number journals through its own

### 4.3 Selection Criteria

Studies considered for inclusion must meet the following criteria set by the author:

a. Published between 1981 and 2010 due to STEM education program variation.

b. Investigate the effects of STEM programs in the United States.

c. Include a program evaluation about science, technology, engineering, or mathematics.

d. Provide outcome variables including engagement, capability or continuity.
e. Consider a target of K-12 students or teachers, college students or teachers, undergraduate students or teachers, or graduate students or teachers.

f. Investigate the effects of STEM programs on the general population.

g. Include a treatment group (with STEM) and non-equivalent control group (without STEM) research design or pretest/posttest design.

h. Report sufficient empirical data on the effectiveness of the STEM program to calculate effect sizes assessing the effectiveness of STEM programs.

Several conceptual criteria formed the basis for the exclusion of a study from the analysis

a. Studies targeting other countries instead of the United State were excluded

b. Studies targeting STEM programs for students with special needs were excluded

c. Studies targeting parents, adults, or administrators of the program were excluded.

4.4 Variables Examined in the Analysis

As suggested by Wolf (1986), the variables used in the studies included in this meta-analysis were coded into categories (Table 4-2) to explain the various sources of error variance. The categories include the independent and dependent variables in operational terms to be examined in the meta-analysis as well as moderator variables.
Dependent Variable

Three most commonly used outcome measures in STEM programs were used as the dependent variable. Effect sizes for engagement, capacity, and continuity of students and teachers reported from the individual studies were the dependent variables.

Independent Variables

The independent variable in the study was the research design for the program evaluation used in individual studies to analyze the effectiveness of STEM. The research designs used in these previous studies are either experimental versus control group or pretest-posttest assessments. Students or teachers who participated in STEM programs that were being observed comprised the experimental groups. Control groups were primarily comprised of students or teachers who did not participate in the STEM program being examined. Between-subject design is usually illustrated by comparing outcomes from these two groups of participants. On the other hand, within-subject design is usually illustrated by comparing outcomes from the same group of participants measured twice. The first measure was taken before they participated in STEM programs and the second measure was taken after the participation.

Moderator Variables

Moderator variables, or study characteristics, are variables that can be expected to change the direction or magnitude of the relationship between dependent and independent variables (Bowen, 1997). To address this issue, this study broadly classified five categories of moderators: STEM program variation, participant characteristics,
setting/environment characteristics, study characteristics of STEM programs, and the method used in the various studies. These five categories of moderator variables consisted of two to five “sub-class” moderators (see Table 4-2 for more details).

The program characteristics moderator variable was sub-classified into four categories: program creator, program strategy, program funder, and program field. The participant characteristic was sub-categorized into student ability and grade levels. Setting characteristics were sub-classified into regional location and school type. The study method characteristics were sub-classified into four categories: reliability evidence, subject assignment, testing instrument, and treatment duration to explore potential explanations for the variance in the overall effectiveness of STEM programs. Finally, study characteristics were sub-classified into two categories: published year and research funder. For instance, Marzano(1998) states that the goals of science programs are achieved through the use of different pedagogical strategies. Teachers should choose appropriate strategies to achieve their goals of improving student knowledge in STEM fields. If teachers choose inquiry-based strategy, students will be encouraged to pursue their own questions about science and to have a sense of the significance of the results of their inquiry. According to Schroeder, Scott, and Tolson’s study (2007), the enhanced context strategy was among the most successful approaches in science programs for student achievement in the United States. Their study created six classifications: enhanced material strategies, assessment strategies, inquiry strategies, enhanced context strategies, instructional technology (IT) strategies, and collaborative learning strategies. Different pedagogical strategies used in the STEM programs were considered a
moderator variable because different pedagogical strategies might create different levels of effectiveness for the STEM programs.
Table 4 - 2 The Study Characteristics by Categories

<table>
<thead>
<tr>
<th>Variables</th>
<th>Levels</th>
<th>Study Feature</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent Variables</strong></td>
<td></td>
<td>Engagement</td>
<td>Capability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Continuity</td>
<td></td>
</tr>
<tr>
<td><strong>Independent Variables</strong></td>
<td>Outcome measures</td>
<td>Research design for the program evaluation</td>
<td>Between-subject design</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Within-subject design</td>
</tr>
<tr>
<td><strong>STEM Program Characteristics</strong></td>
<td>Type of STEM Program strategy</td>
<td>Inquiry strategy</td>
<td>Enhanced context strategy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Instructional technology strategy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Collaborative learning strategy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Enhanced material strategy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A combination of the above</td>
</tr>
<tr>
<td><strong>Field Type</strong></td>
<td>Science</td>
<td></td>
<td>Technology</td>
</tr>
<tr>
<td></td>
<td>Engineering</td>
<td></td>
<td>Mathematics</td>
</tr>
<tr>
<td></td>
<td>STEM (a combination of the above)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Program Funder</strong></td>
<td>Government</td>
<td></td>
<td>Nonprofit organization</td>
</tr>
<tr>
<td></td>
<td>Private school</td>
<td></td>
<td>Mixed</td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td></td>
<td>No information</td>
</tr>
<tr>
<td><strong>Program creator</strong></td>
<td>Professional association</td>
<td></td>
<td>Government</td>
</tr>
<tr>
<td></td>
<td>Government</td>
<td></td>
<td>Non-profit organization</td>
</tr>
<tr>
<td></td>
<td>Private organization</td>
<td></td>
<td>Mixed(cooperative)</td>
</tr>
<tr>
<td></td>
<td>Mixed(cooperative)</td>
<td></td>
<td>No information</td>
</tr>
<tr>
<td><strong>Participant Characteristics</strong></td>
<td>Ability Level</td>
<td>Gifted</td>
<td>Average</td>
</tr>
<tr>
<td></td>
<td>Underachieving/below grade level</td>
<td></td>
<td>Mixed (above)</td>
</tr>
<tr>
<td></td>
<td>Mixed (above)</td>
<td></td>
<td>Other (teacher)</td>
</tr>
<tr>
<td><strong>Grade Level</strong></td>
<td>Elementary school</td>
<td></td>
<td>Middle school</td>
</tr>
<tr>
<td></td>
<td>High school</td>
<td></td>
<td>College</td>
</tr>
<tr>
<td></td>
<td>K-12 student</td>
<td></td>
<td>K-12 student</td>
</tr>
<tr>
<td><strong>Setting Environment</strong></td>
<td>Regional Location</td>
<td>West</td>
<td>Midwest</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Northeast</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>South</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Nationwide</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No information</td>
</tr>
</tbody>
</table>
School Type
- Public school
- Private school
- Private school with a religious affiliation
- Mixed (public and private)
- No information

Methodology

<table>
<thead>
<tr>
<th>School Type</th>
<th>Subject Assignment</th>
<th>Testing Instrument</th>
<th>Reliability Evidence</th>
<th>Treatment Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public school</td>
<td>Random assignment</td>
<td>Instructor design</td>
<td>More than &gt;.7</td>
<td>One class</td>
</tr>
<tr>
<td>Private school</td>
<td></td>
<td>Existing instrument</td>
<td>Less than &lt;.7</td>
<td>Several classes</td>
</tr>
<tr>
<td>Private school</td>
<td></td>
<td></td>
<td></td>
<td>Entire semester</td>
</tr>
<tr>
<td>with a religious</td>
<td>Nonrandom assignment</td>
<td>GPA/course score</td>
<td>No information</td>
<td>More than one semester</td>
</tr>
<tr>
<td>affiliation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed (public and</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>private)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No information</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Study Characteristics

<table>
<thead>
<tr>
<th>Publication Type</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Published studies</td>
<td>1981-1990</td>
</tr>
<tr>
<td>Unpublished studies</td>
<td>1991-2000</td>
</tr>
<tr>
<td></td>
<td>2001-2010</td>
</tr>
</tbody>
</table>

**Types of STEM Programs’ Strategy**

Following Schroeder, Scott, & Tolson (2007), the pedagogical types of STEM programs in studies included in this meta-analysis were placed into six categories: inquiry strategy, enhanced context strategy, instructional technology strategy, collaborative learning strategy, enhanced material strategy, or combinations of one or more of them. Each program described in the individual study belongs to one of the following categories based on the characteristics of the strategy. Inquiry strategy involves student-centered instruction that student answer scientific research questions by analyzing data such as guided inquiry activities or laboratory inquiries. Although inquiry projects are very similar to inquiry strategy, Schroeder, Scott, and Tolson (2007) identify the presence of teamwork in inquiry projects as the distinguishing characteristic. Collaborative
learning strategy refers to teachers arranging students in flexible groups to work on various tasks such as inquiry projects and discussions. Enhanced context strategy involves learning by experience. Instructional technology strategy involves use for technology such as computers or videos to emphasize science concepts. In Enhanced materials strategy, the teacher modifies the instructional materials, such as the teacher annotates text materials. Schroeder, Scott, and Tolson (2007) modified the pedagogical categories of science programs, which were established by Wise (1996). Given the variation in STEM programs’ effectiveness, the studies collected can be used to examine whether the effect of STEM programs were consistent across the pedagogical types.

Field Type

STEM academic fields were identified by five different programs: science, technology, engineering, mathematics and STEM (a combination of these fields). Given the variation in STEM programs’ effectiveness, the studies collected can be used to examine whether the effect of STEM programs were consistent across these different field types.

Program Funder

The program funding types in studies included in this meta-analysis were placed into four categories: the government (all levels), non-profit organizations, private schools, and mixed funding (i.e. combination of two or more organizations). Given the variation in STEM programs’ effectiveness, the studies collected can be used to examine whether the program effects were consistent across these different program funding sources.
Program Creator

STEM programs were created by four different organization types: professional associations (e.g. the National Teacher Association and teachers), government (federal, state, or local government), non-profit organizations (e.g. private and public schools), private organizations, and mixed organizations. Given the variation in STEM programs’ effectiveness, the studies collected can be used to examine whether the effect of STEM programs were consistent across these different program creators.

Ability level of Sample Group

Each participant group in STEM programs in this study were categorized under one of the following categories: gifted, average, underachieving/below grade level, or mixed (e.g. the combination of gifted, average, and underachieving). Given the variation in STEM programs’ effectiveness, the studies collected can be used to examine whether the effect of STEM programs were consistent across these different ability levels.

Educational level of sample group

Grade levels of the participants in samples on STEM programs were classified by seven different levels: elementary school, middle school, high school, two or four years of college, graduate school, K-12 students, and K-12 teachers. Given the variation in STEM programs’ effectiveness, the studies collected can be used to examine whether the effect of STEM programs were consistent across these educational level of sample groups.
**Regional location**

Some studies on science programs (Nunez & Bowers, 2011; Bredderman, 1983) found that a geographical context affects whether students continue to study in science areas. Nunez and Bowers (2011) found that retention in college is more likely in the South region. This result provides insight for this meta-analyst to consider the regional levels as a moderator variable. Regional locations were classified into five regions in which the programs were held: West, Midwest, Northeast, South, and nationwide. Some locations were also classified as having no information. Given the variation in STEM programs’ effectiveness, the studies collected can be used to examine whether the effect of STEM programs were consistent across these regionally different locations.

**School Type**

Types of schools which STEM programs were conducted in were classified into three categories: public school, private school, and private school with a religious affiliation. Given the variation in STEM programs’ effectiveness, the studies collected can be used to examine whether the effect of STEM programs were consistent across these school type.

**Subject Assignment**

Subject assignment in the design of these studies on the effectiveness of STEM programs consisted of random assignment or nonrandom assignment. Given the variation in STEM programs’ effectiveness, the studies collected can be used to examine whether
the effect of STEM programs were consistent between random assignment and nonrandom assignment.

Testing Instrument

The testing instrument used in this collection of studies varied. The instrument is another variable to measure the effectiveness of STEM programs. Of the three types of instrument, one set was derived by the instructors/researchers of specific studies. Another set of assessment measures included state board examinations or standardized course evaluations. The other measures used included GPA or course grades. Given the variation in STEM programs’ effectiveness, the studies collected can be used to examine whether the effect of STEM programs were consistent across the testing instruments.

Reliability of the Evidence

Another variable to consider is the reliability level of the measurement instrument. Cronbach (1970) suggests that the sufficient reliability level is “more than .7 reliability.” The reliability of evidence levels were placed into three categories: “more than .7 reliability”, “less than .7 reliability”, and “no information.” If the reliability level was explicitly stated at a reliability >.7, the studies were identified “more than .7”. If the reliability level was explicitly stated at a reliability <.7, the studies were identified “less than .7”. The studies that did not provide the reliability level of the measurement instrument were identified “no information” categorization.
Treatment Duration

Another variable to consider is treatment duration of STEM programs. The duration of STEM programs varied among the collected studies. STEM programs ranged in duration from a few months to five years. Treatment durations were classified by three different durations: “less one semester”, “entire semester”, and “more than one semester.” Given the variation in STEM programs’ effectiveness, the studies collected can be used to examine whether the effect of STEM programs were consistent across treatment duration.

Publication Type

The publications examined in this meta-analysis consist of published studies and unpublished studies. Studies were coded by type of publication to determine whether published studies were more likely to report a positive effect size than unpublished studies. Given the variation in STEM programs’ effectiveness, the studies collected can be used to examine whether the effect of STEM programs were consistent between publication type.

Publication Time

This study sought to determine whether or not studies published in different years had consistent results. There were three categories of published years: the studies published from 2000 to 2010, those published from 1990 to 1999, and from 1980 to 1989. Given the variation in STEM programs’ effectiveness, the studies collected can be used
to examine whether the effect of STEM programs were consistent across these published years.

4.5 Statistical Calculation

Statistical data from each study was recorded. The recorded data included the number of subjects, mean scores, standard deviations, F-value, t-value, Chi-Square statistics or probability values corresponding to the statistics used in each particular study.

Inter-Coder Reliability of Coding

A random sample of 8 studies (10% of the total sample of studies) was coded by two independent coders to ensure the reliability of the data coding process. Prior to completing the coding process, a one-session meeting was held to familiarize the coders with the study and discuss the items on the coding sheet. After coding, the results from the two coders were compared to determine inconsistencies in the coded information. A discussion session was held following the coding process to determine the cause for any inconsistencies found in the data entry. Adjustments to the items were made thereafter and corrected for the future coding process. Inter-rater reliability was calculated using Cohen’s (1960) formula:

\[ k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \]
Where:

\[ k = \text{kappa value (inter-rater reliability)} \]

\[ \text{Pr}(a) = \text{Observed percentage of agreement}, \]

\[ \text{Pr}(e) = \text{Expected percentage of agreement}. \]

Landis and Koch (1977) characterized the magnitude of Kappa coefficient values < 0 as indicating no agreement, 0 to 0.20 as slight, 0.21 to 0.40 as fair, 0.41 to 0.60 as moderate, 0.61 to 0.80 as substantial, and 0.81 to 1 as almost perfect agreement. Table 4-3 shows Kappa Coefficients for all categorical variables in this study.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Kappa Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent Variable</strong></td>
<td></td>
</tr>
<tr>
<td>Outcome measure</td>
<td>0.950</td>
</tr>
<tr>
<td><strong>Independent Variable</strong></td>
<td></td>
</tr>
<tr>
<td>Research Design</td>
<td>1</td>
</tr>
<tr>
<td><strong>Moderator Variable</strong></td>
<td></td>
</tr>
<tr>
<td>Report ID</td>
<td>1</td>
</tr>
<tr>
<td>Report Type</td>
<td>1</td>
</tr>
<tr>
<td>Research funder</td>
<td>0.673</td>
</tr>
<tr>
<td>Publication Type</td>
<td>1</td>
</tr>
<tr>
<td>Publication Year</td>
<td>1</td>
</tr>
<tr>
<td>STEM Program Strategy</td>
<td>0.945</td>
</tr>
<tr>
<td>Program Target Group</td>
<td>0.588</td>
</tr>
<tr>
<td>Field Type</td>
<td>0.959</td>
</tr>
<tr>
<td>Program Funder</td>
<td>1</td>
</tr>
<tr>
<td>Program Creator</td>
<td>1</td>
</tr>
<tr>
<td>Participant Ability Level</td>
<td>0.500</td>
</tr>
<tr>
<td>Socioeconomic Status (SES)</td>
<td>0.465</td>
</tr>
<tr>
<td>Participant Grade Level</td>
<td>0.833</td>
</tr>
<tr>
<td>Regional Location</td>
<td>1</td>
</tr>
<tr>
<td>School Type</td>
<td>1</td>
</tr>
<tr>
<td>Subject Selection</td>
<td>1</td>
</tr>
<tr>
<td>Subject Assignment</td>
<td>1</td>
</tr>
<tr>
<td>Testing Instrument</td>
<td>1</td>
</tr>
<tr>
<td>Reliability Evidence</td>
<td>0.816</td>
</tr>
<tr>
<td>Treatment Duration</td>
<td>0.940</td>
</tr>
</tbody>
</table>

According to the results of Kappa coefficients for each variable, this study excluded the variables “Participant Ability Level,” “Socioeconomic Status,” “Program Target Group”, and “Research Funder” due to insufficient kappa and retained the rest of the variables with high kappa for further analysis.
The Unit of Analysis

The unit of analysis for this research study is an effect size. In this meta-analysis, the effect size is a standardized measure of STEM program effectiveness in individual studies. Cooper (1998) classified effects sizes of .20 as small, .50 as moderate and .80 as large. Effect size was calculated for those studies based upon a between-subject design by using the formula derived by Hunter and Schmidt (1990)

\[
ES(d) = \frac{X_1 - X_2 (\text{group})}{(SD_p)}
\]

where,

\(ES = \) effect size
\(X_1 = \) the treatment group mean score
\(X_2 = \) the control group mean score
\(SD_p = \) pooled standard deviation of the two groups

The formula for the corresponding pooled standard deviation is:

\[
SD_{pooled} = \sqrt{\frac{(n_1 - 1)sd_1^2 + (n_2 - 1)sd_2^2}{n_1 + n_2 - 2}}
\]

where,

\(n_1 = \) number of subjects in the treatment group
\(n_2 = \) number of subjects in the control group
\(sd_1 = \) standard deviation of the treatment group
For the purpose of this study, a pooled standard deviation was chosen rather than the control group standard deviation because it has less sampling error than the control group standard deviation under the condition of equal sample size (Hunter and Schmidt, 1990).

When means and standard deviation were not reported, the effect sizes were derived from F-value, t-value, or Chi-squared statistics. For such cases, these values were converted to (d) statistics by using the conversion formulas provided by Defife (2009).

\[ (t \text{ test conversion}) \quad d = \frac{t(n_1 + n_2)}{\sqrt{n_1 + n_2 - 2\sqrt{n_1n_2}}} \]

\[ (F \text{ test conversion}) \quad d = \frac{\sqrt{F(n_1 + n_2)}}{\sqrt{df} \sqrt{n_1n_2}} \]

\[ (\chi^2 \text{ test conversion}) \quad d = 2\sqrt{\frac{x^2}{n - x^2}} \]

where,

\( n_1 \) = number of subjects in the treatment group

\( n_2 \) = number of subjects in the control group

In contrast, the effect size for those studies based upon a within-subject design was calculated by using the following formula derived by Becker (1988)
\[ ES(d) = \frac{M_1 - M_2}{Sd} \]

where,

\[ ES = \text{effect size} \]

\[ M_1 = \text{the post-test mean score} \]

\[ M_2 = \text{the pretest mean score} \]

\[ Sd = \text{the mean standard deviation of the pre-post test} \]

The formula for the corresponding mean standard deviation is:

\[ Sd = \sqrt{\frac{Var_1 + Var_2}{2}} \]

where,

\[ Var_1 = \text{the post-test variance score} \]

\[ Var_2 = \text{the pretest variance score} \]

When means and standard deviation were not reported, the effect sizes were derived from the relevant F-value or t-value. For such cases, these values were converted to (d) statistics by using the conversion formulas provided by Rosenthal (1991).

\[
(t \text{ test conversion}) \ d = \frac{2t}{\sqrt{df}}
\]

\[
(F \text{ test conversion}) \ d = 2 \sqrt{\frac{F}{df}}
\]
Bias Adjustment to Effect size

Effect sizes were weighted according to sample size. This was done to prevent effect sizes based upon greater sample sizes from being given more weight in the final analysis than effect sizes based upon smaller sample sizes. Hedges and Olkin (1985) suggested a formula for correcting this effect-size bias for both between-subject design and within subject design.

\[
\text{Unbiased } (d) = d \left[ 1 - \left( \frac{3}{4(n_1 + n_2) - 9} \right) \right]
\]

where,

\( n_1 \) = number of subjects in the first group

\( n_2 \) = number of subjects in the second group

\( d \) = calculated value for effect size

Calculation of Mean Effect Sizes

Effect sizes are combined to obtain an overall effect size estimate. According to Cooper (2010), the first step in calculating the \( d \) index is to “calculate a weighting factor, which is the inverse of the variance associated with each \( d \) index estimate” (p. 137)

\[
\text{wi(WeightedFactor)} = \frac{2(n_1 + n_2)n_1n_2}{2(n_1 + n_2)^2 + n_1n_2d^2}
\]

Where,

\( n_1 \) = number in the first group

69
\[ n_2 = \text{number in the second group} \]
\[ d = \text{calculated value for effect size} \]

The formula for combining effect sizes to obtain an overall effect size estimate is provided below.

\[ d_* = \frac{\sum_{i=1}^{k} d_i w_i}{\sum_{i=1}^{k} w_i} \]

Where,
\[ d_i = \text{calculated value for effect size} \]
\[ w_i = \text{weighted factor} \]

**Test for Homogeneity of the Effect Sizes**

To investigate whether the effect sizes were homogeneous, a homogeneity of the effect sizes is calculated using the \( Q \) statistic by Hedges and Olkin (1985). The homogeneity statistic is distributed as a chi-square variable with degrees of freedom equal to effect size used minus (k-1). If the computed value is greater than the critical value of chi-square, then this suggests heterogeneity in which the effect sizes differ significantly more than that expected by sampling error. In the presence of heterogeneity, the moderator variables were introduced to attempt to account for the heterogeneity.

To investigate the relation between the moderators and the magnitude of the effect sizes, categorical models were tested (Eagly & Jahnson, 1990; Hedges & Olkin, 1985; Johnson, 1989). The method of calculating categorical models provided a between-class
homogeneity ($Q_b$) and within-class homogeneity ($Q_w$). $Q_b$ tests for homogeneity of effect sizes across classes. It has an approximate chi-square distribution with $p - 1$ degrees of freedom, where $p$ is the number of classes. $Q_w$ indicates whether the effect sizes within each class are homogeneous. It has an approximate chi-square distribution with $m - 1$ degrees of freedom, where $m$ is the number of effect sizes in each class.

The formula for the test of non-homogeneity within groups ($Q_t$) is provided below.

$$Q_t = \sum_{i=1}^{k} w_i * d^2 - \frac{\left(\sum_{i=1}^{k} w_i d_i\right)^2}{\sum_{i=1}^{k} w_i}$$

Where,

$d$ = calculated value for effect size

$w_i$ = weighted factor

The formula to determine whether between group comparison (Homogeneity Analysis Between Group) explains the variance in effect sizes is provided below:

$$Q_b = Q_t - Q_w$$

Where,

$Q_w$ = homogeneity factor within the group

$Q_t$ = sum of all $Q$ values
Computational Software

Microsoft Excel was used initially for calculating these analyses. Then the SPSS statistical package and Comprehensive Meta-Analysis (Biostat) was used for verifying the accuracy of the data obtained. We conducted analyses using Comprehensive Meta-Analysis (Biostat) and SPSS (Version 11 for Macintosh OSX).
CHAPTER V

RESULTS

The purpose of this study was to investigate the effectiveness of STEM education programs. This study found that STEM programs have positive effects on the overall effect size of the three measurement outcomes: engagement, capability, and continuity. As mentioned in Chapter 3, the concept of “engagement” is defined as students’ and/or teachers’ interest and initial involvement in STEM-related disciplines. The concept of “capability” is defined as students’ and/or teachers’ knowledge and skills improvement. Finally, “continuity” is defined as the opportunity for students and/or teachers to move ahead to the next level of education and/or work. Although the overall effect sizes of these three outcome variables were positive, they remained heterogeneous, which means each of outcome variables had a high level of variance in effect sizes. Whether there is significant difference of the effect sizes resulting from the moderator variables remains an empirical question. One of the main purposes of this study is to determine whether moderator variables might explain the variation in STEM programs’ effect sizes.
5.1 The Three Overall Effectiveness Measurements

The 91 studies included in this study produced 222 effect sizes for three measurements of STEM education programs: engagement, capability, and continuity. A total of 91 studies were identified as suitable for this analysis, but some studies reported more than one measurement of dependent variables. There were 32 of the 91 studies that reported multiple outcomes: 27 of these studies reported two outcome variables; either engagement and capability, engagement and continuity, or capability and continuity, and 5 studies reported all three outcomes. Overall, 31 studies reported engagement outcomes, 77 studies reported capability, and 17 studies reported continuity.

The publication dates of these studies ranged from 1980 to 2010, with 54 of the studies published between 2000 and 2010, 26 of the studies published between 1990 and 1999, and 11 of the studies published between 1980 and 1989. Eighty-seven of the 91 studies retrieved were from peer-reviewed journals. Only four of the 91 studies were conference papers. Section 5.1 presents the overall effect sizes of the three outcome measurements with the weighted mean effect sizes resulting from the 91 studies. Also, this section provides heterogeneity test results and estimates of the effect of moderator variables on the three outcome measurements.

5.1.1 Weighted Mean Effect Sizes for Three Effectiveness Measurements

Table 5-1 shows the number of effect sizes, weighted average effect sizes, ranges of effect sizes (with 95% confidence intervals) for three different outcomes in STEM
education programs. There were 44 effect sizes on engagement, 153 effect sizes on capability and 25 effect sizes on continuity.

Table 5 - 1 The Effect Sizes and Homogeneity Test Results of Three Outcomes of STEM Education Program

<table>
<thead>
<tr>
<th>Outcome Measure</th>
<th># of Studies Evaluating the Outcome Measure</th>
<th># of Effect Sizes</th>
<th>95% C. I</th>
<th>Mean(d)</th>
<th>Qw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engagement</td>
<td>31</td>
<td>44</td>
<td>0.306 to 0.386</td>
<td>.346</td>
<td>301.93***</td>
</tr>
<tr>
<td>Capability</td>
<td>77</td>
<td>153</td>
<td>0.431 to 0.476</td>
<td>.454</td>
<td>1922.05***</td>
</tr>
<tr>
<td>Continuity</td>
<td>17</td>
<td>25</td>
<td>0.321 to 0.418</td>
<td>.369</td>
<td>125.09***</td>
</tr>
</tbody>
</table>

Note1: *** p < .05

The effect size interpretation was based on Cohen (1988) where d = .20 is considered to be a small effect within the behavioral sciences, d = .50, is a moderate effect, and d = .80, is a large effect when the effect sizes are statistically significant. Cohen (1988) also pointed out that the relatively small effect sizes around d = .20 were most representative of fields that are closely aligned with social science such as in education, personality, social psychology, and clinical psychology.

The 95% confidence interval for the effect sizes for engagement in STEM programs ranged from 0.306 to 0.386 with an average weighted effect size of .346, a small, positive effect. The effect sizes for capability in STEM programs ranged from 0.431 to 0.476 with an average weighted effect size of .454, a moderate, positive effect. The effect sizes for continuity in STEM programs ranged from 0.321 to 0.418 with an average weighted effect size of .369, a moderate, positive effect. The 95% confidence intervals of these average effect sizes did not include “0”, which means the average effect sizes were statistically significant.
To determine whether the “observed variance in the individual effect sizes was significantly different from that expected by sampling error” (Cooper, 2008), a test for non-homogeneity of the effect sizes was conducted. As Table 5-1 shows, a test for non-homogeneity of each of the outcome level’s groupings revealed $Q_w$ values that exceeded the critical value of the homogeneity, which means these results of each three outcome variables were heterogeneous.

The homogeneity analysis for the engagement outcomes resulted in a $Q$ value of 301.93 with 43 degrees of freedom and $p < .05$. The homogeneity analysis for the capability outcomes resulted in a $Q$ value of 1922.05 with 152 degrees of freedom and $p < .05$. The homogeneity analysis for the continuity outcomes resulted in a $Q$ value of 125.09 with 24 degrees of freedom, and $p < .05$. Sampling error concerning each of the three outcomes cannot explain the differences in effect sizes, while further analysis of potential moderators from study characteristics might explain the heterogeneity in effect sizes.

Since the effect sizes were heterogeneous in each of the three outcome variables, I identified moderator variables to explain the heterogeneity. As previously described, this study explored five categories of moderators: STEM program variations, participant characteristics, setting/environment characteristics, study characteristics of STEM programs, and the method used in the various studies. All moderator variables are categorical so categorical models were tested to examine whether the moderator variables could explain the heterogeneity of the effect sizes (Eagly & Johnson, 1990; Hedges & Olkin, 1985; Johnson, 1989).
5.2 Moderator Variables for the Engagement Outcomes

The 31 studies produced 44 effect sizes for the engagement of students and teachers in STEM education programs. The effect sizes were between 0.306 and 0.386, with a weighted average effect size of 0.346. The weighted average effect size indicates that STEM programs were moderately successful in engaging students and teachers. Due to the fact that this moderate positive effect size is significantly heterogeneous, a number of moderator variables are identified to be examined whether they can account for the heterogeneity of the effect sizes. As shown in Table 5.2, the heterogeneity or variance of engagement outcome was not fully explained by the following moderator variables: research design, reliability evidence, program strategy, program creator, program funder, grade level, and regional location. A successful moderator is supposed to produce homogeneous effect sizes within the same level of the variable and heterogeneous effect sizes across different levels of the variable. In other words, the $Q_w$ in Table 5.2 is not supposed to be significant and the $Q_b$ in Table 5.2 is supposed to be significant. The results in Table 5.2 showed none of the moderator variables successfully accounted for the heterogeneity of the effect sizes.
Table 5 - Effect of Moderator Variables on Weighted Mean Effect Sizes in the Engagement of Students and Teachers

<table>
<thead>
<tr>
<th>Moderator Variable</th>
<th>Mean(d)</th>
<th>(n)</th>
<th>Q&lt;sub&gt;w&lt;/sub&gt;</th>
<th>Q&lt;sub&gt;b&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study Method</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research Design</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Subject Design</td>
<td>.213</td>
<td>16</td>
<td>103.13***</td>
<td>67.00**</td>
</tr>
<tr>
<td>Within Subject Design</td>
<td>.557</td>
<td>28</td>
<td>131.38***</td>
<td></td>
</tr>
<tr>
<td>Reliability Evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>More than .7</td>
<td>.304</td>
<td>14</td>
<td>22.07</td>
<td>0.19</td>
</tr>
<tr>
<td>Less than .7</td>
<td>.341</td>
<td>3</td>
<td>5.87(H)</td>
<td></td>
</tr>
<tr>
<td>No Information</td>
<td>.349</td>
<td>27</td>
<td>273.81***</td>
<td></td>
</tr>
<tr>
<td><strong>Program Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STEM Strategy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inquiry</td>
<td>.335</td>
<td>11</td>
<td>20.51**</td>
<td>3.24</td>
</tr>
<tr>
<td>Enhanced Context</td>
<td>.357</td>
<td>23</td>
<td>239.09***</td>
<td></td>
</tr>
<tr>
<td>Instructional Technology</td>
<td>.394</td>
<td>2</td>
<td>7.31***</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>.340</td>
<td>7</td>
<td>32.97***</td>
<td></td>
</tr>
<tr>
<td>Funder</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Government</td>
<td>.185</td>
<td>23</td>
<td>104.43***</td>
<td>88.84***</td>
</tr>
<tr>
<td>Nonprofit Organization</td>
<td>.341</td>
<td>3</td>
<td>0.56(H)</td>
<td></td>
</tr>
<tr>
<td>(private and public school)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Information</td>
<td>.318</td>
<td>13</td>
<td>18.80(H)</td>
<td></td>
</tr>
<tr>
<td>Program Creator</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professional Association</td>
<td>.680</td>
<td>11</td>
<td>86.09***</td>
<td>125.68***</td>
</tr>
<tr>
<td>Government</td>
<td>.155</td>
<td>7</td>
<td>22.54***</td>
<td></td>
</tr>
<tr>
<td>Non-profit Organization</td>
<td>.166</td>
<td>16</td>
<td>50.89***</td>
<td></td>
</tr>
<tr>
<td>(private and public school)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>.278</td>
<td>6</td>
<td>12.18**</td>
<td></td>
</tr>
<tr>
<td>No Information</td>
<td>.360</td>
<td>4</td>
<td>4.54(H)</td>
<td></td>
</tr>
<tr>
<td><strong>Participants Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample Grade Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elementary School</td>
<td>.353</td>
<td>5</td>
<td>29.96***</td>
<td>78.43***</td>
</tr>
<tr>
<td>Middle School</td>
<td>.223</td>
<td>2</td>
<td>1.68(H)</td>
<td></td>
</tr>
<tr>
<td>High School</td>
<td>.103</td>
<td>4</td>
<td>22.27***</td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>.187</td>
<td>7</td>
<td>3.05(H)</td>
<td></td>
</tr>
<tr>
<td>K-12 Students</td>
<td>.574</td>
<td>8</td>
<td>71.74***</td>
<td></td>
</tr>
<tr>
<td>K-12 Teacher</td>
<td>.216</td>
<td>18</td>
<td>94.80***</td>
<td></td>
</tr>
<tr>
<td><strong>Setting Environment Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regional Location</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Midwest</td>
<td>.154</td>
<td>12</td>
<td>19.52(H)</td>
<td>129.25***</td>
</tr>
<tr>
<td>Northeast</td>
<td>.112</td>
<td>6</td>
<td>30.42***</td>
<td></td>
</tr>
<tr>
<td>South</td>
<td>.643</td>
<td>16</td>
<td>91.80***</td>
<td></td>
</tr>
<tr>
<td>Nationwide</td>
<td>.405</td>
<td>6</td>
<td>30.10***</td>
<td></td>
</tr>
<tr>
<td>No Information</td>
<td>.457</td>
<td>3</td>
<td>0.82(H)</td>
<td></td>
</tr>
</tbody>
</table>
### Study Characteristics

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.170</td>
<td>.453</td>
<td>.314</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>17</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>43.56***</td>
<td>184.35***</td>
<td>35.12***</td>
<td></td>
</tr>
</tbody>
</table>

Note: *** refers to weighted mean effect sizes that are heterogeneous at .001 level
** refers to weighted mean effect sizes that are heterogeneous at .05 level
H refers to weighted mean effect sizes that are homogeneous

### Research Designs

Some previous studies used between-subject design by contrasting experimental group versus control group to demonstrate the effect of STEM programs while others used within-subject design by comparing the same group of people before and after STEM programs. Is there a systematic difference in the engagement outcome of STEM programs when using a between-subject research design versus a within-subject design to make the comparison? Both research designs have their own advantages and disadvantages. A within-subject design is advantageous in controlling for individual differences among research participants as well as in requiring fewer participants in a study. However, this research design is exposed to threats to internal validity, such as history, maturation, and pretest effects. A between-subject design avoids the time-related threats, but it is difficult to conduct a random selection and a random assignment of subjects to each group due to practical constraints. In addition, different formulas are needed to calculate the effect size for between-subject design studies and within-subject design studies.

The effect sizes for studies using between-subject design versus studies using within-subject design on the engagement outcome is reported in Table 5-3. When
considered for their research designs, the 31 studies produced 44 effect sizes in engagement outcomes. There were 16 effect sizes produced by between-subject design and 28 effect sizes by within-subject design.

When the within-subject design was compared to the between-subject design, the between-subject design had a higher average effect sizes than the within-subject design. The effect sizes from the between-subject design ranged from 0.264 to 0.161 in the engagement outcomes. The average weighted effect size was 0.213, a small, positive effect. The effect size from the within-subject design ranged from 0.622 to 0.492 in the engagement outcomes. The average weighted effect size was 0.557, a moderate, positive effect. The 95% confidence intervals for these average effect sizes did not include “0”, which means the average effect sizes were statistically significant.

Table 5 - 3 The Effect Sizes of STEM Programs on the Engagement by Research Design

<table>
<thead>
<tr>
<th>Research Design</th>
<th># of Study</th>
<th># of Effect Size</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between-Subject Design</td>
<td>12</td>
<td>16</td>
<td>0.264 to 0.161</td>
<td>.213</td>
<td>103.13***</td>
<td>67.00**</td>
</tr>
<tr>
<td>Within-Subject Design</td>
<td>19</td>
<td>28</td>
<td>0.622 to 0.492</td>
<td>.557</td>
<td>131.38***</td>
<td></td>
</tr>
</tbody>
</table>

Note: Qw refers to the non-homogeneity within each group

The homogeneity test for the two types of research designs revealed Qb of 67.00 and p < .05. The effectiveness of STEM programs on engagement outcomes was heterogeneous across different research designs. A test for homogeneity within each of the research designs revealed Qw values that exceeded the 95th percentile of the chi-square distribution. This result was heterogeneous, which means that there was a high level of variance within each of the research designs in the engagement outcomes.
STEM Education Strategies

Is there a difference in the engagement outcomes of STEM programs when considering STEM education strategies? To answer this question, this study grouped pedagogical strategy into four categories: Inquiry, Enhanced Context, Instructional Technology, and Mixed Strategies. Table 5-4 presents the distribution of effect sizes on the engagement outcome by program strategy. There were eight studies of STEM programs applying an inquiry strategy yielding 11 effect sizes. The most widely reported strategy used in STEM programs was an enhanced context strategy, with 23 effect sizes from 15 studies. There were two studies applying an instruction technology strategy yielding two effect sizes. Five studies that applied a mixed strategy yielded seven effect sizes.

The effect sizes from the inquiry strategy ranged from 0.185 to 0.485 with an average weighted effect size of 0.335, a small, positive effect. The enhanced context strategy had effect sizes ranging from 0.308 to 0.406 with an average weighted effect sizes of 0.357, a small, positive effect. The effect size for the instructional technology strategy ranged from -0.042 to 0.830 with an average weighted effect size of 0.394. Finally, the effect sizes of the mixed strategies ranged from 0.238 to 0.442 with an average weighted effect size of 0.340, a small, positive effect. The majority of the 95% confidence intervals for these average effect sizes did not include “0”, which means the average effect sizes were statistically significant.
Table 5 - 4 The Effect Sizes of STEM Programs on the Engagement by Program Strategy

<table>
<thead>
<tr>
<th>STEM strategies</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inquiry</td>
<td>8</td>
<td>11</td>
<td>0.185 to 0.485</td>
<td>.335</td>
<td>20.32**</td>
<td>3.24</td>
</tr>
<tr>
<td>Enhanced Context</td>
<td>15</td>
<td>23</td>
<td>0.308 to 0.406</td>
<td>.357</td>
<td>239.09***</td>
<td>3.49</td>
</tr>
<tr>
<td>Instructional Technology</td>
<td>2</td>
<td>2</td>
<td>-0.042 to 0.830</td>
<td>.394</td>
<td>7.31***</td>
<td>3.08</td>
</tr>
<tr>
<td>Mixed</td>
<td>5</td>
<td>7</td>
<td>0.238 to 0.442</td>
<td>.340</td>
<td>32.97***</td>
<td>3.08</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>43</td>
<td>-0.042 to 0.830</td>
<td>.346</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The STEM education strategy was not a successful moderator because it did not create a homogeneous within group variance and a heterogeneous between group variance. The homogeneity test for STEM education strategies revealed $Q_b$ of 3.24 and $p > .05$. This suggests that the effectiveness of STEM programs on engagement outcomes was homogeneous across different pedagogical strategies. Also, a homogeneity test within each of the strategies revealed $Q_w$ values that exceeded the 95th percentile of the chi-square distribution, which means there was a high level of variance in the engagement levels within each strategy.

**Funding Source**

Is there a difference in the capability outcomes of STEM programs when considering different funding sources? To answer this question, this study conducted comparisons across different funding sources. This meta-analysis identified three funding sources of STEM programs: government, nonprofit organization (public and private school), and no information. Table 5-5 shows the distribution of effect sizes by funding sources. Of the 27 studies that evaluated engagement, the 14 studies of programs that
were funded by all levels of government produced 23 effect sizes. There were two studies of programs funded by non-profit organizations produced two effect sizes. The 11 studies that produced 13 effect sizes did not provide any information about STEM-program funding source.

The 23 effect sizes from government-funded STEM programs ranged from 0.130 to 0.239 with an average weighted effect size of 0.185, a small, positive effect. The three effect sizes from the STEM programs funded by nonprofit organizations ranged from 0.075 to 0.608 with an average weighted effect size of 0.341, a small, positive effect. Finally, the 13 effect sizes from STEM programs with no information ranged from 0.226 to 0.410 with an average weighted effect size of .318, a small, positive effect. The 95% confidence intervals for these average effect sizes did not included “0”, which means the average effect sizes were statistically significant.

Table 5 - 5 The Effect Sizes of STEM Programs on the Engagement by Founding Source

<table>
<thead>
<tr>
<th>STEM Areas</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government</td>
<td>14</td>
<td>23</td>
<td>0.130 to 0.239</td>
<td>.185</td>
<td>104.43***</td>
<td>88.84***</td>
</tr>
<tr>
<td>Nonprofit Organization</td>
<td>2</td>
<td>2</td>
<td>0.075 to 0.608</td>
<td>.341</td>
<td>0.56(H)</td>
<td></td>
</tr>
<tr>
<td>No Information</td>
<td>11</td>
<td>13</td>
<td>0.226 to 0.410</td>
<td>.318</td>
<td>18.80(H)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>27</td>
<td>38</td>
<td>0.075 to 0.608</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The homogeneity test for program funding source revealed Qb of 88.84 and p < .05. It showed heterogeneity across different program funding sources. A homogeneity test within government-funded programs revealed Qw values that exceeded the 95th percentile of the chi-square distribution, which means there was a high level of variance in the engagement levels within government-funded programs. However, a homogeneity
test within nonprofit organization revealed Q× values that did not exceed the 95th percentile of the chi-square distribution. The mean effect size .341 is a typical representation of engagement outcome of STEM programs within nonprofit organization funding source.

**STEM Program Creator**

Is there a difference in the engagement outcomes of STEM programs when considering the program creators? To answer this question, this meta-analysis compared five different program creators: governments (federal, state, and local government), non-profit organizations (private and public schools), professional associations (e.g. the National Teacher Association), mixed organizations, and no information. Table 5-6 shows the distribution of effect sizes by program creator. Of the 31 studies that focused on the engagement outcomes, the five studies were created by governmental entities yielded seven effect sizes. There were 12 studies created by non-profit entities that produced 16 effect sizes. There were six studies created by professional associations that yielded 11 effect sizes. There were five studies created by multiple creators (i.e. programs created jointly by governmental and non-profit entities) that yielded six effect sizes. There were three studies with no information about the STEM program creators that generated four effect sizes.

Programs that were created by governments and non-profit organizations exhibited small positive effect sizes. The seven effect sizes from programs created by governmental entities ranged from 0.080 to 0.230 with an average weighed effect size of 0.155, a small, positive effect. The 16 effect sizes from programs created by non-profit
entities ranged from 0.081 to 0.250 with an average weighed effect size of 0.166, a small, positive effect. On the other hand, STEM programs created by professional associations had higher effect sizes than those created by others. The 11 effect sizes from programs created by professional associations ranged from 0.607 to 0.752 with an average weighted effect size of 0.680, a large, positive effect. The six effect sizes from programs created by multiple entities ranged from 0.161 to 0.395 with an average weighed effect size of 0.278, a small positive effect. The four effect sizes from programs with no creator information ranged from 0.194 to 0.526 with an average weighed effect size of 0.360, a small, positive effect. The 95% C.I. for all average effect sizes did not include 0, which means these average effect sizes were statistically significant.

Table 5 - 6 The Effect Sizes of STEM Programs on the Engagement by Program Creator

<table>
<thead>
<tr>
<th>Program Creator</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government</td>
<td>5</td>
<td>7</td>
<td>0.080 to 0.230</td>
<td>0.155</td>
<td>22.54***</td>
<td>125.68***</td>
</tr>
<tr>
<td>Non-profit Organization</td>
<td>12</td>
<td>16</td>
<td>0.081 to 0.250</td>
<td>0.166</td>
<td>50.89***</td>
<td></td>
</tr>
<tr>
<td>Professional Association</td>
<td>6</td>
<td>11</td>
<td>0.607 to 0.752</td>
<td>0.680</td>
<td>86.09***</td>
<td></td>
</tr>
<tr>
<td>Mixed (cooperative)</td>
<td>5</td>
<td>6</td>
<td>0.161 to 0.395</td>
<td>0.278</td>
<td>12.18**</td>
<td></td>
</tr>
<tr>
<td>No Information</td>
<td>3</td>
<td>4</td>
<td>0.194 to 0.526</td>
<td>0.360</td>
<td>4.54(H)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>44</td>
<td>0.080 to 0.752</td>
<td>0.346</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Professional associations include the National Teacher Association and educators (teacher). Government includes federal, state, and local government. Nonprofit organization includes private and public schools. H refers to weighted mean effect sizes that are homogeneous

The program creator was not a successful moderator because it did not create a homogeneous within group variance and a heterogeneous between group variance. It showed a heterogeneous engagement outcome both within and between different program
creators. The homogeneity test of the different types of STEM program creators revealed $Q_b$ of 125.68 and $p < .001$. There was a significant difference among the four creator types. The homogeneity test for each program creator revealed $Q_w$ values that exceeded the 95th percentile of the chi-square distribution, with the exception of “no information” classification.

**Regional Location**

Is there a difference in the engagement outcomes of STEM programs when considering the regions in which STEM programs were conducted? To answer this question, this meta-analysis classified all STEM programs into four regions: South, Midwest, Northeast, and nationwide. The West was removed due to one observation. Table 5-7 shows the distribution of effect sizes by regional location. In the South, there were 11 studies producing 16 effect sizes. In the Midwest, eight studies produced 12 effect sizes. In the Northeast, five studies produced six effect sizes. Across the nation, four studies produced six effect sizes. Finally, two studies with three effect sizes did not provide regional information.

The Southern region produced the highest effect sizes of the four American regions. The 16 effect sizes from STEM programs implemented in the South ranged from 0.575 to 0.712 with an average weighted effect size of 0.643, a large, positive effect. In contrast, STEM programs carried out in the Midwest had small, positive average effect sizes. The 12 effect sizes from STEM programs implemented in the Midwest ranged from 0.092 to 0.216 with an average weighted effect size of 0.154, a small, positive effect. The six effect sizes from STEM programs implemented in the Northeast ranged from -0.017 to 0.241 with an average weighted effect size of 0.112. The six effect sizes from STEM
programs implemented nationwide ranged from 0.265 to 0.544 with an average weighted effect size of 0.405, a moderate, positive effect. The majority of the 95% confidence interval did not include 0, which means these average effect sizes were statistically significant.

Table 5 - 7 The Effect Sizes of STEM Programs on the Engagement by Regional Location

<table>
<thead>
<tr>
<th>Regional Location</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>South</td>
<td>11</td>
<td>16</td>
<td>0.575 to 0.712</td>
<td>.643</td>
<td>91.80***</td>
<td>129.25***</td>
</tr>
<tr>
<td>Midwest</td>
<td>8</td>
<td>12</td>
<td>0.092 to 0.216</td>
<td>.154</td>
<td>19.52(H)</td>
<td></td>
</tr>
<tr>
<td>Northeast</td>
<td>5</td>
<td>6</td>
<td>-0.017 to 0.241</td>
<td>.112</td>
<td>30.42***</td>
<td></td>
</tr>
<tr>
<td>Nationwide</td>
<td>4</td>
<td>6</td>
<td>0.265 to 0.544</td>
<td>.405</td>
<td>30.10***</td>
<td></td>
</tr>
<tr>
<td>No information</td>
<td>2</td>
<td>3</td>
<td>0.220 to 0.695</td>
<td>.457</td>
<td>0.84(H)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>43</td>
<td>-0.017 to 0.712</td>
<td>.344</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Sample size (n) is 43 because one observation from the West was removed. H refers to weighted mean effect sizes that are homogeneous.

The homogeneity test of regional locations revealed Qb of 129.25 and p < .001. There was a significant difference among the five regional classifications. The test for homogeneity within each of the regions revealed Qw values that exceeded the 95th percentile of the chi-square distribution, with the exception of the “Midwest” and “no information” classification. The mean effect size .154 is a typical representation of engagement outcome of STEM program within Midwest region.

Educational Levels of Sample Group

Is there a difference in the engagement outcomes of STEM programs when considering students’ education levels in STEM programs? To answer this question, this...
study classified six sample groups: elementary school, middle school, high school, college, K-12 students, and K-12 teachers. Table 5-10 shows the distribution of effect sizes by educational levels. Of the 31 studies, three studies measuring elementary students yielded five effect sizes, two studies sampling middle students yielded two effect sizes, four studies sampling high school students produced four effect sizes, four studies sampling college students yielded seven effect sizes, five studies measuring K-12 students yielded eight effect sizes and 14 studies measuring K-12 teachers’ engagement yielded 18 effect sizes.

Elementary, middle, and K-12 students had higher average weighted effect sizes than high and college students. Elementary students had effect sizes ranging from 0.235 to 0.471, with an average weighted effect size of 0.353, a small, positive effect. Middle school students had effect sizes ranging from -0.068 to 0.515, with an average weighted effect size of .223. On the other hand, high school and college students tended to have lower average weighted effect sizes. High school students had effect sizes ranging from -0.013 to 0.220, with an average weighted effect size of 0.103. College students had effect sizes ranging from 0.016 to 0.358, with an average weighted effect size of 0.187, a small, positive effect. K-12 students had effect sizes ranging from 0.507 to 0.641, with an average weighed effect size of 0.574, a moderate, positive effect. The effect sizes of K-12 teachers ranged from 0.145 to 0.286, with an average weighted effect size of 0.216, a small, positive effect. The majority of the 95% confidence intervals did not include 0, which means these average effect sizes were statistically significant.
Table 5 - 8 The Effect Sizes of STEM Programs on the Engagement by Educational Level

<table>
<thead>
<tr>
<th>Educational Level</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary Students</td>
<td>3</td>
<td>5</td>
<td>0.235 to 0.471</td>
<td>.353</td>
<td>29.96***</td>
<td>78.43**</td>
</tr>
<tr>
<td>Middle School Students</td>
<td>2</td>
<td>2</td>
<td>-0.068 to 0.515</td>
<td>.223</td>
<td>1.68(H)</td>
<td></td>
</tr>
<tr>
<td>High School Students</td>
<td>4</td>
<td>4</td>
<td>-0.013 to 0.220</td>
<td>.103</td>
<td>22.27***</td>
<td></td>
</tr>
<tr>
<td>College Students</td>
<td>4</td>
<td>7</td>
<td>0.016 to 0.358</td>
<td>.187</td>
<td>3.05(H)</td>
<td></td>
</tr>
<tr>
<td>K-12 Students</td>
<td>5</td>
<td>8</td>
<td>0.507 to 0.641</td>
<td>.574</td>
<td>71.74***</td>
<td></td>
</tr>
<tr>
<td>K-12 Teachers</td>
<td>13</td>
<td>18</td>
<td>0.145 to 0.286</td>
<td>.216</td>
<td>94.80***</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>44</td>
<td>-0.068 to 0.641</td>
<td>.346</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: H refers to weighted mean effect sizes that are homogeneous

The homogeneity test of education levels of sample groups revealed Qb of 78.43 and p < .05. This suggests that the effect sizes of STEM programs on engagement outcome were heterogeneous among the six educational levels. A homogeneity test within each of the educational levels revealed Qw values that exceeded the 95th percentile of the chi-square distribution in elementary, high school students, K-12 student, and K-12 teacher. However, a homogeneity test within middle and college students showed homogeneous results. The mean effect size .223 is a typical representation of engagement outcome of STEM programs among middle school students. The mean effect size .187 is a typical representation among college school students.

**Publication Year**

Is there a difference in engagement outcomes of STEM programs when considering the publication year? To answer this question, this meta-analysis compared effect sizes by publication year. Table 5-9 shows the distribution of effect sizes by
publication year. More than half of the studies that analyzed the engagement outcome in this study were published between 2000 and 2010. The 16 studies in that period produced 18 effect sizes. Nine studies published between 1990 and 1999 produced 17 effect sizes. Six studies published between 1980 and 1989 had nine effect sizes.

The 95% confidence interval of effect sizes from the studies published from 2000 to 2010 ranged from 0.100 to 0.239 with an average weighted effect size of 0.170, a small, positive effect. In contrast, the 95% confidence interval of effect sizes from 1990 to 1999 ranged from 0.397 to 0.510 with an average weighted effect size of 0.453, a moderate, positive effect. The 95% confidence interval of effect sizes for the studies published from 1980 to 1989 ranged from 0.274 to 0.474 with an average weighted effect size of 0.314, a small, positive effect. The 95% confidence intervals for these average effect sizes did not include “0”, which means the average effect sizes were statistically significant.

Table 5 - 9 The Effect Sizes of STEM Programs on the Engagement by Publication Year

<table>
<thead>
<tr>
<th>Year</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Q_w</th>
<th>Q_b</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000-2010</td>
<td>16</td>
<td>18</td>
<td>0.100 to 0.239</td>
<td>.170</td>
<td>43.56***</td>
<td>38.90</td>
</tr>
<tr>
<td>1990-1999</td>
<td>9</td>
<td>17</td>
<td>0.397 to 0.510</td>
<td>.453</td>
<td>184.35***</td>
<td></td>
</tr>
<tr>
<td>1980-1999</td>
<td>6</td>
<td>9</td>
<td>0.274 to 0.474</td>
<td>.314</td>
<td>35.12***</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>44</td>
<td>0.100 to 0.510</td>
<td>.346</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The publication year was not a successful moderator because it did not create a homogeneous within group variance and a heterogeneous between group variance. The homogeneity test of the publication year of individual studies revealed Q_b of 38.90 and p > .05. There was no significant difference among the publication year. The homogeneity
test within each of the published years revealed $Q_w$ values that exceeded the 95\textsuperscript{th} percentile of the chi-square distribution. The results mean that there was a high level of variations within the published year groups.

**Reliability Evidence for Instrument**

Is there a difference in the engagement outcomes of STEM programs when considering whether or not the studies provided reliability evidence in the study? To answer this question, this meta-analysis classified three instrument reliability levels: “more than .7 reliability”, “less than .7 reliability”, and “no information.” The studies that did not provide the reliability level of the measurement instrument were identified as “no information.” Table 5-10 presents the distribution of effect sizes on the engagement outcome by reliability evidence. Of the 31 studies on engagement, the 11 studies indicating “more than .7 reliability” produced 14 effect sizes. Only two studies explicitly stated that the reliability was less than .7 and yielded three effect sizes. The 18 classified as “no information” produced 27 effect sizes.

The effect sizes from studies with “more than .7 reliability” ranged from 0.090 to 0.517 with an average weighted effect size of 0.304, a small, positive effect. Alternatively, the effect sizes from studies with “less than .7 reliability” ranged from 0.261 to 0.420 with an average weighted effect size of .341, a small, positive effect. The effect sizes from the “no information” classification of the measurement instrument ranged from 0.302 to 0.397, with an average weighted effect size of 0.349, a small, positive effect. The 95% confidence intervals for these average effect sizes did not included “0”, which means the average effect sizes were statistically significant.
Table 5 - 10 The Effect Sizes of STEM Programs on the Engagement by Reliability Evidence

<table>
<thead>
<tr>
<th>Reliability</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>More than .7</td>
<td>11</td>
<td>14</td>
<td>0.090 to 0.517</td>
<td>.304</td>
<td>22.07(H)</td>
<td>0.19</td>
</tr>
<tr>
<td>Less than .7</td>
<td>2</td>
<td>3</td>
<td>0.261 to 0.420</td>
<td>.341</td>
<td>5.87(H)</td>
<td></td>
</tr>
<tr>
<td>No Information</td>
<td>18</td>
<td>27</td>
<td>0.302 to 0.397</td>
<td>.349</td>
<td>273.81***</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>44</td>
<td>0.090 to 0.517</td>
<td>.346</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The reliability evidence was not a successful moderator because it did not create a homogeneous within group variance and a heterogeneous between group variance. The homogeneity test of the instrument reliability levels revealed $Q_b$ of 0.19 and $p > .05$. This suggests that the effect sizes of STEM programs on engagement outcomes were homogeneous across the instrument reliability levels. Although the within group variance in “no information” was heterogeneous, the within group variance in “more than .7” and “less than .7” were homogeneous.

5.3 Moderator Variables for the Capability Outcome

There were 77 studies which produced 153 effect sizes for the capability outcomes of the STEM programs. The effect sizes ranged from 0.431 to 0.476, with a weighted averaged effect size of 0.454, a moderate, positive effect. This average effect size indicates that STEM education programs were moderately successful in improving the capability of students and teachers in STEM fields. Despite the positive capability outcome, an empirical issue remains as to whether the independent and moderator
variables have effects on the capability outcome. As shown in Table 5-11, the heterogeneity of the variance of capability outcome was not fully explained by the following moderator variables: research design, reliability evidence, program strategy, funder, program creator, program funder, grade level, and regional location. In other words, the heterogeneity of between-class effect sizes in these moderators, including the independent variable, accounted for the overall mean effect size of the capability outcome. A successful moderator is supposed to produce homogeneous effect sizes within the same level of the variable and heterogeneous effect sizes between different levels of the variable. In other words, the Q_w in Table 5.11 is not supposed to be significant and the Q_b in Table 5.13 is supposed to be significant.
Table 5 - 11 Effect of Moderator Variables on Weighted Mean Effect Sizes in the Capability of Students and Teachers

<table>
<thead>
<tr>
<th>Moderator Variable</th>
<th>Mean (d)</th>
<th>(n)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study Method</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research Design</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Subject Design</td>
<td>.287</td>
<td>79</td>
<td>547.19***</td>
<td>263.54***</td>
</tr>
<tr>
<td>Within Subject Design</td>
<td>.662</td>
<td>74</td>
<td>1111.32***</td>
<td></td>
</tr>
<tr>
<td>Reliability Evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Information</td>
<td>.316</td>
<td>84</td>
<td>510.08***</td>
<td>240.66***</td>
</tr>
<tr>
<td>More than .7</td>
<td>.556</td>
<td>61</td>
<td>1110.56***</td>
<td></td>
</tr>
<tr>
<td>Less than .7</td>
<td>1.087</td>
<td>8</td>
<td>51.74***</td>
<td></td>
</tr>
<tr>
<td><strong>Program Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program Strategies</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inquiry</td>
<td>.707</td>
<td>23</td>
<td>171.94***</td>
<td>304.89***</td>
</tr>
<tr>
<td>Enhanced Context</td>
<td>.553</td>
<td>63</td>
<td>826.53***</td>
<td></td>
</tr>
<tr>
<td>Instructional Technology</td>
<td>.274</td>
<td>28</td>
<td>122.98***</td>
<td></td>
</tr>
<tr>
<td>Collaborative Learning</td>
<td>.139</td>
<td>16</td>
<td>48.65***</td>
<td></td>
</tr>
<tr>
<td>Enhanced Materials</td>
<td>.744</td>
<td>5</td>
<td>30.27***</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>.327</td>
<td>16</td>
<td>411.63***</td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>.493</td>
<td>2</td>
<td>5.17**</td>
<td></td>
</tr>
<tr>
<td>Funder</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Government</td>
<td>.582</td>
<td>68</td>
<td>1111.32***</td>
<td>136.06</td>
</tr>
<tr>
<td>Nonprofit Organization</td>
<td>.298</td>
<td>35</td>
<td>516.10***</td>
<td></td>
</tr>
<tr>
<td>Mixed(cooperative)</td>
<td>.405</td>
<td>7</td>
<td>80.51***</td>
<td></td>
</tr>
<tr>
<td>No Information</td>
<td>.416</td>
<td>43</td>
<td>214.12***</td>
<td></td>
</tr>
<tr>
<td>Program Creator</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professional Association</td>
<td>.653</td>
<td>19</td>
<td>101.66***</td>
<td>210.54***</td>
</tr>
<tr>
<td>Government</td>
<td>.164</td>
<td>17</td>
<td>35.61***</td>
<td></td>
</tr>
<tr>
<td>Non-profit Organization</td>
<td>.450</td>
<td>89</td>
<td>1319.04***</td>
<td></td>
</tr>
<tr>
<td>Private Organization</td>
<td>.059</td>
<td>3</td>
<td>50.45***</td>
<td></td>
</tr>
<tr>
<td>Mixed(cooperative)</td>
<td>.585</td>
<td>20</td>
<td>179.54***</td>
<td></td>
</tr>
<tr>
<td>No Information</td>
<td>.217</td>
<td>5</td>
<td>25.21***</td>
<td></td>
</tr>
<tr>
<td><strong>Participant Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elementary School</td>
<td>.889</td>
<td>20</td>
<td>364.89***</td>
<td>634.66***</td>
</tr>
<tr>
<td>Middle School</td>
<td>.880</td>
<td>12</td>
<td>161.26***</td>
<td></td>
</tr>
<tr>
<td>High School</td>
<td>.243</td>
<td>19</td>
<td>77.93***</td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>.414</td>
<td>52</td>
<td>361.99***</td>
<td></td>
</tr>
<tr>
<td>K-12 Students</td>
<td>.124</td>
<td>11</td>
<td>150.68***</td>
<td></td>
</tr>
<tr>
<td>K-12 Teacher</td>
<td>.428</td>
<td>39</td>
<td>170.64***</td>
<td></td>
</tr>
<tr>
<td><strong>Setting Environment Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regional Location</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>West</td>
<td>.663</td>
<td>24</td>
<td>624.96***</td>
<td>298.08***</td>
</tr>
<tr>
<td>Midwest</td>
<td>.261</td>
<td>38</td>
<td>450.17***</td>
<td></td>
</tr>
<tr>
<td>Region</td>
<td>Effect Size</td>
<td>N</td>
<td>Mean Effect Size</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------</td>
<td>-----</td>
<td>------------------</td>
<td></td>
</tr>
<tr>
<td>Northeast</td>
<td>.415</td>
<td>21</td>
<td>62.96***</td>
<td></td>
</tr>
<tr>
<td>South</td>
<td>.555</td>
<td>50</td>
<td>406.67***</td>
<td></td>
</tr>
<tr>
<td>Nationwide</td>
<td>.214</td>
<td>13</td>
<td>69.83***</td>
<td></td>
</tr>
<tr>
<td>No Information</td>
<td>.836</td>
<td>7</td>
<td>9.09(H)</td>
<td></td>
</tr>
</tbody>
</table>

**Study Characteristics**

<table>
<thead>
<tr>
<th>Published Year</th>
<th>Effect Size</th>
<th>N</th>
<th>Mean Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000-2010</td>
<td>.542</td>
<td>85</td>
<td>1229.46***</td>
</tr>
<tr>
<td>1990-1999</td>
<td>.304</td>
<td>55</td>
<td>555.13***</td>
</tr>
<tr>
<td>1980-1989</td>
<td>.419</td>
<td>13</td>
<td>41.89***</td>
</tr>
</tbody>
</table>

Note: *** refers to weighted mean effect sizes that are heterogeneous at .001 level
** refers to weighted mean effect sizes that are heterogeneous at .05 level
H refers to weighted mean effect sizes that are homogeneous

**Research Designs**

Is there a difference in the capability outcomes of STEM programs when using a between-subject research design versus a within-subject design? To answer this question, this study compared the effect sizes between different research designs concerning the capability outcome. As mentioned in the engagement outcome section, each design has its own advantages and disadvantages. Table 5-12 presents the effect sizes for studies using between-subject design versus studies using within-subject design on the capability outcome. There were 44 studies using between-subject designs which produced 79 effect sizes, and 33 studies using within-subject designs which produced 74 effect sizes.

When the within-subject design was compared to the between-subject design, the within-subject design had a much higher average effect sizes than the between-subject design. The 79 effect sizes from the between-subject design ranged from 0.257 to 0.317 with a weighted average effect size of 0.287, a small, positive effect. On the other hand, the effect sizes from the within-subject design of STEM ranged from 0.629 to 0.695 with the average weighted effect size of 0.662, a large, positive effect. The 95% confidence
intervals for these average effect sizes did not include “0”, which means the average effect sizes were statistically significant.

Table 5 - 12 The Effect Sizes of STEM Programs on the Capability by Research Design

<table>
<thead>
<tr>
<th>Research Design</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between-Subject Design</td>
<td>44</td>
<td>79</td>
<td>0.257 to 0.317</td>
<td>.287</td>
<td>547.19***</td>
<td>263.54***</td>
</tr>
<tr>
<td>Within-Subject Design</td>
<td>33</td>
<td>74</td>
<td>0.629 to 0.695</td>
<td>.662</td>
<td>1111.32***</td>
<td></td>
</tr>
</tbody>
</table>

The homogeneity test for the two types of research designs revealed Qb of 263.54 and p < .001. The effectiveness of STEM programs on capability outcomes was heterogeneous across different research designs. On the other hand, a homogeneity test within each of the research designs revealed Qw values that exceeded the 95th percentile of the chi-square distribution, which means there was a high level of variance in capability outcomes within each research design.

**STEM Education Strategies**

Is there a difference in the capability outcomes of STEM programs when considering STEM education strategies? To answer this question, this study grouped pedagogical strategy into seven categories: Inquiry, Enhanced Context, Enhanced Materials, Assessment, Instructional Technology, Collaborative Learning and Mixed Strategies. Table 5-13 presents the distribution of effect sizes on the capability outcome by program strategy. There were 16 studies of STEM programs applying an inquiry strategy yielding 23 effect sizes. The most widely reported strategy used in STEM programs was an enhanced context strategy, with 63 effect sizes from 27 studies. Three
studies evaluated programs which applied an enhanced materials strategy; these three yielded five effect sizes. Two studies of applying an assessment strategy yielded two effect sizes. There were 18 studies applying an instruction technology strategy which yielded 28 effect sizes. Four studies applying a collaborative learning strategy yielded 16 effect sizes. Seven studies that applied a mixed strategy yielded 16 effect sizes.

Comparing strategies used in STEM programs gave a variety of effect sizes with the highest average effect size generated by the inquiry strategy, while the collaborative learning strategy had the lowest average effect size. The effect sizes from the inquiry strategy ranged from 0.658 to 0.757 with an average weighted effect size of 0.707, a large, positive effect. The enhanced context strategy had effect sizes ranging from 0.515 to 0.592 with an average weighted effect sizes of 0.553, a moderate, positive effect. The effect sizes of the enhanced materials strategy ranged from 0.541 to 0.947 with an average effect size of 0.744. The assessment strategy had an average effect size of .493. Meanwhile, the instructional technology strategy and collaborative learning strategy had small effects on the capability outcome. The effect size for the instructional technology strategy ranged from 0.201 to 0.347 with an average weighted effect size of 0.274, a small, positive effect. The effect sizes from the collaborative learning strategy ranged from 0.083 to 0.194 with an average weighted effect size of 0.139, a small, positive effect. Finally, the effect sizes of the mixed strategies ranged from 0.272 to 0.382 with an average weighted effect size of 0.327. The 95% confidence intervals for these average effect sizes did not include “0”, which means the average effect sizes were statistically significant.
Table 5 - 13 The Effect Sizes of STEM Programs on the Capability by Program Strategy

<table>
<thead>
<tr>
<th>STEM strategies</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inquiry</td>
<td>16</td>
<td>23</td>
<td>0.658 to 0.757</td>
<td>.707</td>
<td>171.94***</td>
<td>304.89***</td>
</tr>
<tr>
<td>Enhanced Context</td>
<td>27</td>
<td>63</td>
<td>0.515 to 0.592</td>
<td>.553</td>
<td>826.53***</td>
<td></td>
</tr>
<tr>
<td>Enhanced Materials</td>
<td>3</td>
<td>5</td>
<td>0.541 to 0.947</td>
<td>.744</td>
<td>30.27***</td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>2</td>
<td>2</td>
<td>0.247 to 0.738</td>
<td>.493</td>
<td>5.17**</td>
<td></td>
</tr>
<tr>
<td>Instructional Technology</td>
<td>18</td>
<td>28</td>
<td>0.201 to 0.347</td>
<td>.274</td>
<td>122.98***</td>
<td></td>
</tr>
<tr>
<td>Collaborative Learning</td>
<td>4</td>
<td>16</td>
<td>0.083 to 0.194</td>
<td>.139</td>
<td>48.65***</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>7</td>
<td>16</td>
<td>0.272 to 0.738</td>
<td>.327</td>
<td>411.63***</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>77</td>
<td>153</td>
<td>0.083 to 0.947</td>
<td>.454</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The homogeneity test for STEM education strategies revealed Qb of 304.89 and p < .001. The effectiveness of STEM programs on capability outcomes was heterogeneous across different pedagogical strategies. However, a homogeneity test within each of the strategies revealed Qw values that exceeded the 95th percentile of the chi-square distribution, which means there was a high level of variance in the capability levels within each strategy.

**Funding Source**

Is there a difference in the capability outcomes of STEM programs when considering different funding sources? To answer this question, this study conducted comparisons across different funding sources. This meta-analysis identified three funding sources of STEM programs: government, non-profit organization, and mixed (cooperative). Table 5-14 shows the distribution of effect sizes by funding sources. Of the 77 studies that evaluated capability, the 31 studies of programs that were funded by all...
levels of government produced 68 effect sizes. There were 15 studies of programs funded by non-profit organizations produced 35 effect sizes and two studies funded by the mixed organizations produced seven effect sizes. Finally, the 29 studies that produced 43 effect sizes did not provide any information about STEM-program funding source.

The 68 effect sizes from government funded STEM programs ranged from 0.550 to 0.582 with an average weighted effect size of 0.566, a moderate, positive effect. The 35 effect sizes from the STEM programs funded by nonprofit organizations ranged from 0.248 to 0.331 with an average weighted effect size of 0.298, a small, positive effect. Finally, the seven effect sizes from STEM programs funded by the mixed organizations (cooperative) ranged from 0.380 to 0.526 with an average weighted effect size of 0.405, a moderate, positive effect. The 95% confidence intervals for these average effect sizes did not include “0”, which means the average effect sizes were statistically significant.

Table 5 - 14 The Effect Sizes of STEM Programs on the Capability by Founding Source

<table>
<thead>
<tr>
<th>Funding Source</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government</td>
<td>31</td>
<td>68</td>
<td>0.550 to 0.582</td>
<td>.582</td>
<td>1111.32***</td>
<td>136.06</td>
</tr>
<tr>
<td>Nonprofit Organization</td>
<td>15</td>
<td>35</td>
<td>0.248 to 0.331</td>
<td>.298</td>
<td>516.10***</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>2</td>
<td>7</td>
<td>0.380 to 0.526</td>
<td>.405</td>
<td>80.51***</td>
<td></td>
</tr>
<tr>
<td>No Information</td>
<td>29</td>
<td>43</td>
<td>0.350 to 0.500</td>
<td>.416</td>
<td>214.12***</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>77</td>
<td>153</td>
<td>0.248 to 0.582</td>
<td>.454</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The program funding source was not a successful moderator because it did not create a homogeneous within group variance and a heterogeneous between group variance. It actually showed a heterogeneous engagement outcome within each funding source and homogeneity across different program funding sources. The homogeneity test
for funding source revealed $Q_b$ of 136.06 and $p > .05$. A homogeneity test within each of the strategies revealed $Q_w$ values that exceeded the 95th percentile of the chi-square distribution. The effect sizes within funding sources were thus heterogeneous, which means there was a high level of variance in the capability levels within each funding source.

**STEM Program Creator**

Is there a difference in the capability outcomes of STEM programs when considering the program creators? To answer this question, this meta-analysis compared five different organization types: professional associations (e.g. the National Teacher Association), private organizations (e.g. Apple computer and IBM Corporation), governments (federal, state, and local government), non-profit organizations (private and public schools), mixed organizations, and no information. Even though the creator category is similar to the funder category, there is a distinction between the program creators which actually participated in creating programs, and program funders which provides funds for programs. Although the creator and funder can be the same person or entity, they do not need to be.

Table 5-15 shows the distribution of effect sizes by program creator. Of the 77 studies, eight studies created by professional associations yielded 19 effect sizes, three studies created by private organization produced three effect sizes, eight studies created by governmental entities yielded 17 effect sizes, 46 studies created by non-profit entities produced 89 effect sizes, eight studies created by cooperative creators (e.g. programs created jointly by governmental and non-profit entities) yielded 20 effect sizes, and four studies with no information about the STEM program creator produced five effect sizes.
STEM programs created by professional associations, private organizations, and mixed (cooperative) creators had larger effect sizes than those created by governments and non-profit organizations. The 19 effect sizes from programs created by professional associations ranged from 0.596 to 0.711 with an average weighted effect size of 0.653, a large, positive effect. On the other hand, the three effect sizes from programs created by private organizations ranged from -0.390 to 0.272 with an average weighed effect sizes of 0.059. The 17 effect sizes from programs created by governmental entities ranged from 0.114 to 0.214 with an average weighed effect size of 0.164, a small, positive effect. The 89 effect sizes from programs created by non-profit entities ranged from 0.470 to 0.530 with an average weighed effect size of 0.450, a moderate, positive effect. The 20 effect sizes from programs created by multiple, cooperative entities ranged from 0.450 to 0.672 with an average weighed effect size of 0.585, a moderate, positive effect. The six effect sizes from programs with no creator information ranged from 0.039 to 0.395 with an average weighed effect size of 0.217, a small, positive effect. The majority of the 95% confidence intervals for these average effect sizes did not include “0”, which means the average effect sizes were statistically significant.
Table 5-15: The Effect Sizes of STEM Programs on the Capability by Program Creator

<table>
<thead>
<tr>
<th>Program Creator</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional Association</td>
<td>8</td>
<td>19</td>
<td>0.596 to 0.711</td>
<td>.653</td>
<td>101.66***</td>
<td>210.54***</td>
</tr>
<tr>
<td>Private Organization</td>
<td>3</td>
<td>3</td>
<td>-0.390 to 0.272</td>
<td>.059</td>
<td>50.45***</td>
<td>210.54***</td>
</tr>
<tr>
<td>Government</td>
<td>8</td>
<td>17</td>
<td>0.114 to 0.214</td>
<td>.164</td>
<td>35.61***</td>
<td>210.54***</td>
</tr>
<tr>
<td>Non-profit Organization</td>
<td>46</td>
<td>89</td>
<td>0.470 to 0.530</td>
<td>.450</td>
<td>1319.04***</td>
<td>210.54***</td>
</tr>
<tr>
<td>Mixed (cooperative)</td>
<td>8</td>
<td>20</td>
<td>0.450 to 0.672</td>
<td>.585</td>
<td>179.54***</td>
<td>210.54***</td>
</tr>
<tr>
<td>No Information</td>
<td>4</td>
<td>5</td>
<td>0.039 to 0.395</td>
<td>.217</td>
<td>25.21***</td>
<td>210.54***</td>
</tr>
<tr>
<td>Total</td>
<td>77</td>
<td>153</td>
<td>-0.390 to 0.711</td>
<td>.454</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Professional associations include the National Teacher Association and educators (teacher). Government includes federal, state, and local government. Nonprofit organization includes private and public schools.

The homogeneity test of the different types of STEM program creators revealed Qb of 210.54 and p < .001. This suggests that the effectiveness of STEM programs on capability outcomes was heterogeneous across different programs creators. However, a homogeneity test for each program creator revealed Qw values that exceeded the 95th percentile of the chi-square distribution. The results mean that different types of creators generated heterogeneous capability outcomes.

**Educational Levels of Sample Group**

Is there a difference in the capability outcomes of STEM programs when considering education levels? To answer this question, this study classified six sample groups: elementary, middle, high, college, K-12 students, and K-12 teachers. Table 5-16 shows the distribution of effect sizes by educational levels. Of the 79 studies, 11 studies measuring elementary students’ capability yielded 20 effect sizes, six studies measuring middle school students yielded 12 effect sizes, 11 studies sampling high school students
produced 19 effect sizes, 23 studies measuring college students yielded 52 effect sizes, six studies sampling K-12 students yielded 11 effect sizes, and 20 studies sampling K-12 teachers yielded 39 effect sizes.

Elementary and middle school students experienced the most effective STEM outcomes in terms of capability. Elementary students had effect sizes ranging from 0.840 to 0.938, with an average weighted effect size of 0.889, a large, positive effect. Middle school students had effect sizes ranging from 0.786 to 0.974, with an average weighted effect size of 0.880, a large, positive effect. On the other hand, the remaining groups had moderate effects. High school students had effect sizes ranging from 0.153 to 0.335, with an average weighted effect size of 0.243, a small, positive effect. College students had effect sizes ranging from 0.371 to 0.458, with an average weighted effect size of 0.414, a moderate, positive effect. K-12 students had effect sizes ranging from 0.081 to 0.167, with an average weighed effect size of 0.124, a small, positive effect. K-12 teachers had effect sizes ranging from 0.366 to 0.490, with an average weighted effect size of 0.428, a moderate, positive effect. The 95% confidence intervals for all average effect sizes did not include “0”, which means the average effect sizes were statistically significant.
Table 5-16 The Effect Sizes of STEM Programs on the Capability by Educational Level

<table>
<thead>
<tr>
<th>Educational Level</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary Students</td>
<td>11</td>
<td>20</td>
<td>0.840 to 0.938</td>
<td>.889</td>
<td>364.89***</td>
<td>634.66***</td>
</tr>
<tr>
<td>Middle School Students</td>
<td>6</td>
<td>12</td>
<td>0.786 to 0.974</td>
<td>.880</td>
<td>161.26***</td>
<td></td>
</tr>
<tr>
<td>High School Students</td>
<td>11</td>
<td>19</td>
<td>0.153 to 0.335</td>
<td>.243</td>
<td>77.93***</td>
<td></td>
</tr>
<tr>
<td>College Students</td>
<td>23</td>
<td>52</td>
<td>0.371 to 0.458</td>
<td>.414</td>
<td>361.99***</td>
<td></td>
</tr>
<tr>
<td>K-12 Students</td>
<td>6</td>
<td>11</td>
<td>0.081 to 0.167</td>
<td>.124</td>
<td>150.68***</td>
<td></td>
</tr>
<tr>
<td>K-12 Teachers</td>
<td>20</td>
<td>39</td>
<td>0.366 to 0.490</td>
<td>.428</td>
<td>170.64***</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>77</td>
<td>153</td>
<td>0.081 to 0.974</td>
<td>.454</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The homogeneity test of education levels of sample groups revealed Qb of 634.66 and p < .001. The effect sizes of STEM programs on capability outcome were heterogeneous among the six educational levels. A homogeneity test within each of the educational levels revealed Qw values that exceeded the 95th percentile of the chi-square distribution in all sub-classes. Therefore, all sub-classes had heterogeneity, which means there was a high level of variance in the capability levels within each education level.

**Regional Location**

Is there a difference in the capability outcomes of STEM programs when considering the regions that STEM programs were conducted in? To answer this question, this study classified into five regions: West, South, Midwest, Northeast, nationwide, and no information. Table 5-17 shows the distribution of effect sizes by regional location. There were eight studies in the West produced 24 effect sizes, 26 studies in the South produced 50 effect sizes, 18 studies in Midwest produced 38 effect sizes, 14 studies in Northeast produced 21 effect sizes, six studies conducted nationwide produced 13 effect sizes, and five studies with seven effect sizes did not provide regional information.
The Western and Southern regions produced the highest effect sizes of the six American regions. The 24 effect sizes from STEM programs implemented in the West ranged from 0.607 to 0.720 with an average weighted effect size of 0.663, a large, positive effect. The 50 effect sizes from STEM programs implemented in the South ranged from 0.512 to 0.598 with an average weighted effect size of 0.555, a moderate, positive effect. In contrast, STEM programs carried out in the Midwest and Northeast each had relatively small, positive average effect sizes. The 38 effect sizes from STEM programs implemented in the Midwest ranged from 0.218 to 0.304 with an average weighted effect size of 0.261, a small, positive effect. The 20 effect sizes from STEM programs implemented in the Northeast ranged from 0.336 to 0.495 with an average weighted effect size of 0.415, a moderate, positive effect. The 14 effect sizes from STEM programs implemented nationwide ranged from 0.154 to 0.273 with an average weighted effect size of 0.214, a small, positive effect. The 95% confidence intervals for these average effect sizes did not include “0”, which means the average effect sizes were statistically significant.

Table 5 - 17 The Effect Sizes of STEM Programs on the Capability by Regional Location

<table>
<thead>
<tr>
<th>Regional Location</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>West</td>
<td>8</td>
<td>24</td>
<td>0.607 to 0.720</td>
<td>.663</td>
<td>624.96***</td>
<td>298.08***</td>
</tr>
<tr>
<td>South</td>
<td>26</td>
<td>50</td>
<td>0.512 to 0.598</td>
<td>.555</td>
<td>406.97***</td>
<td></td>
</tr>
<tr>
<td>Midwest</td>
<td>18</td>
<td>38</td>
<td>0.218 to 0.304</td>
<td>.261</td>
<td>450.17***</td>
<td></td>
</tr>
<tr>
<td>Northeast</td>
<td>14</td>
<td>21</td>
<td>0.336 to 0.495</td>
<td>.415</td>
<td>62.96***</td>
<td></td>
</tr>
<tr>
<td>Nationwide</td>
<td>6</td>
<td>13</td>
<td>0.154 to 0.273</td>
<td>.214</td>
<td>69.83***</td>
<td></td>
</tr>
<tr>
<td>No information</td>
<td>5</td>
<td>7</td>
<td>0.754 to 0.918</td>
<td>.836</td>
<td>9.09(H)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>77</td>
<td>153</td>
<td>0.154 to 0.918</td>
<td>.454</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The homogeneity test of regional locations where STEM programs were carried out revealed $Q_b$ of 298.08 and $p < .001$. The effect sizes of STEM programs on capability outcomes were heterogeneous across different regional locations. A test for homogeneity within each of the regions revealed $Q_w$ values that exceeded the 95th percentile of the chi-square distribution, which means the effect sizes had a high level of variance within each region.

**Publication Year**

Is there a difference in capability outcomes of STEM programs when considering the publication year? To answer this question, this meta-analysis compared effect sizes by publication year. Table 5-18 shows the distribution of effect sizes by publication year. More than half of the studies that analyzed the engagement outcome in this study were published between 2000 and 2010. The 45 studies in that period produced 85 effect sizes. The 23 studies published between 1990 and 1999 produced 55 effect sizes. Nine studies published between 1980 and 1989 had 13 effect sizes.

The effect sizes from the studies published from 2000 to 2010 ranged from 0.514 to 0.571 with an average weighted effect size of 0.542, a moderate, positive effect. In contrast, the effect sizes from 1990 to 1999 ranged from 0.265 to 0.342 with an average weighted effect size of 0.304, a small, positive effect. The effect sizes for the studies published from 1980 to 1989 ranged from 0.324 to 0.515 with an average weighted effect size of 0.419, a moderate, positive effect. The 95% confidence intervals for these average effect sizes did not include “0”, which means the average effect sizes were statistically significant.
Table 5 - The Effect Sizes of STEM Programs on the Capability by Publication Year

<table>
<thead>
<tr>
<th>Year</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>( Q_w )</th>
<th>( Q_b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000-2010</td>
<td>45</td>
<td>85</td>
<td>0.514 to 0.571</td>
<td>.542</td>
<td>1229.46***</td>
<td>95.58</td>
</tr>
<tr>
<td>1990-1999</td>
<td>23</td>
<td>55</td>
<td>0.265 to 0.342</td>
<td>.304</td>
<td>555.13***</td>
<td></td>
</tr>
<tr>
<td>1980-1999</td>
<td>9</td>
<td>13</td>
<td>0.324 to 0.515</td>
<td>.419</td>
<td>41.89***</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>77</td>
<td>153</td>
<td>0.265 to 0.571</td>
<td>.454</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The publication year was not a successful moderator because it did not create a homogeneous within group variance and a heterogeneous between group variance. The homogeneity test of the publication year of individual studies revealed \( Q_b \) of 95.58 and \( p > .05 \). This suggests that there was no significant difference across the publication year. The homogeneity test within each of the published years revealed \( Q_w \) values that exceeded the 95th percentile of the chi-square distribution. The results mean that there was a high level of variations within the published year groups.

Reliability Evidence for Instrument

Is there a difference in the capability outcomes of STEM programs when considering whether or not the studies provided reliability evidence in the study? To answer this question, this meta-analysis classified three instrument reliability levels: “more than .7 reliability”, and “less than .7 reliability”, and “no information”. The studies that did not provide the reliability level of the measurement instrument were identified “no information”. Table 5-19 presents the distribution of effect sizes on the capability outcome by reliability evidence. Of the 77 studies on capability, 27 studies indicating “more than .7 reliability” produced 61 effect sizes. Only three studies explicitly stated
that the reliability was less than .7 and yielded eight effect sizes. The 47 classified as “no information” produced 84 effect sizes.

The studies that explicitly stated reliability less than .7 had higher effect sizes than those classified as “no information” or “more than .7”. The effect sizes from studies with “more than .7 reliability” ranged from 0.522 to 0.590 with an average weighted effect size of 0.556, a moderate, positive effect. Alternatively, the effect sizes from studies with “less than .7 reliability” ranged from 0.978 to 1.196 with an average weighted effect size of 1.087, a large, positive effect. The effect sizes from the “no information” classification of the measurement instrument ranged from 0.285 to 0.347, with an average weighted effect size of 0.316, a small, positive effect. The 95% confidence intervals for these average effect sizes did not include “0”, which means the average effect sizes were statistically significant.

Table 5 - 19 The Effect Sizes of STEM Programs on the Capability by Reliability Evidence

<table>
<thead>
<tr>
<th>Reliability</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>More than .7</td>
<td>27</td>
<td>61</td>
<td>0.522 to 0.590</td>
<td>.556</td>
<td>1110.56***</td>
<td>240.66***</td>
</tr>
<tr>
<td>Less than .7</td>
<td>3</td>
<td>8</td>
<td>0.978 to 1.196</td>
<td>1.087</td>
<td>51.74***</td>
<td></td>
</tr>
<tr>
<td>No Information</td>
<td>47</td>
<td>84</td>
<td>0.285 to 0.347</td>
<td>.316</td>
<td>519.08***</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>77</td>
<td>153</td>
<td>0.285 to 1.196</td>
<td>.454</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The homogeneity test of the instrument reliability levels revealed $Q_b$ of 240.66 and $p < .001$. This suggests that the effect sizes of STEM programs on capability outcomes were heterogeneous across the instrument reliability levels. A homogeneity test within each of the sub-classes revealed $Q_w$ values that exceeded the 95th percentile of the
chi-square distribution. The sub-classes were heterogeneous, which means there was a high level of variance of capability outcomes within each instrument reliability level.

**5.4 Moderator Variables for the Continuity Outcome**

The 17 studies included in this analysis produced 25 effect sizes for the continuity outcomes of STEM programs. The effect sizes ranged from 0.321 to 0.418, with a weighted averaged effect size of 0.369. The weighted effect size indicates that STEM programs were moderately successful in the students’ continuity outcome. Whether there is significant difference of the effect sizes resulting from the moderator variables remains an empirical question. One of the main purposes of this study is to determine whether moderator variables can explain the heterogeneous effect sizes on the continuity outcomes of STEM programs. Table 5-20 presents the effects of the independent variable and significant moderator variables on the continuity outcome. A successful moderator is supposed to produce homogeneous effect sizes within the same level of the moderator variable and heterogeneous effect sizes across different levels of the variable. In other words, the $Q_w$ in Table 5.24 is not supposed to be significant, and the $Q_b$ in Table 5.24 is supposed to be significant.
<table>
<thead>
<tr>
<th>Study Method</th>
<th>Moderator Variable</th>
<th>Mean ( (d) )</th>
<th>(n)</th>
<th>( Q_w )</th>
<th>( Q_b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research Design</td>
<td>Between Subject Design</td>
<td>.337</td>
<td>15</td>
<td>47.40***</td>
<td>4.59</td>
</tr>
<tr>
<td></td>
<td>Within Subject Design</td>
<td>.457</td>
<td>10</td>
<td>73.60***</td>
<td></td>
</tr>
<tr>
<td>Evidence</td>
<td>More than&gt; .7</td>
<td>.815</td>
<td>4</td>
<td>33.04***</td>
<td>41.78**</td>
</tr>
<tr>
<td></td>
<td>Less than&lt; .7</td>
<td>.269</td>
<td>3</td>
<td>4.47(H)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No information</td>
<td>.314</td>
<td>18</td>
<td>45.80***</td>
<td></td>
</tr>
<tr>
<td>Program Characteristics</td>
<td>Program Strategies</td>
<td>Inquiry</td>
<td>.201</td>
<td>4</td>
<td>2.25(H)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Enhanced Context</td>
<td>.294</td>
<td>14</td>
<td>38.51***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Collaborative Learning</td>
<td>.411</td>
<td>3</td>
<td>2.78(H)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mixed</td>
<td>1.024</td>
<td>3</td>
<td>15.54***</td>
</tr>
<tr>
<td></td>
<td>Funder</td>
<td>Government</td>
<td>.368</td>
<td>6</td>
<td>11.39(H)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nonprofit Organization</td>
<td>.286</td>
<td>13</td>
<td>22.73**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No Information</td>
<td>.710</td>
<td>5</td>
<td>48.62***</td>
</tr>
<tr>
<td></td>
<td>Program Creator</td>
<td>Government</td>
<td>.144</td>
<td>4</td>
<td>.93(H)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nonprofit Organization</td>
<td>.407</td>
<td>15</td>
<td>103.42***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No Information</td>
<td>.231</td>
<td>5</td>
<td>8.83(H)</td>
</tr>
<tr>
<td>Participant Characteristics</td>
<td>Grade Level</td>
<td>High School</td>
<td>.621</td>
<td>2</td>
<td>8.39***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>College</td>
<td>.354</td>
<td>17</td>
<td>94.57***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K-12 Students</td>
<td>.315</td>
<td>5</td>
<td>7.51(H)</td>
</tr>
<tr>
<td>Setting Environment</td>
<td>Regional Location</td>
<td>West</td>
<td>.228</td>
<td>7</td>
<td>9.92(H)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Midwest</td>
<td>.460</td>
<td>8</td>
<td>23.36***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Northeast</td>
<td>.148</td>
<td>4</td>
<td>3.24(H)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>South</td>
<td>.782</td>
<td>5</td>
<td>40.05***</td>
</tr>
<tr>
<td>Study Characteristics</td>
<td>Published year</td>
<td>2000-2010</td>
<td>.338</td>
<td>17</td>
<td>66.19***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1990-1999</td>
<td>.698</td>
<td>5</td>
<td>33.73***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1980-1989</td>
<td>.175</td>
<td>3</td>
<td>1.14(H)</td>
</tr>
</tbody>
</table>
Research Designs

Is there a difference in the continuity of students in STEM fields when using a between-subject research design versus a within-subject design in STEM education programs? To answer this question, this study compared the effect sizes between different research designs. Table 5-21 presents the distribution of effect sizes for studies using between-subject design versus studies using within-subject design on the continuity outcome. There were 17 studies evaluating STEM programs on continuity outcomes produced 25 effect sizes, ten studies using a between-subject design produced 15 effect sizes and seven studies using a within-subject design produced 10 effect sizes.

The effect sizes ranged from 0.280 to 0.394 for the between-subject design with an average weighted effect size of 0.337, a small, positive effect. The effect sizes from the within-subject design ranged from 0.364 to 0.551 with an average weighted effect size of 0.457, a moderate, positive effect. The 95% confidence intervals for these average effect sizes did not include “0”, which means the average effect sizes were statistically significant.

<table>
<thead>
<tr>
<th>Research Design</th>
<th># of Studies</th>
<th># of effect sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Q_w</th>
<th>Q_b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between-Subject Design</td>
<td>10</td>
<td>15</td>
<td>0.280 to 0.394</td>
<td>.337</td>
<td>47.40***</td>
<td>4.59</td>
</tr>
<tr>
<td>Within-Subject Design</td>
<td>7</td>
<td>10</td>
<td>0.364 to 0.551</td>
<td>.457</td>
<td>73.60***</td>
<td></td>
</tr>
</tbody>
</table>

Note: *** refers to weighted mean effect sizes that are heterogeneous at .001 level
** refers to weighted mean effect sizes that are heterogeneous at .05 level
H refers to weighted mean effect sizes that are homogeneous
The research design was not a successful moderator because it did not create a homogeneous within group variance and a heterogeneous between group variance. The homogeneity test of the two research design types revealed $Q_b$ of 4.59 and $p > 0.5$. A homogeneity test within each of the research designs revealed $Q_w$ values that exceeded the 95th percentile of the chi-square distribution. This result indicated that there was a high level of variance in the students’ continuity outcomes within each research design.

**STEM Education Strategies**

Is there a difference in the continuity outcomes of STEM programs when considering the educational strategies in the STEM programs? To answer this question, this study grouped pedagogical strategy into four categories: enhanced context, inquiry, collaborative learning, and mixed strategies. Table 5-22 shows the distribution of effect sizes on the continuity outcomes by educational strategies. The most widely reported strategy used for the students’ continuity in STEM programs was an enhanced context strategy, with 14 effect sizes from eight studies. There were four studies using an inquiry strategy yielded four effect sizes, two studies using a collaborative learning strategy produced three effect sizes, and two studies using a mixed strategy produced three effect sizes.

The effect sizes from the individual research studies ranged from 0.226 to 0.362 for the enhanced context strategy with a mean of 0.294, a small, positive effect. The effect sizes of inquiry strategy ranged from 0.056 to 0.346 with an average weighted effect size of 0.201, a small, positive effect. The effect sizes of collaborative learning strategy ranged from 0.312 to 0.510 with an average of 0.411, a moderate, positive effect. The effect sizes of the mixed strategy ranged from 0.939 to 1.505 with an average of
1.024. The 95% confidence intervals for these average effect sizes did not include “0”, which means the average effect sizes were statistically significant.

<table>
<thead>
<tr>
<th>STEM strategies</th>
<th># of Study</th>
<th># of Effect Size</th>
<th>95% C. I. Range</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enhanced Context</td>
<td>8</td>
<td>14</td>
<td>0.226 to 0.362</td>
<td>.294</td>
<td>38.16***</td>
<td>66.36***</td>
</tr>
<tr>
<td>Inquiry</td>
<td>4</td>
<td>4</td>
<td>0.056 to 0.346</td>
<td>.201</td>
<td>2.25(H)</td>
<td></td>
</tr>
<tr>
<td>Collaborative Learning</td>
<td>2</td>
<td>3</td>
<td>0.312 to 0.510</td>
<td>.411</td>
<td>2.78(H)</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>2</td>
<td>3</td>
<td>0.939 to 1.505</td>
<td>1.024</td>
<td>15.54***</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>24</td>
<td>0.056 to 1.505</td>
<td>.374</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Sample size (n) varies because of only one observation (Instructional Technology Strategy was removed). H refers to weighted mean effect sizes that are homogeneous.

The homogeneity test of STEM program strategies revealed Qb of 66.36 and p <.001, which indicated that the effect sizes of STEM programs on continuity outcomes were heterogeneous across different strategies. A homogeneity test within the enhanced context and mixed strategies revealed Qw values that exceeded the 95th percentile of the chi-square distribution, which means there was a high level of variance within each strategy. In contrast, the homogeneity tests within the inquiry and collaborative learning strategies revealed Qw values that did not exceed the 95th percentile of the chi-square distribution, which means the effect sizes were homogeneous.

**Funding Source**

Is there a difference in the continuity outcomes of STEM programs when considering different funding sources? To answer this question, this study conducted comparisons across different funding sources. This meta-analysis identified four funding sources of STEM programs: government, non-profit organization, and mixed.
(cooperative). Table 5-23 shows the distribution of effect sizes by funding sources. Of the 16 studies that evaluated continuity, five studies of programs that were funded by all levels of government produced six effect sizes. There were seven studies of programs funded by non-profit organizations produced 13 effect sizes. Finally, four studies that produced five effect sizes did not provide any information about STEM-program funding source.

The six effect sizes from government-funded STEM programs ranged from 0.242 to 0.494 with an average weighted effect size of 0.368, a small, positive effect. The 13 effect sizes from the STEM programs funded by nonprofit organizations ranged from 0.226 to 0.347 with an average weighted effect size of 0.286, a small, positive effect. Finally, the five effect sizes from STEM programs with “no information” ranged from 0.576 to 0.845 with an average weighted effect size of 0.710, a large, positive effect. The 95% confidence intervals for these average effect sizes did not include “0”, which means the average effect sizes were statistically significant.

<table>
<thead>
<tr>
<th>Funding Source</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government</td>
<td>5</td>
<td>6</td>
<td>0.242 to 0.494</td>
<td>.368</td>
<td>11.39(H)</td>
<td>42.32**</td>
</tr>
<tr>
<td>Nonprofit Organization</td>
<td>7</td>
<td>13</td>
<td>0.226 to 0.347</td>
<td>.286</td>
<td>22.73**</td>
<td></td>
</tr>
<tr>
<td>No Information</td>
<td>4</td>
<td>5</td>
<td>0.576 to 0.845</td>
<td>.710</td>
<td>48.62***</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>24</td>
<td>0.242 to 0.845</td>
<td>.435</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The program funding source was not a successful moderator because it did not create a homogeneous within group variance and a heterogeneous between group variance. The homogeneity test for funding source revealed Qb of 42.32 and p < .05. This
suggests that the effect sizes of STEM programs’ continuity outcomes were heterogeneous across funding sources. A homogeneity test within governments revealed $Q_w$ values that did not exceed the 95th percentile of the chi-square distribution. The mean effect size .368 is a typical representation of continuity outcome of STEM programs among governments. In contrast, the homogeneity tests within nonprofit organizations and “no information” revealed $Q_w$ values that exceeded the 95th percentile of the chi-square distribution.

**STEM Program Creator**

Is there a difference in the continuity outcomes of STEM programs when considering the program creators? To answer this question, this meta-analysis compared three different organization types: governments (federal, state, and local government), non-profit organizations (private and public schools), and no information. Table 5-24 shows the distribution of effect sizes by program creator. Of the 16 studies, three studies created by governmental entities yielded four effect sizes, eight studies created by non-profit entities produced 15 effect sizes, and five studies with no information about the STEM program creator produced five effect sizes.

The four effect sizes from programs created by governmental entities ranged from -0.025 to 0.313 with an average weighted effect size of 0.144. On the other hand, the effect sizes from programs created by nonprofit organizations ranged from 0.352 to 0.462 with an average weighed effect size of 0.407, a moderate, positive effect. The effect sizes from programs with no creator information ranged from 0.075 to 0.387 with an average
weighed effect size of 0.231, a small, positive effect. The majority of the 95% C.I. did not include 0, which means these average effect sizes were statistically significant.

<table>
<thead>
<tr>
<th>Program Creator</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government</td>
<td>3</td>
<td>4</td>
<td>-0.025 to 0.313</td>
<td>.144</td>
<td>0.93(H)</td>
<td>12.11</td>
</tr>
<tr>
<td>Non-profit Organization</td>
<td>8</td>
<td>15</td>
<td>0.352 to 0.462</td>
<td>.407</td>
<td>103.42***</td>
<td></td>
</tr>
<tr>
<td>No Information</td>
<td>5</td>
<td>5</td>
<td>0.075 to 0.387</td>
<td>.231</td>
<td>8.83(H)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>24</td>
<td>-0.025 to 0.462</td>
<td>.326</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Government includes federal, state, and local government. Nonprofit organization includes private and public schools.

The program creator was not a successful moderator because it did not create a homogeneous within group variance and a heterogeneous between group variance. The homogeneity test of the different types of STEM program creators revealed Qb of 12.11 and p > .05. This suggests that the effectiveness of STEM programs on continuity outcomes was homogeneous across different programs creators. Although a homogeneity test within nonprofit organization creator revealed Qw values that exceeded the 95th percentile of the chi-square distribution, the homogeneity tests within government revealed Qw values that did not exceeded the 95th percentile of the chi-square distribution.

**Educational Levels of Sample Group**

Is there a difference in the continuity outcomes of STEM programs when considering education levels? To answer this question, this study classified four sample groups: high, college, and K-12 students. Table 5-25 shows the distribution of effect sizes by educational levels. Of the 15 studies, two studies sampling high school students...
produced two effect sizes, nine studies measuring college students yielded 17 effect sizes, and four studies sampling K-12 students yielded five effect sizes.

High school and college students had moderator effects in terms of continuity outcome. High school students had effect sizes ranging from 0.181 to 0.741, with an average weighted effect size of 0.621, a moderate, positive effect. College students had effect sizes ranging from -0.065 to 1.882, with an average weighted effect size of 0.354, a small, positive effect. K-12 students had effect sizes ranging from 0.090 to 0.582, with an average weighed effect size of 0.315, a small, positive effect. The 95% confidence intervals for these average effect sizes did not include “0”, which means the average effect sizes were statistically significant.

<table>
<thead>
<tr>
<th>Educational Level</th>
<th># of Studies</th>
<th># of Effect Sizes</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>High School Students</td>
<td>2</td>
<td>2</td>
<td>0.466 to 0.776</td>
<td>.621</td>
<td>8.39***</td>
<td>14.61</td>
</tr>
<tr>
<td>College Students</td>
<td>9</td>
<td>17</td>
<td>0.298 to 0.411</td>
<td>.354</td>
<td>94.57***</td>
<td></td>
</tr>
<tr>
<td>K-12 Students</td>
<td>4</td>
<td>5</td>
<td>0.185 to 0.445</td>
<td>.351</td>
<td>7.51(H)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>24</td>
<td>0.185 to 0.776</td>
<td>.367</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The homogeneity test of education levels of sample groups revealed Qb of 14.61 and p > .05. This suggests that the effect sizes of STEM programs on continuity outcome were homogeneous among the four educational levels. A homogeneity test within high school and college students revealed Qw values that exceeded the 95th percentile of the chi-square distribution, which means there was a high level of variance in the continuity levels within each education level. In contrast, the homogeneity tests within K-12 students revealed Qw values that did not exceed the 95th percentile of the chi-square distribution, which means the effect sizes were homogeneous.
Regional Location

Is there a difference in the continuity outcomes of STEM programs when considering the different regional areas where STEM programs were conducted? To answer this question, this meta-analysis classified program locations into four regions: South, West, Midwest, and Northeast. Table 5-26 shows the distribution of effect sizes on the continuity outcome by four regions. In the South, four studies produced five effect sizes. In the West, three studies produced seven effect sizes. In the Midwest, six studies produced eight effect sizes. In the Northeast, three studies produced four effect sizes.

The South produced the highest effect sizes of the four US regions. The effect sizes of STEM programs implemented in the South ranged from 0.615 to 0.949 with an average weighted effect size of 0.782, a large, positive effect. In contrast, the STEM programs in the West and Northeast each had small average weighted effect sizes. The effect sizes of STEM programs implemented in the West ranged from 0.151 to 0.304 with an average weighted effect size of 0.228, a small, positive effect. The effect sizes from STEM programs implemented in the Northeast ranged from -0.028 to 0.324 with an average weighted effect size of 0.148. The effect sizes of STEM programs implemented in the Midwest ranged from 0.383 to 0.538 with an average weighted effect size of 0.460, a moderate, positive effect. The majority of the 95% C.I. did not include 0, which means these average effect sizes were statistically significant.
The homogeneity test of the regional location where STEM programs were conducted revealed $Q_w$ of 48.53 and $p < .05$. This suggested that the effect sizes of STEM programs on continuity outcome were heterogeneous across the four regional locations. A test for homogeneity within the Midwest and South revealed $Q_w$ values that exceeded the 95th percentile of the chi-square distribution, which means there was a high level of variance within these regions. In contrast, the homogeneity test within the West and Northeast revealed $Q_w$ values that did not exceed the 95th percentile of the chi-square distribution, which means the effect sizes within that region were homogeneous. The mean effect size .228 is a typical representation of continuity outcome of STEM programs in West region. The mean effect size .148 is a typical representation in Northeast region.

**Reliability Evidence for Instrument**

Is there a difference in the continuity outcomes of STEM programs when considering the reliability of the measurement? To answer this question, this meta-analysis classified three instrument reliability levels: “more than .7”, “less than .7” and

---

**Table 5 - The Effect Sizes of STEM Programs on the Continuity by Regional Location**

<table>
<thead>
<tr>
<th>Regional Location</th>
<th># of Study</th>
<th># of Effect Size</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>$Q_w$</th>
<th>$Q_b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>South</td>
<td>4</td>
<td>5</td>
<td>0.615 to 0.949</td>
<td>.782</td>
<td>40.05***</td>
<td>48.53**</td>
</tr>
<tr>
<td>West</td>
<td>3</td>
<td>7</td>
<td>0.151 to 0.304</td>
<td>.228</td>
<td>9.92(H)</td>
<td></td>
</tr>
<tr>
<td>Northeast</td>
<td>3</td>
<td>4</td>
<td>-0.028 to 0.324</td>
<td>.148</td>
<td>3.24(H)</td>
<td></td>
</tr>
<tr>
<td>Midwest</td>
<td>6</td>
<td>8</td>
<td>0.383 to 0.538</td>
<td>.460</td>
<td>23.36***</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>24</td>
<td>-0.028 to 0.949</td>
<td>.391</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Sample size (n) varies because of only one observation (no information about region was removed.) H refers to weighted mean effect sizes that are homogeneous.
“no information”. Table 5-27 shows the distribution of effect sizes on the continuity outcome by reliability levels. Of the 17 studies on continuity outcome, three studies with “more than .7 reliability” produced five effect sizes. The only two studies with “less than .7 reliability” produced three effect sizes. The 12 studies in which the reliability of the measurement instrument was classified as “no information” produced 18 effect sizes.

Studies with “more than .7 reliability” had higher effect sizes than those studies with “less than .7” or with the “no information” classification of measurement instruments. The effect sizes from studies with “more than .7 reliability” ranged from 0.671 to 0.959 with an average weighted effect size of 0.815, a large, positive effect. In contrast, the effect sizes from the studies with “less than .7 reliability” ranged from 0.056 to 0.482 with an average weighted effect size of 0.269, a small, positive effect. The effect sizes from the studies identified as “no information” ranged from 0.261 to 0.368 with an average weighted effect size of 0.314, a small, positive effect. The 95% confidence intervals for these average effect sizes did not include “0”, which means the average effect sizes were statistically significant.

<table>
<thead>
<tr>
<th>Five Division</th>
<th># of Study</th>
<th># of Effect Size</th>
<th>95% C. I.</th>
<th>Mean (d)</th>
<th>Qw</th>
<th>Qb</th>
</tr>
</thead>
<tbody>
<tr>
<td>More than &gt;.7</td>
<td>3</td>
<td>4</td>
<td>0.671 to 0.959</td>
<td>.815</td>
<td>33.04***</td>
<td>41.78**</td>
</tr>
<tr>
<td>Less than &lt;.7</td>
<td>2</td>
<td>3</td>
<td>0.056 to 0.482</td>
<td>.269</td>
<td>4.47(H)</td>
<td></td>
</tr>
<tr>
<td>“No Information”</td>
<td>12</td>
<td>18</td>
<td>0.261 to 0.368</td>
<td>.314</td>
<td>45.80***</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>25</td>
<td>0.056 to 0.959</td>
<td>.369</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The homogeneity test of instrument reliability revealed $Q_b$ of 41.78 and $p < .05$. This suggests that the effect sizes of STEM programs on continuity outcomes were
heterogeneous across the instrument reliability levels. A homogeneity test within the “more than .7 reliability” and “no information” sub-classes revealed $Q_w$ values that exceeded the 95\textsuperscript{th} percentile of the chi-square distribution. There was a high level of variance in the continuity outcomes within each reliability level. In contrast, a homogeneity test within “less than .7 reliability” classification revealed $Q_w$ that did not exceed the 95\textsuperscript{th} percentile of the chi-square distribution. The mean effect size .269 is a typical representation of the continuity outcome of STEM programs within “less than .7” classification.

**Publication Year**

Is there a difference in continuity outcomes of STEM programs when considering the year of publication? To answer this question, this meta-analysis compared effect sizes by publication year. Table 5-28 shows the distribution of effect sizes on the continuity outcome by published year. There were 11 studies published between 2000 and 2010 produced 17 effect sizes, four studies published between 1990 and 1999 produced five effect sizes and two studies published from 1980 to 1989 produced three effect sizes.

The 95\% confidence interval of effect sizes from the studies published from 2000 to 2010 ranged from 0.284 to 0.391 with an average weighted effect size of 0.338, a small, positive effect. In contrast, the 95\% confidence interval of effect sizes from 1990 to 1999 ranged from 0.552 to 0.844 for the students’ continuity with an average weighted effect size of .698, a moderate, positive effect. The effect sizes from 1980 to 1989 ranged from -0.037 to 0.387 with an average weighted effect size of 0.175.
The publication year was not a successful moderator because it did not create a homogeneous within group variance and a heterogeneous between group variance. The homogeneity test of publication years revealed $Q_b$ of 24.02 and $p > .05$. There were no significant differences in effect sizes across the publication year. Homogeneity tests within each decade of 1990-1999 and 2000-2010 revealed $Q_w$ values that exceeded the 95th percentile of the chi-square distribution. There was a high level of variance within the published years. In contrast, the homogeneity test within publication years between 1980 to 1989 revealed $Q_w$ values that did not exceed the 95th percentile of the chi-square distribution, which means the effect sizes were homogeneous.
CHAPTER VI

DISCUSSION

The findings of this meta-analysis provide some practical guidelines for researchers as well as important policy implications. Overall, STEM programs generated moderate and positive effect sizes on engagement, capability and continuity outcomes. However, such average effect sizes do not represent all the STEM programs due to the heterogeneous nature of the results.

6.1 Implication for Research

One of the main purposes of this study was to investigate whether or not research design has had an effect on the overall effect sizes of the three outcome variables: engagement, capability, and continuity. As it turns out, the between-subject design versus within-subject design has had an effect on the engagement outcomes of STEM programs, with between-subject design having produced a weighted mean effect size of .213 and within-subject design having produced a weighted mean effect size of .557. The
between-subject design versus within-subject design has had a significant effect on the capability outcomes of STEM programs, with between-subject design having generated a weighted mean effect size of .287 and within-subject design having produced a weighted mean effect size of .662. Apparently, individual differences in capability across experimental group and control group weaken the effect sizes. Within-subject design measures the same participants both before and after the implementation of the STEM programs. Such repeated measures effectively eliminated the potential problems due to pre-existing individual differences that were irrelevant to the STEM programs.

The findings in this study are consistent with several other meta-analyses that demonstrated within-subject designs producing larger effect sizes than between-subject designs (Olian, Schwab, & Haberfeld 2004; Dupaul & Eckert, 1997; Viswesvaran & Ones, 1999). For instance, Dupaul and Eckert (1997) reported that the overall effect sizes for academic and social skills from STEM program were positive, but different results occurred across research designs, with between-subject design at .45 and within-subject design at .64. Similarly, this study found that effect sizes for within-subject design were higher than those for between-subject design on the overall effect results on the capability outcome. However, when comparing STEM programs on the continuity outcomes, there were no significant differences in the effect sizes due to research design.

In a between-subject design, random assignment of participants to the control group and experimental group is desirable but the reality often prevents researchers from doing so due to practical, ethical, and political issues. As the predominant nonequivalent design for research in STEM program evaluation, the non-random assignment may have jeopardized the sample equivalence of some of studies included in this meta-analysis.
Most studies on STEM program effectiveness used non-randomly selected subjects. For instance, many of the research studies with a between subject design relied on the use of intact groups, such as students’ home room or classes they were enrolled, to form the groups.

The finding in this study that within-subject capability effect sizes are larger than the between-subject design might be an individual differences variable. That is, some individuals are more predisposed toward actively learning STEM related skills and knowledge than are others. Especially the majority of studies using between subject designs in this meta-analysis non-randomly selected and assigned subjects to groups. The non-random selection tends to increase the important individual difference between control group and experimental group. In contrast, participants in within subject design are always the same by dealing with the same group of participants in STEM programs.

“Testing” effect might be related to an capability improvement on the post-test due to taking the pre-test. For instance, students learned the types of STEM questions (e.g. Algebra) on the pretest and therefore knew how to do better by posttest. It is possible that within-subjects capability effect sizes are larger than the corresponding between-subject effect sizes. In this case, the larger effect size from within-subject would not be STEM programs that had an effect but the experience of taking the test once that led to the improvement. Future research is needed to address this concern.

Unlike its impact on the capability outcomes, the research design (between-subject design versus within-subject design) did not statistically affect the overall continuity outcomes. Between-subject design generated a weighted mean effect size of .337 and within-subject design generated a weighted mean effect size of .457. The
potential individual differences across the control group and the experimental group did not make a difference in opportunity to pursue STEM program further but they made a difference in the psychological engagement or capability outcomes. STEM programs are evaluated by participants’ capability, studies that used within-subject design shows larger effect sizes than those used between-subject design by ruling out irrelevant individual differences. However, these results should be viewed with caution since the number of studies included was small. Future researchers studying STEM program effectiveness should consider using a within-subject design due to the elimination of individual variances. First, a within subject design is more powerful than a between subject due to the elimination of individual differences. Within subject design dealt with same group of participants and small variance, therefore, the mean effect size in the within-subject design was more likely to be higher than the mean effect size in studies with the between subject designs. Second, most of the studies with between-subject designs in STEM program evaluations were conducted with non-randomly assigned subjects. The non-randomization of group assignment increases the risk of unobserved differences between individuals within the experimental group and control group, which will cause variation in the outcome variable. If researchers have to choose a between subject design for reasons of various practical realities such as limited time, future research utilizing a between subject design should randomly select the control group in a research design to minimize pre-existing differences between control group and experimental group.

Although a within subject design is desirable for a researcher to evaluate STEM program effectiveness, an ideal research design is a combination of a within-subject design and a between-subject design in a single study. Utilizing a between subject design
with a pretest and a posttest allows future researcher to identify the magnitude of initial group differences. Most of the studies using between-subject design in this meta-analysis selected non-equivalent groups, where participant characteristics may not be balanced equally among the control and experiment group. But, when future researchers use the between-subject design with a pretest and a posttest, researchers can identify the differences of pretest scores between a control group and an experimental group before the STEM programs are implemented. Smaller differences in pretest scores indicate that smaller differences may exist between two groups. Additionally, comparing test scores between each group’s pretest and posttest helps researchers analyze data and interpret the results.

6. 2 Implication for STEM Education Policy

The results of this study have important policy implications for developing successful STEM programs.

6.2.1 Policy Implication related to the program strategy

STEM program strategy had a significant effect on the capability of students and teachers, and continuity of students. If decision makers aim to achieve the most effective STEM programs, according to the results of this study, they should first continue to make use of the following pedagogical strategies: the Inquiry, Enhanced Context, and Mixed strategies. First, the Inquiry Strategy had a strong effect size on the capability outcome.
This Inquiry Strategy states that students should be involved in the asking and solving of questions in science lessons (Chiappetta, 1997). This inquiry-based program emphasizes an understanding of the nature of science and science as a process (Anthony, 1973; Bruner, 1961). For example, the students in the more inquiry-based activity are not told what the expected results are. Students are simply told to investigate how global warming affects plant growth. In other words, teachers should facilitate students’ own investigations by encouraging them to question, predict, collect and analyze data, and explain the results, rather than providing answers for students. In a non-inquiry experiment, students might simply be told that “higher temperatures caused by global warming will cause plants to photosynthesize more” (Kirschner, Sweller and Clark; 2006.) Teachers should make learning relevant by encouraging students to answer scientific research questions through such means as the conduct of laboratory exercises.

Another finding was that the Enhanced Context Strategy also showed a high effect on the capability outcome of students and teachers in STEM areas. This result suggests that experiential learning can play an important role as one of the successful tools for obtaining knowledge in the STEM fields. The Enhanced Context Strategy engages students’ interest by relating learning to the student’s or school’s environment or experiential learning (Boud, Keogh, & Walker, 1985; Kolb & Fry, 1975). If students are placed in an environment in which they can actively connect the instruction to their experience, their capabilities will accelerate.

Some examples of experience-based approaches used in STEM programs include: problem-based learning, taking field trips and using the schoolyard for lesson, and experiencing research internships. To learn scientific knowledge such as paleontology,
visiting a museum and examining fossil is an effective approach. Also, research internships provided to college students in STEM areas should be encouraged in the United States. These results are consistent with a previous finding about the same strategies. Schroeder, Scott, et al (2007) reported that the Enhanced Context Strategy had the highest effect size (1.480) on the students’ achievement.

Although the Inquiry and Enhanced Context Strategies had relatively high effect sizes for capability outcome, the strategies had relatively low effect sizes for continuity. This result is supported by Bronford, Brown, & Cocking (2000), who suggested that the educator must select from among the various strategies to accomplish the particular goal for learning. On the other hand, the Mixed Strategy approach, a combination of different strategies, provided the largest effect size on the continuity outcomes. Wise (1996) asserts that no single strategy is as powerful as an array of them.

### 6.2.2 Policy Implication related to the program creator

STEM programs created by professional associations, such as American Mathematical Association, National Science Teacher Association, and American Association of the Advancement of Science, are the most successful in engagement and capability measurements. These results suggest that professional associations participate in creating STEM programs to achieve the greatest effect. However, governments, such as the U.S. Department of Education, produced far less successful results. Dierking (2010) states that because professional associations better understand and support STEM learning, they are able to create new programs representing how to transform education practice. Cavanaugh, Gillan, Kromrey, Hess, and Blomeyer (2004) suggest that educators
as a professional group involved in science education should develop and articulate K-12 science and technology education programs.

Also, this study found that STEM programs created by cooperative groups had the second largest effect sizes on the capability of students and teachers. Cooperative group refers to a combination of two or more organizations or groups, such as professional associations, non-profit organizations, governments, K-12 school districts, or teacher groups. The result suggests that cooperation among the groups is important to create effective STEM programs. Zhang, McInerney, and Frechtling (2010) pointed out that the idea of partnership includes not only among institutions of higher education and K-12 school districts, but also among the STEM faculty and other project participants. For instance, the Merck Institute for Science Education (MISE) program has collaborated with local school districts, superintendents, and educators to support the STEM education objectives and to create effective science programs. However, there is no empirical study on the effectiveness of programs by different creators to compare with the results of this study.

6.2.3 Policy Implication related to the grade level

The effect of education programs on the engagement and capability may vary by participants’ grade levels (Duncan, Brooks-Gunn, & Klebenov, 1994; Lerner, 1991). According to this study, the largest mean effect size on the capability outcome was attributed to elementary and middle school students at .889 and .880 respectively. These findings are consistent with several other studies (Becker & Park; 2011, Sander; 2009; Tekbiyik and Akdeniz 2010). Tekbiyik and Akdeniz (2010) found that the most effective
grade level in science education programs was the elementary level. The most ineffective was the high school level.

The result suggests that STEM program funders including governments and school districts in the United States should invest in younger learners such as elementary and middle school students over adult learners such as college students, to more effectively obtain higher STEM skills and knowledge provided by STEM programs. This study supports Sanders’ claim (2009) that:

Elementary grades are absolutely the place to begin these programs. If America hopes to effectively address the ‘STEM pipeline’ program, we must find ways of developing young learners’ interest early on in the educational process (p.35).

The total number of effect sizes of programs focusing on the college students was 52 out of 153 effect sizes in capability outcome. American governments have been increased funds to STEM programs to increase college students’ interests of STEM areas over the past decades (National Science Board, 2006). To obtain a higher return for the federal funds, the government should concentrate on increasing the interest of young learners in STEM fields. The governments should provide more opportunities for elementary students to participate in STEM programs in order to improve children’s interest and knowledge throughout early childhood.

According to this meta-analysis, the most effective STEM programs had several key attributes. First, the programs designed by professional association organizations had higher effects related to the engagement and capability outcomes. This seemingly contradicts the fact that the majority of STEM programs observed in this study were created by non-profit organizations, and professional associations created a relatively low
number of STEM programs. Second, the most effective programs used a student-centered inquiry and experiential-based strategy to improve knowledge and skills in STEM areas. These strategies were also the most frequently used, and thus, teachers should continue to use them because of their efficacy in the capability outcome. But, on the other hand, a mixed strategy was the most strategy in terms of STEM program engagement. Third, the most successful STEM programs with respect to student capability targeted elementary and middle school students rather than high-school and college students. Thus, if the goal is to increase student capability, more resources should be invested in programs from elementary and middle school students to achieve the best results, or alternative strategies should be attempted in the future to achieve better capability results with respect to high school and college students. But, despite these results above, any programs that targeted specific groups did not make a difference on the continuity outcome.

To evaluate future STEM programs, researchers should consider using a combination of between-subject and with-subject design. All of the studies on STEM program effectiveness utilized within this meta-analysis have used either within-subject design or between-subject design, but not both. The combination of both research designs allows researchers to identify the difference of pretest scores between a control group and an experimental group with the pretest

6. 3 Limitations of This Study

The majority of research studies on STEM program effectiveness failed to report some pertinent information about each program and participant characteristics. A meta-
analysis cannot improve the quality or reporting of the original studies. In other words, more specific information about the funding source and gender ratio needs to be collected in order to clearly analyze the overall mean effect size. Especially, a lack of information about the funding source and an inconsistent measurement about socio-economic status and participant ability level from the existing studies restricted the meta-analysis of these moderator variables in this study.

The following sub-categories missed information in the studies of STEM program evaluations:

- Despite findings that the funding source variable was statistically significant for students’ and teachers’ enhancement, information about the funding source has not been reported in many previous studies on STEM effectiveness. Out of the 89 studies, the 33 studies that did not provide information about the funding source which had a mean effect size of .857, a high effect size.

- Although a majority of the studies provided information on socioeconomic status of their participants, they did not allow for consistent categorization in the meta-analysis: percent of students receiving subsidized lunch (Bachman, Bischoff, &Gallagher; 2008, Lambert & Whelan Ariza;2008), participants’ ethnicities (e.g. Know, Moynihan, & Markowitz; 2003, Born, Revelle, & Pinto; 2002; Rullilove & Treisman; 1990 ), measuring parent income levels (e.g. Kaplan & Black; 2003, Krajcik, McNeill, & Reiser; 2007), and education levels (e.g. Hughes; 2000) . In addition, information
about gender ratio was not sufficient to conduct the meta-analysis in this study.

- In methodology characteristics, most studies did not use randomly selected subjects (only 15% of the total effect sizes collected). Many research studies relied on the use of intact groups, such as students’ home room or classes they enrolled, to form the groups.

To date, the majority of studies on STEM program effectiveness continue to be qualitative in nature. Much STEM program research remains in the form of case studies, opinion papers and “how to” articles (Merisotis, 1999). Meta-analysis is a useful method for finding all effective factors influencing STEM performance in the United States by synthesizing rigorous studies of program effectiveness. However, a meta-analysis can certainly benefit from a more uniform way of reporting STEM programs’ effectiveness from individual studies.

6.4 Conclusion

The purpose of the current study was to investigate engagement, capability, and continuity outcomes of STEM education programs by conducting a comprehensive meta-analysis. This study found that all three outcome variables had positive effect sizes at the moderate level of effect sizes: the weighted mean effect sizes were .346 for engagement, .454 for capability, and .369 for continuity measurements. The capability outcome had the highest effect among the three outcome variables. The continuity outcome was slightly higher than the engagement outcome, which had the lowest mean
effect size of the three measurements. Thus on the basis of the 91 studies evaluated herein, the evidence indicates that investments in STEM programs tend to improve engagement, capability and continuity of students and teachers within the fields.

At the same time, estimates of overall STEM program effectiveness partially depend on research design and some moderator variables. Although the research design variable did not have an effect on results of the continuity outcomes, it was significant with the engagement and capability outcomes. The studies using a within subject design had much higher effect sizes on the engagement and capability outcome than those using between subject design. The reason why the research design variable was not related to the continuity outcomes could be possibly explained by its small sample size.

Various moderators of STEM program characteristics were also investigated in terms of whether they were associated with the effectiveness of STEM programs. Program creators made a difference in the engagement and capability outcomes. First, the STEM programs created by professional association organizations as compared to programs created by governments and non-profit-organizations had the highest effect on students’ and teachers’ engagement as well as capability. According to the results of this study, the professional association organizations as the program creators are important for effective improvement of students’ and teachers’ skills and knowledge.

Second, the type of program strategy was also related at a statistically significant level to the capability. Inquiry and enhanced context strategies were successful one for improving skills and knowledge in STEM areas. The Inquiry Strategy is based on analyzing scientific research questions with data. The Enhanced Context Strategy is typical of an activity-based, hands-on approach to science. According to this analysis,
one of the prevailing approaches to develop various capabilities is an Enhanced Context Strategy, which is an experiential learning approach, and includes taking field trips, or participating in internships. The total number of effect sizes of programs using this strategy was 63 out of 153 effect sizes. The Inquiry Strategy, which emphasizes answering scientific research questions by analyzing data (e.g. using laboratory inquiries), is also commonly utilized for improving STEM students’ achievements. The Inquiry Strategy produced 23 effect sizes.

Finally, the participants’ grade level was related at a statistically significant level to the engagement and capability measurements. According to this study, elementary and middle school students had higher average weighed effect sizes than high school and college students on the engagement outcome. Additionally, the largest mean effect size on the capability outcome was attributed to elementary and middle school students. These results suggest that reaching out to student earlier leads to achieving maximum effects in STEM programs. American governments need to continue to provide opportunities for young learners to get interested in STEM areas and improve their skills and knowledge of STEM areas.
6.5 Recommendation for Future Research

Given the abundance of STEM programs currently being implemented, there is a need to continue to empirically evaluate their outcomes and record their results in order to conduct an extended meta-analysis. First, future research needs to continue to focus on the integrated STEM education programs. A concern with the integrative characteristics of STEM education has become more important due to a demand for integrative teaching approaches to draw on the foundation of each of the constituent disciplines to form a cohesive overall course of instruction (Hays, Blaine, & Lantz, 2009; Morrison, 2006; NHSA, 2008). If there are enough studies on the effectiveness of integrated STEM programs, a future meta-analysis can be conducted of the integrated STEM programs, furthermore, as well as by field combination, such as science and technology, or science and mathematics.

In addition, a future study could productively investigate whether there is a statistically significant gender difference between males and females in regards to STEM program effectiveness. Recent studies suggest that the gender gap is narrowing (Greenfield, 1996; Jovanovic and Dreves, 1998; Riesz et al., 1997). However, there are other studies that suggest that males have had more positive attitudes towards science than have females (Catsambis, 1995; O’Brien et al., 1999; Simpson and Oliver, 1985; Weinburgh, 1995). Additional studies targeting STEM education in all school levels will enable future meta-analyses of this literature to have a stronger literature base, providing more effect sizes for analysis.
Finally, future research needs to concentrate on a continuity measurement which assesses the extent to which STEM education leads to STEM careers. Many employers in STEM fields require continuing education credits as a way to ensure that employees obtain and maintain knowledge and skills related to STEM (STEM Education Coalition, 2008;). STEM education in the United States is responsible for providing the STEM workforce (STEM Education Caucus, 2008; Department of Labor; 2007, the Business-Higher Education Forum; 2010). It would be useful for research to investigate whether STEM education programs has an effect on the STEM workforce. Currently, there is limited STEM education research available targeting the continuity between education and workforce.
Reference

Overall studies


National Science Foundation (NSF), Division of Science Resource Statistics (1996).


Randy, J. M. (2002). A synthesis and reflection on the research findings from a statewide undergraduate program to prepare specialist mathematics and science teachers (the Maryland Collaborative for Teacher Preparation), Department of Curriculum & Instruction.


Richardson, W.S., Wilsonm, M.S., Nishikawa, J., Hayward, R.S.A. (1995). The well-


Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude: A psychophysical method and some experiments with a scale for measuring attitude toward the church*. Chicago, IL: The University of Chicago Press.


**Studies for a meta-analysis**


Adamson, S.L., Bank, D., Burtch, M. Frank Cox III, Judson, E., Turley, J.B., Benford, R.


evaluation of admissions applications for a minority focused STEM research Program. *Journal of STEM Education*, 9(1), 40-47.


Meichtry, Y.J. (1992). Influencing student understanding of the nature of science: Data


# Appendix

## Computer searched journals

**Journal of Science Education and Technology**

<table>
<thead>
<tr>
<th>Year</th>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Stage 3</th>
<th>Stage 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006-2010</td>
<td>26</td>
<td>16</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>2001-2005</td>
<td>42</td>
<td>19</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>1996-2000</td>
<td>25</td>
<td>9</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>1992-1995</td>
<td>14</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>107</td>
<td>48</td>
<td>35</td>
<td>11</td>
</tr>
</tbody>
</table>

**Science Education (1)**

<table>
<thead>
<tr>
<th>Year</th>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Stage 3</th>
<th>Stage 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006-2010</td>
<td>19</td>
<td>6</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2001-2005</td>
<td>23</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>1996-2000</td>
<td>18</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>60</td>
<td>14</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>

**Science Education (http://www3.interscience.wiley.com/journal/32122/home) (2)**

<table>
<thead>
<tr>
<th>Year</th>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Stage 3</th>
<th>Stage 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006-2010</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2001-2005</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1996-2000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1991-1995</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1986-1990</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1981-1985</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
### Educational Evaluation and Policy Analysis

<table>
<thead>
<tr>
<th>Year</th>
<th>Total number of titles</th>
<th>Number retained after reading titles</th>
<th>Number retained after reading abstract</th>
<th>Number retained after methodology or result</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006-2010</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2001-2005</td>
<td>18</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1996-2000</td>
<td>18</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1991-1995</td>
<td>28</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1986-1990</td>
<td>28</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1981-1985</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>total</td>
<td>134</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### American Journal of Evaluation(1)

<table>
<thead>
<tr>
<th>Year</th>
<th>Total number of titles</th>
<th>Number retained after reading titles</th>
<th>Number retained after reading abstract</th>
<th>Number retained after methodology or result</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006-2010</td>
<td>61</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2001-2005</td>
<td>80</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1996-2000</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>total</td>
<td>151</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Journal of Negro Education

<table>
<thead>
<tr>
<th>Year</th>
<th>Total number of titles</th>
<th>Number retained after reading titles</th>
<th>Number retained after reading abstract</th>
<th>Number retained after methodology or result</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2001-2005</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1996-2000</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1991-1995</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1986-1990</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1981-1985</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>total</td>
<td>46</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
### Manually searched journals

#### Journal of STEM education

<table>
<thead>
<tr>
<th>Year</th>
<th>Number retained after reading titles</th>
<th>Number retained after reading abstract</th>
<th>methodology or results</th>
<th>Number used in final Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006-2010</td>
<td>14</td>
<td>14</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>2001-2005</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>17</td>
<td>16</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

#### Journal of Elementary Science Education

<table>
<thead>
<tr>
<th>Year</th>
<th>Number retained after reading titles</th>
<th>Number retained after reading abstract</th>
<th>methodology or result</th>
<th>Number used in final Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006-2010</td>
<td>12</td>
<td>9</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>2001-2005</td>
<td>9</td>
<td>9</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>1996-2000</td>
<td>15</td>
<td>9</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>1991-1995</td>
<td>11</td>
<td>9</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>1989-1990</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>51</td>
<td>38</td>
<td>21</td>
<td></td>
</tr>
</tbody>
</table>

#### Journal of Engineering Education

<table>
<thead>
<tr>
<th>Year</th>
<th>Number retained after reading titles</th>
<th>Number retained after reading abstract</th>
<th>methodology or result</th>
<th>Number used in final Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009-2010</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>2007-2008</td>
<td>13</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2005-2006</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>23</td>
<td>10</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

#### Journal of Research in Science Teaching

<table>
<thead>
<tr>
<th>Year</th>
<th>Number retained after reading titles</th>
<th>Number retained after reading abstract</th>
<th>methodology or result</th>
<th>Number used in final Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006-2010</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2001-2005</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>1996-2000</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1986-1990</td>
<td>13</td>
<td>12</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>1981-1985</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>41</td>
<td>38</td>
<td>35</td>
<td></td>
</tr>
</tbody>
</table>