Spring June 1st, 2019

A Meta-Analysis of the Effects of Reflective Self-Assessment on Academic Achievement in Primary and Secondary Populations

Jeffrey James Youde
Seattle Pacific University

Follow this and additional works at: https://digitalcommons.spu.edu/soe_etd

Part of the Curriculum and Instruction Commons, Educational Assessment, Evaluation, and Research Commons, Educational Methods Commons, Elementary Education Commons, and the Secondary Education Commons

Recommended Citation
Youde, Jeffrey James, 'A Meta-Analysis of the Effects of Reflective Self-Assessment on Academic Achievement in Primary and Secondary Populations' (2019). Education Dissertations. 48.
https://digitalcommons.spu.edu/soe_etd/48

This Dissertation is brought to you for free and open access by the Education, School of at Digital Commons @ SPU. It has been accepted for inclusion in Education Dissertations by an authorized administrator of Digital Commons @ SPU.
A Meta-Analysis of the Effects of Reflective Self-Assessment on Academic Achievement in Primary and Secondary Populations

by

Jeffrey J. Youde

Dissertation presented to the Faculty of the Graduate School of Education at Seattle Pacific University in partial fulfillment of the requirements for the degree of Doctor of Education

Seattle Pacific University

2019
A Meta-Analysis of the Effects of Reflective Self-Assessment on Academic Achievement in Primary and Secondary Populations

by

Jeffrey J. Youde

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Education

Seattle Pacific University

2019

Approved by

Arthur Ellis, EdD, Chair, Dissertation Committee

John Bond, EdD, Dissertation Committee

David Denton, EdD, Dissertation Committee

Program Authorized to Offer Degree

SCHOOL OF EDUCATION

Date JUNE 2019

Nyaradzo Mvududzi, EdD, Dean, School of Education
Copyright Page

In presenting this dissertation in partial fulfillment of the requirements for the Doctor of Philosophy in Education degree at Seattle Pacific University, I agree that the library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with “Fair Use” as prescribed in U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to University Microfilms, 1490 Eisenhower Place, PO Box 975, Ann Arbor, MI 48106, to whom the author has granted “the right to reproduce and sell (a) copies of the manuscript in microfilm and/or (b) printed copies of the manuscript from microfilm.”

Signature

Date June 2019

__________________________
# Table of Contents

Abstract .......................................................................................................................... 1

Chapter 1: Introduction ................................................................................................... 2

Purpose of the Study ....................................................................................................... 2

Significance of the Study ............................................................................................... 2

Definitions ....................................................................................................................... 2

   Metacognition ............................................................................................................. 3

   Formative Assessment ............................................................................................... 3

   Reflective Assessment .............................................................................................. 3

Research Questions and Null Hypotheses ..................................................................... 4

Research Questions ....................................................................................................... 4

Null Hypotheses ............................................................................................................. 4

Content of the Following Chapters ............................................................................. 5

Chapter 2: Review of Literature .................................................................................... 6

   Theoretical Underpinnings ....................................................................................... 6

   Empirical Research .................................................................................................. 9

      High School Mathematics .................................................................................... 10

      High School Science ........................................................................................... 10

      Middle School Mathematics .............................................................................. 11

   Conclusion ................................................................................................................. 14

Chapter 3: Research Methods ...................................................................................... 16

   Criticisms and Limitations of Meta-Analysis ......................................................... 16

   Literature Search .................................................................................................. 20
Inclusion Criteria ................................................................. 21
Calculating Effect Size .......................................................... 23
Calculating Meta Analysis Effect Size ...................................... 25
Limitations and Delimitations.................................................. 27
Summary ............................................................................. 28
Chapter 4: Results .................................................................. 29
Study Characteristics .............................................................. 29
Summary Mean Effect ............................................................. 32
Effect of Reflection by Subject Area .......................................... 34
  Effect of Reflection on Math Learning .................................. 34
  Effect of Reflection on Science Learning .............................. 35
  Effect of Reflection on Social Studies Learning ...................... 35
Effect of Reflection by Grade Level ........................................... 36
Effect of Reflection by Publication Type .................................... 38
Publication Bias Analysis .......................................................... 40
  Funnel Plot Analysis .......................................................... 40
  Orwin's Fail-Safe N ........................................................... 41
  Dubal and Tweedie's Trim and Fill ........................................ 42
Chapter 5: Discussion ............................................................... 43
Null Hypothesis Testing ............................................................. 43
  Null Hypothesis 1, Overall Mean Effect .............................. 43
  Null Hypothesis 2, Moderator Effects ................................. 44
  Academic Subject .............................................................. 44
Math Learning ................................................................................................................. 45
Science Learning ............................................................................................................ 45
Social Studies Learning ................................................................................................. 45
Grade Level .................................................................................................................... 46
Publication Type .................................................Error! Bookmark not defined........ 49
Limitations ..................................................................................................................... 49
Recommendations for Future Research ........................................................................ 51
Summary ......................................................................................................................... 53
References ....................................................................................................................... 56
List of Figures

Figure 1. Stem and leaf plot of sample sizes from 19 studies........................................29

Figure 2. Stem and leaf plot of effect sizes from 19 studies...........................................32

Figure 3. Forest plot of observed and mean effects across authors and subjects..............33

Figure 4. Forest plot of observed and mean effects by subject, math...............................34

Figure 5. Forest plot of observed and mean effects by subject, science..........................35

Figure 6. Forest plot of observed and mean effects by subject, social studies...............36

Figure 7. Forest plot of observed and mean effects by level, middle school....................37

Figure 8. Forest plot of observed and mean effects by level, high school.......................38

Figure 9. Forest plot of observed and mean effects across journal-published studies......39

Figure 10. Forest plot of observed and mean effects across unpublished dissertations ...40

Figure 11. Funnel plot of potential publication bias.......................................................41
List of Tables

Table 1. Search Terms by Database.................................................................................. 21
Table 2. Included Studies.................................................................................................................. 23
Acknowledgements

I would like to thank my committee chair and mentor, Dr. Arthur Ellis, for his steadfast advice, encouragement, and expertise during my journey through this doctoral program. His kindness, enthusiasm, and willingness to engage with people, places and ideas both domestic and foreign have served as both an inspiration to me and a template for a life well-lived. I owe a debt of gratitude to my committee members, Dr. John Bond and Dr. David Denton. Both provided extremely useful critical analysis and feedback on various drafts of my dissertation. Other faculty members who contributed importantly to my learning during this program include Dr. Bill Nagy, Dr. Chris Sink, Dr. Andrew Lumpe, and Dr. Greg Fritzberg. Megan Hamshar, Graduate Programs Manager at the SPU School of Education, responded quickly to questions and requests.

I benefitted from a cohort of high quality individuals who made the coursework and projects richer and more enjoyable. I especially appreciate the friendship and collegiality of Dr. Nalline Balriam, Dr. Douglas Asbjornsen, and the fabulous Flores brothers, Dr. Omar Flores and Dr. Homero Flores.

I would like to thank the two people who have shaped me from day one, my parents James and Judith Youde. By believing in me, you continue to inspire me to challenge myself.

Finally, I give a full measure of love and gratitude to my wife of 25 years, Kris. She provided the encouragement for me to undertake this journey. Her willingness to stand by me gave me the emotional and intellectual sustenance to pursue this and other dreams.
Seattle Pacific University

Abstract

A Meta-Analysis of the Effects of Reflective Self-Assessment on Academic Achievement in Primary and Secondary Populations

By Jeffrey J. Youde

Chairperson of the Dissertation Committee: Dr. Arthur K. Ellis

Recent empirical research studies indicate that reflective self-assessment as a classroom approach can have a positive impact on student achievement. Reflective self-assessment, a form of metacognition, allows a student to think about past, current, and future learning performance. Although several discrete empirical studies have supported such hypotheses, a quantified exploration and summary of the relationship between classroom techniques of reflective self-assessment and student academic achievement is needed. The results of the current study, a meta-analysis of surveyed empirical studies from the past 26 years, indicate that reflective self-assessment has an overall effect size of .46 on academic achievement across grade levels and subjects. This effect size is considered moderate. Overall, such findings indicate that an increased use of reflective self-assessment in classrooms may provide students a chance to improve academic achievement.

Keywords: meta-analysis, metacognition, reflection, self-assessment, academic achievement, effect size.
Chapter One: Introduction

Purpose of the Study

Over the past quarter century, several studies have examined the use of reflective self-assessment as a classroom technique employed by teachers and students. The studies have involved a variety of grade levels and academic subjects. At this time, a quantified synthesis of such studies is needed, and a meta-analytic method is an appropriate approach through which to examine the effectiveness of this classroom strategy. Simply put, meta-analysis is “data analysis applied to quantitative summaries of individual experiments” (Glass, McGaw, & Smith, 1981, p. 21). A meta-analysis collects related quantitative studies and combines the findings of these studies into a calculated effect size and variance for a given intervention. This approach helps researchers understand each individual study in the context of similar studies in a way that a narrative literature review does not (Borenstein, Hedges, Higgins, & Rothstein, 2009). The purpose of the current study is to synthesize the findings of related studies with the goal of establishing an overall impact, or effect size, for reflective self-assessment.

Significance of the Study

Current and future teachers have a variety of classroom practices at their disposal, including techniques related to metacognition, formative assessment, and reflective self-assessment. The current study is significant because it contributes to the knowledge base on the use of reflective self-assessment.

Definitions
**Metacognition.** Flavell (1979) defined metacognition as “knowledge and cognition about one’s cognitive phenomena” (p. 906). A more succinct definition might be Costa’s “thinking about thinking” (Costa, 2001, p. 21).

**Formative assessment.** Black and Wiliam (2009) defined formative assessment as:

Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited. (p. 9)

**Reflective assessment.** For Ellis (2001), reflective assessment is “stepping back from what you’re doing in order to achieve some measure of perspective. It means thinking, talking, and otherwise expressing your feelings, the things you’ve learned, the growth you’ve achieved, and the sense you have of accomplishing something” (p. 5). Bond (2003) described reflective assessment as “a form of metacognition using a formative approach that places students at the center of assessment practice” (p. 46). Similarly, Baliram (2016) defined reflective assessment as “active contemplation on the cognitive process of knowledge, skills, situations or experiences with some kind of measurement, typically formative” (p. 13). In a classroom context, metacognition describes the broader category of reflective thought about one’s thinking or learning, while reflective assessment indicates a technique that allows students and teacher to measure or describe their own thought process or learning experiences.
The practice of self-assessment in the classroom, in its simplest form, might be a teacher asking students to complete an “I Learned” statement, which is a strategy for teachers to facilitate student reflection on what has been learned, while finding out if their lesson objectives have been attained. Typically done during the last few minutes of a lesson or activity, students are asked to think about what they have learned during the lesson and then write a sentence or paragraph that begins with the phrase “I learned.” This simply done closure activity facilitates student reflection on their learning and also informs the teacher regarding individual and group progress. (Ellis & Bond, 2016, p. 144)

While other techniques for reflective assessment exist, the studies included in this analysis utilize a mechanism similar to the *I learned* prompt, as well as other similar techniques. It should be noted criteria for studies involving metacognition and self-reflection included in this meta-analysis are detailed in Chapter 3, Research Methods.

**Research Questions**

1. What is the mean summary effect of reflective self-assessment on student individual academic achievement compared to instruction that does not include such metacognition?

2. To what extent is this mean effect, if any, moderated by subject, grade level, or publication type?

**Null Hypotheses**

The null hypotheses derived from the above research questions include:
H₀: There is no statistically significant difference in the effects of reflective self-assessment on academic achievement compared to instruction that does not include such metacognition.

H₀: There are no statistically significant differences in the effects of reflective self-assessment on academic achievement as moderated by subject, grade level, or publication type.

For both of the null hypotheses, the independent variable is the use of reflective self-assessment, and the dependent variable is academic achievement, as measured by various classroom and content-based assessments.

Content of the Following Chapters

The remainder of this dissertation is divided into four chapters, titled Review of Literature, Research Methods, Results, and Discussion. The Literature Review includes theoretical underpinnings of the construct, and a selection of representative empirical studies. The Research Methods chapter includes research design, criteria for including or excluding studies in the meta-analysis, methodology, and data analysis. The Results chapter presents descriptive statistics and results related to the research questions and hypotheses. In the Discussion chapter, the results are evaluated in relation to the research questions and hypotheses.
Chapter Two: Review of Literature

Theoretical Underpinnings

The construct of reflective assessment originates with theorists, such as Rousseau, Dewey, Piaget, Vygotsky, Bandura, and Habermas. Others who have contextualized the practice of reflective assessment include Pintrich, Flavell, Black and Wiliam, and Ellis.

In his 1762 classic on educational philosophy, *Emile*, Rousseau emphasized that the individual learner must grow to understand the world in his or her own way. Rather than relying on external motivation and direction from a teacher, a student should be self-directed and intrinsically motivated. Thus, a student-centered methodology makes the most sense, including discovery learning (Rousseau, 1762/2004). Rousseau’s perspectives on the centrality of the individual learner place him squarely in the progressivist camp, and presage who is considered the father of progressive education, Dewey.

John Dewey’s *Experience and Education* (1938) emphasized the importance of each individual learner, noting that there will never be one singular learning experience that will be useful for all students. An individual learner works to find a solution to a new problem, while bringing her or his past experiences to bear. Reflection gives students a chance as individual learners to connect past learnings with the current problem at hand (Dewey, 1938). In *How We Think*, Dewey (1933) described student reflection as “active persistent and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it and the further conclusion to which it tends” (p. 212).

Piaget’s (1976) stage learning theorizes that intellectual development occurs in specific stages. Knowledge is organized and adapted in relation to prior knowledge, and
reflection allows a student to refine and strengthen understanding of concepts encountered during instruction (Piaget, 1976). Piaget’s theory encourages peer-to-peer interaction and collaboration, and places value on student-centered pedagogical approaches, such as discovery learning.

Lev Vygotsky, Albert Bandura, and Jürgen Habermas’ each proposed theories regarding the central role that student reflection plays in learning. Vygotsky’s (1978) socio-cultural theory encourages social interaction during the learning process. As students interact with one another and with their teacher, and reflect on their experiences, individual learning is clarified (Vygotsky, 1978). Bandura’s (1977) social learning theory argues that the most powerful learning takes place in social settings, when students learn by observing and interacting with others. For Bandura, learning should be self-directed and self-reflective (Bandura, 1977). Philosopher Jürgen Habermas’ emancipatory domain suggests the importance of reflection, as “knowledge is gained by self-emancipation through reflection leading to a transformed consciousness” (Habermas, 1968/1971, p. 78). These theorists’ philosophies, while diverse in their approaches, do share a common progressive theme: the learner as an individual in a social context, and reflection as a way for a learner to individualize his or her own learning.

Some more recent researchers have developed contexts for explaining, and applying, metacognitive practices for use in classroom settings, including Pintrich, Flavell, Black and William, and Ellis.

Pintrich (2002) identified three types of metacognition: student knowledge of learning and thinking strategies, knowledge of cognitive tasks and the tools to use with
them, and knowledge of self, including the strengths and weaknesses of each strategy (Pintrich, 2002).

Flavell (1979), called the father of metacognition and an influencer of Pintrich’s work, identified three knowledge types related to metacognition as well: person, task, and strategy. Person, what you believe about yourself and others; task, information you possess that allows you to complete a cognitive task; and strategy, knowing which thinking strategies will allow you to complete a task (Flavell, 1979). Flavell and Pintrich’s models of metacognition are roughly equivalent: self-knowledge or belief (Who am I?); task knowledge or belief (What do I know?); and strategies (How should I approach a problem?).

Black and Wiliam (2009) categorized metacognitive practices in classroom settings as one type of formative assessment. Formative assessment in a classroom occurs when:

Evidence about student achievement is elicited, interpreted, and used by teacher, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited (Black & Wiliam, 2009 p. 5).

Formative assessment may take two forms: synchronous, allowing a teacher to make real-time adjustment to a lesson; or asynchronous, which can include a teacher using “evidence derived from homework, or from students’ own summaries made at the end of a lesson (e.g., ‘exit passes’) to plan a subsequent lesson” (Black & Wiliam, 2009, p. 5).
For Ellis (2001), an example of an asynchronous formative assessment is the classroom practice which uses an “I learned…” or other prompt. At the end of a class session, students respond to the prompt, “I learned…” giving the teacher immediate end-of-lesson diagnostic information about what students learned, what they think they learned, misperceptions, and other formative feedback. These student responses to such prompts can help a teacher to evaluate: how effective the current lesson was (or wasn’t); what the class should potentially emphasize the next day; what re-teaching should be done; and which individual students may need additional support or review (Ellis, 2001).

According to theoretical underpinnings, metacognition via formative, reflective assessment may provide students with opportunities to better understand their own learning. At the same time, metacognitive practices in the classroom may give teachers formative feedback to better understand what was effective or ineffective about a lesson. Overall, metacognition, as described by the preceding theorists, has its roots in the progressivist tradition, and plays a central role in student-centered approaches to learning and teaching (Ellis, 2001; Flavell, 1979; Pintrich, 2002).

**Empirical Research**

The current meta-analysis investigates the impact of reflective self-assessment on student academic achievement. Of the 19 studies selected for inclusion, four of the studies are reviewed in the following section. They provide a representative sample of empirical studies done in the subject areas of math \( (n = 7) \) and science \( (n = 5) \), and include middle school \( (n = 9) \) and high school \( (n = 9) \) grade levels. All studies have in common the classroom use of formative self-reflection techniques like the “I Learned” statement. Research designs include both true experimental (random assignment) as well
as quasi-experimental (intact groups) designs; pre-test-treatment-post-test as well as post-test-only procedures; student subject grade levels from elementary to high school; and academic subject areas including math, science, language arts, foreign language, and social studies. Within the range of subject areas and student ages, all employ a similar intervention: students are asked to write a brief self-reflection at a lesson’s conclusion.

**High school mathematics.** Baliram (2016) examined the effects of reflective assessment on geometry achievement of high school students. The researcher used a quasi-experimental, non-equivalent control group design, in which a convenience sample of 75 ninth and tenth grade students in five intact classes were randomly assigned, two to the experimental group and three to the control group. All groups received instruction in the Pythagorean Theorem and special triangles over four weeks. The experimental group received metacognition training at the end of each daily lesson, while the control group reviewed or began on homework. The metacognition training included “I Learned” statements (Ellis, 2001), Strategic Questioning (Mevarech & Kramarski, 1997) and Clear and Unclear Windows (Ellis, 2001). Student responses to metacognitive prompts were collected by the teacher, and reviewed by the investigator, who provided feedback to students based on what they had written. Both groups received a pre-test, post-test, and a retention test four weeks after completing the geometry lessons. Statistically significant scores on the post-test supported the hypothesis that reflection does increase academic achievement in geometry classes. The study offers support for the notion that reflection as part of daily geometry instruction can provide benefits to student achievement.

**High school science.** Bianchi (2007) examined the effects of reflective assessment on biology achievement of high school students. The researcher used a quasi-
experimental, pre-test/post-test control group design, in which a convenience sample of 126 students in four intact classrooms were randomly assigned to either the experimental group or the control group. All groups received instruction on the topic of osmosis from the investigator over a three-week period. The experimental group received self-reflective training at the end of each lesson, while the control group received guided practice. Both groups received a pre-test, post-test, and a retention test 14 weeks after completing the osmosis unit. Statistically significant results from a three-way, repeated-measures ANOVA indicated that the Reflective group outperformed the Non-Reflective group ($F = 10.258, p = .002$), on both the post-test and retention test, although only 9% of variance was attributable to Group. These results confirmed the hypothesis that reflection does increase academic achievement, albeit very modestly (9% of variance). The notion that reflection as part of daily biology instruction can provide benefits to student achievement was supported through this study.

**Middle school mathematics.** Bond and Ellis (2013) examined the effects of reflective assessment on the mathematics achievement of suburban fifth and sixth graders. The researchers used a true experimental, post-test only control group design, in which 141 students were randomly assigned to one of three groups. These groups (reflective assessment, non-reflective, and control) each had one of six teachers randomly assigned to instruct them. Each of the two experimental groups received identical instructional lessons during a mathematics unit on statistics, which included an opportunity for the first group to practice reflective self-assessment at the conclusion of each daily lesson. These reflections consisted of completing “I Learned” statements in writing, as well as verbal “thinking aloud” tasks (Ellis, 2001). The second group, instead
of reflecting, spent that end of lesson time reviewing the lesson activities and objective, while the control group was taught lessons as part of a separate geometry unit. Results supported the hypothesis that reflection does increase academic achievement. A one-way ANOVA indicated a statistically significant ($p < .05$) main effect size of .273, suggesting that 27% of the variance in achievement between groups could be accounted for by the use of reflection. Results from a retention test six weeks later indicated the group using reflective techniques scored statistically significantly higher than the non-reflective group. However, there was no statistically significant difference between the two groups’ post-test and retention test scores. The study offers support for the notion that reflection as part of daily mathematics instruction can provide benefits to student achievement.

Johnson (2004) examined the effects of reflective self-assessment on the mathematics achievement of 74 fourth, fifth, and sixth graders, using an experimental, post-test only control group design, and adding a covariate (STAR Math grade equivalency scores used to adjust disparities over groups). Each of the two experimental groups received identical instructional lessons during a mathematics unit on statistics, which included an opportunity for the first group to practice reflective self-assessment at the conclusion of each daily lesson. The second group spent end-of-lesson time reviewing, while the control group received geometry instruction. The results of the study reported a non-statistically significant difference between experimental and control groups. The study, according to Johnson (2004), may not have shown an increase in student achievement though these practices, but they were not shown to be ineffective either…reflective practices did not affect student achievement in a positive or negative way. The results only indicate
that in this particular study the implementation of reflective assessment strategies
did not improve student math achievement any more than non-reflective
activities. Students still improved their math achievement scores. (p. 105)

As most studies included in this meta-analysis used a quasi-experimental design,
it is important to note the potential threats to validity. Campbell and Stanley (1963) listed
the following threats to internal validity (i.e., factors other than the treatment that may
impact study outcomes) in pre-test-treatment-post-test designs:

a. History - events other than treatment that affected subjects after the pre-test and before
the post-test

b. Maturation - a subject naturally grows and changes during a study, with such changes
unrelated to the treatment

c. Testing - the act of administering a pre-test to a subject changes their response to the
post-test in some way

d. Instrumentation - the instrument used in the post-test changes in some way to invalidate
the results

e. Regression - a high pre-test score tends to regress toward mean scores on the post-test

f. Experimenter bias - conscious or unconscious expectations impacting effects, especially
when the primary investigator acts as the classroom teacher implementing the treatment

g. Attrition/mortality - subjects may drop out of the study before the post-test is
administered

The use of a post-test-only design generally minimizes several of these internal
threats, including history, maturation, testing, and regression (Campbell & Stanley,
1963). Each of the studies included were reviewed for these potential threats to validity,
and were examined for steps or efforts that each researcher took to address or minimize these threats. If it appeared that threats to validity were either unaddressed, or were severe enough to call internal or external validity into question, that study was excluded from the meta-analytic synthesis.

**Conclusion**

Educational theorists have long placed emphasis on the learner as individual. Rousseau (1762/2004) emphasized the importance of a learner developing an individualized understanding of the world. Dewey (1938) and Bandura (1977) viewed a personalized learning experience as more effective than a one-size-fits-all approach, and that if a student struggled with the subject matter, new learning might be connected to prior knowledge and understanding. Formative self-reflection may satisfy both these theorists, as students who are asked to reflect on a prior lesson must struggle to incorporate new learning experiences with prior understanding. For Habermas (1971), this reflection is a path to increased knowledge and a transformed consciousness. Flavell (1979) and Pintrich (2002) classified reflection as a form of metacognition, or thinking about one’s own thinking. Black and William (2009) viewed these metacognitive classroom practices as a type of formative assessment, measuring student understanding at a lesson’s midpoint or endpoint. Ellis (2001) applied end-of-lesson formative assessment in the form of student self-reflection via an *I learned* statement.

Educational researchers are interested in the impacts of formative assessment via self-reflection on student learning and achievement. Experimental and quasi-experimental quantitative studies, conducted with elementary, middle, and high school students, in various subject areas, measure the impact of this practice. These experiments
comprise a growing knowledge base of how the theories related to personalized learning, metacognitive practices, and student reflection may be applied to positive effect. The use of these reflective prompts can be linked to improved student performance in several academic areas.

The theoretical underpinnings provide strong reasons to value student reflection. A review of empirical research on metacognitive practices using student self-reflection in various classroom settings indicates that there is support for the use of end-of-lesson reflective prompts.

Beyond the theoretical and empirical considerations, the practice is relatively simple to implement, as it requires very little in terms of classroom time or resources. Given the expectations that educators do more with less, efficient and effective educational practices need to be identified, understood, and implemented effectively. This meta-analysis hopes to increase understanding of the impact of formative self-reflection.
Chapter 3: Research Methods

Criticisms and Limitations of Meta-Analysis

Meta-analysis is a form of research synthesis where an investigator searches for, collects, and synthesizes quantitative research on a topic. By synthesizing the experimental research on the impact that reflective self-assessment has on student achievement, broader conclusions can be drawn. Rosenthal and DiMatteo (2001) pointed out that a well-designed and executed meta-analysis can provide insight into the impact that a treatment has on a sampled population. Specifically, the present study seeks to quantify and calculate an overall effect size for a collection of related empirical research studies on reflective self-assessment.

The history of research synthesis dates back to at least 1904, when Karl Pearson collected data from 11 studies on typhoid vaccine. Pearson calculated a correlation coefficient for each study, averaged these treatment effect measurements, and drew conclusions about the effectiveness of the vaccines (Pearson, 1904). Three decades later, Ronald Fisher argued that by combining the $p$ values from independent tests of the same hypothesis, statistically significant results could be obtained (Fisher, 1932).

In the 1970s, Glass and Smith (1979) examined the relationship between class size and academic achievement, which included 725 studies with approximately 900,000 students. Glass coined the term “meta-analysis” to mean statistically integrating the findings of individual studies to yield a summary of the effect.

More recently, John Hattie used meta-analysis to examine the effectiveness of various classroom practices. In his 2012 book “Visible Learning for Teachers,” Hattie included over 900 meta-analyses, which summarized the overall effect sizes and ranked
150 classroom influences on student achievement (Hattie, 2012). Practically speaking, Hattie’s meta-analytic research provides a cost-benefit analysis of classroom influences. Educators may be interested in identifying the classroom practice that shows the most potential for increased student academic achievement for a given outlay of resources. A less resource-heavy practice with a lower effect size may even be preferable to one with a greater effect size at a greater cost. Hattie gives the example of an often-discussed intervention: reduced class sizes. His research shows that reducing a class size from 25-30 students to a class size of 15-20 students has an effect size of .22. His research also shows that implementing a program to prepare students to take tests has an effect size of .27 (Hattie, 2012). Since a test-preparation intervention requires a much lower outlay in funds than reducing class sizes, the lower-cost intervention may provide sufficient benefit to students, and therefore may be the preferred policy (Hattie, 2012, p. 14). Hattie’s work is germane to the current study, since formative self-reflection could be considered to be a low-cost intervention, in terms of both time required and money expended.

Meta-analysis as a research tool provides advantages, including the building of a statistical model which samples a much larger population than could be included in an individual experiment (Field & Gillett, 2010). A meta-analysis has at least two advantages over a typical literature review. First, when a researcher locates and collects studies for a meta-analysis, both published and unpublished research can be included, thus yielding a fuller picture of the impacts of a particular treatment or intervention. Secondly, a researcher conducting a literature review may be biased to include studies that support a specific theoretical position or outlook, and thus again a meta-analytic approach has the potential to yield a less-biased view (Rosenthal & DiMatteo, 2001).
Meta-analyses are not without criticism. Theoretically, a post-positivist or constructivist position would argue against the reductive nature of quantitative research overall. Quantitative research has epistemological origins in behaviorists like B. F. Skinner, and uses experimental method to quantify phenomena (Skinner, 1976). A post-positivist or constructivist might criticize this approach as reductionist, as it attempts to build knowledge using true experiments outside of social contexts, including educational institutions. John Creswell, a proponent of mixed-methods research designs, argued that knowledge gained via experimental studies outside of real-world contexts may lack applicability to real-world situations, like schools. If an experiment randomly assigns subjects to treatment and control groups, such a study is lacking ecological validity. That is, the test may in fact accurately measure a treatment in a controlled, randomized setting, but such a setting would likely never be encountered in a real-life school. Therefore, Creswell argued, the usefulness and import of experimental studies can be overstated when applied to classroom settings (Creswell, 2003).

Glass, McGraw, and Smith (1981) classified common criticisms of meta-analysis into four general categories:

1. “Apples and oranges” (combining studies with different measurements, samples, and interventions makes no sense);

2. Suspect data from subpar studies (using data from poorly-designed or executed studies calls results into question, essentially a “garbage in, garbage out” metaphor);

3. Selection and publication bias (differences between findings published in journals and those studies, such as dissertations, that go unpublished);
4. “Lumpy” data (non-independent data leading to reduced reliability of results) (Glass et al., 1981, p. 218). While each of these criticisms has some validity, Glass et al. (1981) offered critical commentary on each.

The first category of criticism is when a researcher includes different types of studies (apples and oranges) in a meta-analysis. A critic might argue that only research studies having similar research designs should be included in a meta-analysis. If only studies with very similar research designs and populations were included in a synthesis, the results could be expected to be very similar. Thus, there would be no need to combine and integrate findings from identical studies (Glass et al., 1981, p. 219).

The second category of criticism involves the use of suspect data from subpar studies. A critic might point out that most meta-analyses rely on quantity of data, rather than seeking studies with high methodological quality. However, there is evidence of very little difference between the findings when so called “high quality” and “low quality” studies are compared. The authors found that “there is seldom much more than .1 standard deviation difference between average effects for high-validity and low-validity experiments” (Glass et al., 1981, p. 226).

The third category of criticism involves selection bias and publication bias. Reviewers of published research are more likely to include studies with statistically significant findings. A meta-analytic review which only considers published research will most likely be skewed by over-estimating effect sizes. The meta-analytic researcher conducting a search which includes published as well as unpublished studies on a given topic can address publication bias (Glass et al., 1981, p. 227). However, if a meta-analysis includes published studies with statistically significant results, as well as
unpublished dissertations with and without statistically significant results, a researcher still can fail to locate unpublished studies that academics choose to toss in a file drawer (hence the term “file drawer” problem) because they don’t show statistically significant results (Slavin, 2008).

The fourth category of criticism focuses on reduced reliability due to non-independent data. Of the four criticisms, the authors considered this to have the most validity. This case would occur if multiple results are made from subjects of a study, thus rendering the data non-independent, and potentially skewing the results. One solution is to average the findings within each study, and then to conduct the meta-analytic calculations with “studies” as the unit of comparison, rather than using “within-studies” findings (Glass et al., 1981, p. 229).

These and other criticisms can be addressed and minimized by conducting meta-analyses with robust design and implementation. According to Field and Gillet (2010), a properly conducted meta-analytic process has six steps: 1. Conduct a literature search; 2. Choose and apply inclusion criteria; 3. Calculate effect sizes for each included study; 4. Calculate meta-analysis effect size; 5. Do additional analysis (publication bias, moderator variables, etc.); and 6. Write up the results (Field & Gillet, 2010, p. 666). The current study follows this six-step process for meta-analysis as outlined by Field and Gillet (2010). Steps one through five will be described in the remaining part of Chapter 3, and step 6 is presented in Chapter 4, Results.

**Literature Search**

The purpose of this current literature search was to locate relevant literature, both published and unpublished (grey) studies, on the topic of reflective assessment. Five
electronic databases were searched for published studies, including Academic Search Complete, Education Source, ERIC, PsychINFO, and ProQuest Dissertations & Theses. The literature search included the search terms reflective assessment, metacognition, and self-regulated learning. Within the databases, additional selections narrowed the results further. See Table 1 for a list of databases and search terms, as well as within-database selections to limit the results. The basic search terms returned 28,412 results, and the selections within database narrowed the field to 245 publications.

Table 1

*Search Terms by Database*

<table>
<thead>
<tr>
<th>Database</th>
<th>Exact Terms and Phrases</th>
<th>Selections within Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic Search Complete</td>
<td>reflective assessment or</td>
<td>1991-2017</td>
</tr>
<tr>
<td>Education Source</td>
<td>metacognition or</td>
<td>academic achievement</td>
</tr>
<tr>
<td>ERIC</td>
<td>self-regulated learning</td>
<td>self-evaluation</td>
</tr>
<tr>
<td>PsychINFO</td>
<td></td>
<td>junior high students</td>
</tr>
<tr>
<td>ProQuest Dissertation &amp; Theses</td>
<td></td>
<td>middle school students</td>
</tr>
<tr>
<td></td>
<td></td>
<td>secondary education</td>
</tr>
<tr>
<td></td>
<td></td>
<td>elementary school students</td>
</tr>
<tr>
<td></td>
<td></td>
<td>high school students</td>
</tr>
</tbody>
</table>

*Note.* All database searches were limited to 1991-2017, which reduced the scope of this search to a quarter century to provide temporal continuity to the student subjects of the studies.

The investigator examined each located study to determine its suitability for inclusion in this meta-analysis. Other studies were located by examining the reference list and literature search sections of related studies.

**Inclusion Criteria**

The search terms yielded a preliminary pool of 245 studies. From this pool, the studies were examined and screened to limit the included studies. The current meta-
analytic review was limited to quasi-experimental or experimental English language studies that employed control or comparison groups (comparing students that utilized self-reflection to those which did not), and that included a quantitative measurement of the impact of self-reflection on student achievement. Population groups were limited to students in school grades Kindergarten through twelfth grade. Only studies published from 1991 to 2017 were included in this meta-analysis. As noted above, publication bias in social science research has the potential to skew a meta-analysis, as peer-reviewed journals are more likely to publish studies with statistically significant results, and such journals are more likely to publish studies in which the null hypothesis was rejected (Glass et al., 1981). To reduce the risk of such bias, unpublished doctoral dissertations were included in this meta-analysis. Further discussion of the threats of publication bias is detailed in Chapter 4 and Chapter 5.

Glass et al. (1981) pointed out that there is no such thing as a perfect literature search (p. 63), in that somewhere beyond where the researcher looks, there exists one or more studies that should have been included in the meta-analysis. Given the search criteria, the researcher’s reasonable thoroughness during the search and selection phase yielded a total of 19 studies that met the above inclusion criteria. The investigator coded key metrics from each study, as well as covariates, which included grade level, academic subject, and publication type. Table 2 below lists the included studies, including author(s), date, sample size, subject, grade level, and treatment duration.
Table 2

Included Studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample Size</th>
<th>Subject</th>
<th>Grade Level</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bajar-Sales, Avilla, and Camacho (2015)</td>
<td>60</td>
<td>Science</td>
<td>High school</td>
<td>1-3 weeks</td>
</tr>
<tr>
<td>Baliram (2016)</td>
<td>75</td>
<td>Math</td>
<td>High school</td>
<td>4-7 weeks</td>
</tr>
<tr>
<td>Bianchi (2007)</td>
<td>110</td>
<td>Science</td>
<td>High school</td>
<td>1-3 weeks</td>
</tr>
<tr>
<td>Bond and Ellis (2013)</td>
<td>93</td>
<td>Math</td>
<td>Middle school</td>
<td>4-7 weeks</td>
</tr>
<tr>
<td>Boyer (2015)</td>
<td>114</td>
<td>Social Studies</td>
<td>High school</td>
<td>4-7 weeks</td>
</tr>
<tr>
<td>Denton (2010)</td>
<td>210</td>
<td>Social Studies</td>
<td>Middle school</td>
<td>4-7 weeks</td>
</tr>
<tr>
<td>Denton and Ellis (2017)</td>
<td>46</td>
<td>Social Studies</td>
<td>Middle school</td>
<td>4-7 weeks</td>
</tr>
<tr>
<td>Edwards (2008)</td>
<td>54</td>
<td>Math</td>
<td>High school</td>
<td>8+ weeks</td>
</tr>
<tr>
<td>Evans (2009)</td>
<td>163</td>
<td>ELA</td>
<td>High school</td>
<td>4-7 weeks</td>
</tr>
<tr>
<td>Glaubman and Glaubman (1997)</td>
<td>25</td>
<td>ELA</td>
<td>Elementary</td>
<td>8+ weeks</td>
</tr>
<tr>
<td>Johnson (2004)</td>
<td>46</td>
<td>Math</td>
<td>Middle school</td>
<td>8+ weeks</td>
</tr>
<tr>
<td>King (1991)</td>
<td>30</td>
<td>Other</td>
<td>Middle school</td>
<td>1-3 weeks</td>
</tr>
<tr>
<td>Kourilenko (2013)</td>
<td>57</td>
<td>Foreign Lang.</td>
<td>High school</td>
<td>4-7 weeks</td>
</tr>
<tr>
<td>Mevarech and Kramarski (1997)</td>
<td>122</td>
<td>Math</td>
<td>Middle school</td>
<td>4-7 weeks</td>
</tr>
<tr>
<td>Moore (2010)</td>
<td>49</td>
<td>Science</td>
<td>Middle school</td>
<td>1-3 weeks</td>
</tr>
<tr>
<td>O’Neal (2015)</td>
<td>39</td>
<td>Math</td>
<td>High school</td>
<td>4-7 weeks</td>
</tr>
<tr>
<td>Özsoy and Ataman (2009)</td>
<td>47</td>
<td>Math</td>
<td>Middle school</td>
<td>8+ weeks</td>
</tr>
<tr>
<td>Shoop (2006)</td>
<td>134</td>
<td>Science</td>
<td>High school</td>
<td>1-3 weeks</td>
</tr>
<tr>
<td>Yin et al. (2008)</td>
<td>254</td>
<td>Science</td>
<td>Middle school</td>
<td>8+ weeks</td>
</tr>
</tbody>
</table>

Calculating Effect Sizes

An effect size is a “standardized measure of the magnitude of observed effect” (Fields & Gillett, 2010, p. 668) and gives information about the impact of a treatment in terms of standard deviation units. As a standardized measure, effect sizes from different studies that may have measured different variables can be compared. Common effect measures include Pearson’s r, Cohen’s d, and Hedge’s g.

This study used the software program Comprehensive Meta-Analysis, Version 3 (Biostat, 2015) to analyze the effects of self-reflection on academic achievement. The dependent variable, academic achievement, was measured using a variety of researcher-created and standardized assessments. Scores for control and treatment groups as reported included means, standard deviations, and sample sizes.
Cohen’s $d$, Hedges’ $g$, and Glass’ $\Delta$ are the measures typically used to illustrate group differences. Standardized effect sizes are then calculated by dividing the difference in means by the pooled standard deviation of each condition. To measure a group difference, the mean difference is divided by the combined standard deviation, which yields the effect size (Ferguson, 2009). The following formula is used,

$$ES = \frac{\bar{x}_1 - \bar{x}_2}{SD}$$

with the assignment of $x_1$ and $x_2$ means considered arbitrary. In general practice, desired effects (in this case increased academic achievement) should use positive terms, and undesired effects, negative.

These three measures of group differences do differ in how standard deviation is used in the equation’s denominator. Cohen’s $d$ uses a pooled standard deviation of experimental and control groups. Since both groups are given equal weight in the Cohen’s $d$ formula, differences in group sizes may skew the standard deviation, and thus the effect size. Cohen’s $d$ also has the potential to overestimate the calculated effect size in small samples (Borenstein et al., 2009). To address differences in standard deviation between control and experimental groups, a researcher can use Glass’ $\Delta$. Glass’ $\Delta$ uses the standard deviation of the control group only, based on the idea that the control group standard deviation should be closer to the entire population than the experimental group (Ferguson, 2009). An overestimation bias in small samples can be addressed by using Hedges’ $g$, which yields a less-biased estimate by using a pooled and weighted standard deviation in the formula’s denominator (Borenstein et al., 2009, p. 27).

$$\text{Hedges' } g = \frac{M_1 - M_2}{SD_{pooled}}$$
Of the 19 studies included in this meta-analysis, all reported measures of group differences using mean, standard deviation, and sample size for treatment and control groups. This investigator obtained the results from each study, and calculated the effect size for each, using Hedges’ $g$.

**Calculating Meta-Analysis Effect Size**

Once a common effect size is calculated for each of the selected studies, the next step is to calculate a combined effect size for all studies. Before this is calculated, the researcher conducting the meta-analysis must view the data through the lens of either a fixed-effects model, or a random-effects model. The researcher makes this determination based on assumptions that can be made about the population, sampling, study characteristics, and overall conclusions that hope to be drawn (Borenstein et al., 2009).

A fixed-effects model is appropriate when similar research designs are used in included studies, and assumes that all studies represent a population with a fixed-effect size. Thus, any differences in effect sizes can be attributed to sampling error (Field & Gillett, 2010). Since the fixed-effect model generates a weighted average of effect size estimates, each individual participant is considered to be the unit of analysis.

A random-effects model, in contrast, considers each study to be the unit of analysis, as not all studies have similar treatments, and not all are drawn from similar populations. Any differences observed in a random-effects model can be attributed to variations between included studies, as well as sampling error (Field & Gillett, 2010). In educational studies, these differences typically might be attributable to variables including grade level, student socio-economic status, and teacher expertise.
In the fixed-effects model, included studies with larger sample sizes have a larger impact in the overall mean effect calculation, as these studies are assigned higher weights. Conversely, a random-effects model assigns weights proportionately, but in a much smaller range. Thus, studies with larger sample sizes are given less weight, and individual studies have less overall impact on the overall summary effect (Borenstein et al., 2009).

In terms of drawing overall conclusions, a random-effects model allows broader conclusions to be drawn, as generalizing the effect size beyond the sampled population is possible. Any inferences one might draw from a fixed-effects model are limited to only the studies, and their population, included in the selected studies (Field & Gillett, 2010).

In the present study, the investigator chose to proceed with the assumption of a random-effects model. It is an appropriate model to use for two main reasons. First, while the studies selected have some commonality in research design (an experimental or treatment group using self-reflection compared to a control group), it is unlikely that the studies could be considered functionally equivalent. It is more likely that the there were differences in the studies that likely impacted the results. That is, real differences exist in effect sizes across studies that are not based solely on sampling error. Therefore, a common effect size should not be assumed, and a random-effects model is justified. Second, the random-effects model allows for generalizations to be drawn beyond the populations included, which is useful in terms of policy recommendations and practical applications (Borenstein et al., 2009, pp. 83-84).

The investigator used Comprehensive Meta-Analysis, Version 3 (Biostat, 2015) to analyze the effects of reflective self-assessment on academic achievement when
considering the 19 included studies. Variation within and between studies was analyzed using forest plots, which visually represent the effect and confidence interval, and allows for comparison of studies with one another (Borenstein et al., 2009). Additionally, Comprehensive Meta-Analysis, Version 3 was used to calculate relationships between moderator variables and summarized effect sizes, as well as to look for evidence of publication bias.

**Limitations and Delimitations**

One limitation to this study is the potential presence of confounding variables. Within the treatment classrooms in each included study, the potential exists for confounding variables to impact the results. These might include variations in teaching style, classroom or building dynamics, and the fidelity to which teachers participating in interventions followed each investigator’s instructions and training.

Another limitation is institutional diversity. Most studies included in this meta-analysis originated in the graduate college school of education at a single institution, Seattle Pacific University. Some of the included studies are peer-reviewed, and some are not, including unpublished doctoral dissertations. The studies from this single source generally share a common research design, potentially allowing the results of this synthesis to shed some light on their effectiveness. At the same time, an overreliance on studies from a single university is a potential limitation in terms of institutional and geographic diversity.

A third limitation is small sample size. The overall limited number of studies (19) is a potential limitation, although as Valentine, Pigott, and Rothstein (2010 pointed out, the minimum number of studies required for a meta-analysis is logically two, as the intent
and value of research synthesis is the synthesis of multiple studies. Cooper (2003) argued that if one were to include only two research studies in a meta-analysis, to be useful the two studies should be replications of one another. Thus, a meta-analysis delimited to studies using similar reflective techniques in similar classroom settings, as the current case exhibits, seems sensible to this researcher.

One category of delimitation is the type of studies included. The studies in this meta-analysis involved elementary, middle, and high school students. The interventions were reflective self-assessment prompts (I learned statements, etc.) administered by the teacher as part of a classroom lesson. A quantitative measure compared the academic achievement of treatment and control groups. Ideally, this synthesis of a set of studies delimited as such should provide some insight into the effectiveness of these reflective techniques.

**Summary**

The investigator conducted a random-effects meta-analysis of the impact of reflective self-assessment on student achievement in grades K-12. Quantitative studies, primarily quasi-experiments, were included in the meta-analysis. The investigator conducted extensive electronic searches to obtain both published and unpublished studies, and potential publication bias was assessed by analyzing descriptive statistics and funnel plots. In the next chapter, the results of the meta-analysis are presented, and the results are analyzed in the context of current theory, existing research, and the current study’s research hypotheses and questions.
Chapter 4: Results

Study Characteristics

The 19 studies included in this meta-analysis measured the effect of self-reflection on academic achievement by students in grades K-12. Each study comparison produced an effect size using Hedges’ $g$, which is a standardized difference between the treatment and control groups in standard deviation units. The formula used to calculate

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>30</td>
<td>0 9</td>
</tr>
<tr>
<td>40</td>
<td>6 6 7 9</td>
</tr>
<tr>
<td>50</td>
<td>4 7</td>
</tr>
<tr>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>70</td>
<td>5</td>
</tr>
<tr>
<td>80</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>3</td>
</tr>
<tr>
<td>100</td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>0 4</td>
</tr>
<tr>
<td>120</td>
<td>2</td>
</tr>
<tr>
<td>130</td>
<td>4</td>
</tr>
<tr>
<td>140</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>8</td>
</tr>
<tr>
<td>160</td>
<td>3</td>
</tr>
<tr>
<td>170</td>
<td></td>
</tr>
<tr>
<td>180</td>
<td></td>
</tr>
<tr>
<td>190</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
</tr>
<tr>
<td>210</td>
<td></td>
</tr>
<tr>
<td>220</td>
<td></td>
</tr>
<tr>
<td>230</td>
<td></td>
</tr>
<tr>
<td>240</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>4</td>
</tr>
<tr>
<td>260</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 1. Stem and leaf plot of sample sizes from 19 studies.*

$$Hedges' \ g = \frac{M_1 - M_2}{SD_{pooled}}$$
where the numerator is the difference in means between treatment and control groups, and the denominator is the pooled and weighted standard deviation.

Publication dates for the 19 studies ranged from 1991 to 2017. Grade levels included one from elementary school (grades K-3), nine from middle school (grades 4-8), and nine from high school (grades 9-12). Academic subject categories included math \((n = 7)\), science \((n = 5)\), social studies \((n = 3)\), language arts \((n = 2)\), one foreign language, and one in the “other” category (problem solving). The median sample size of the studies was 79.5 students with the smallest study including 25 students (Glaubman, Glaubman, and Ofir, 1997), and the largest study including 254 students (Yin et al., 2008). See Figure 1 for a stem and leaf plot showing the sample size distribution of the 19 studies. Most sample sizes \((n = 11)\) cluster between \(n = 25\) and \(n = 75\).

Of the 19 studies included in this meta-analysis, 11 were from doctoral dissertations. It could be argued that including unpublished doctoral dissertations is less desirable compared to studies published in peer-reviewed academic journals. Dissertations are conducted by less-experienced graduate students as compared to most publications. However, as each dissertation is supervised by an advisory committee, dissertations may receive as much scrutiny, and in some cases more scrutiny, than peer-reviewed articles in academic journals. The inclusion of unpublished doctoral dissertations and the relation to publication bias will be discussed later in this chapter.

The studies employed a variety of reflective assessment strategies, which were administered to treatment groups. These strategies provided each student with an opportunity to reflect on the preceding lesson, and to either write, draw, or talk about their learning experience. Students were given the last five minutes of a class period or
lesson to respond in writing to a reflective prompt. In some cases, students were asked to discuss what they had written with another student, a reflective strategy known as “think aloud.”

Several types of prompts included:

1. I Learned Statement (Ellis, 2001).
3. It seems important to note... (Conner & Gunstone, 2004).
5. A question I have is ... (Conner & Gunstone, 2004).

Control groups did not have exposure to the reflective assessment prompts, but instead typically were given an alternate activity for the last five minutes of each lesson, including vocabulary review, or responding to content-based comprehension questions.

The duration of treatments for the included studies was between 2 and 12 weeks, with the most frequent treatment duration being 4-6 weeks (9 studies). Sixteen studies employed a pre-test/post-test design, and 3 employed a post-test only design. Twelve of the 19 studies conducted a retention test 8-12 weeks later. The dependent variable, academic achievement, was measured either by researcher designed instruments ($n = 5$), or preexisting curricular instruments ($n = 14$). Individual effect sizes ranged widely, from a low of $g = -1.186$ (O'Neal, 2015) to a high of $g = 1.964$ (Özsoy & Ataman, 2009). A stem and leaf plot (Figure 2) shows the effect size distribution of the 19 studies. The majority of effect sizes ($n = 13$) cluster between $g$ -0.38 and +0.69.
The inclusion of all studies yielded a summary mean effect of $g = 0.46$, with a 95% confidence interval (CI) of $g = 0.163$ at the lower limit (LL) and $g = 0.764$ at the upper limit (UL) (see Figure 3). Tests of statistical significance indicate support for rejection of the null hypothesis, since the mean effect of predictors in included studies was most likely not zero ($Z = 3.022, p = .003$).

*Figure 2.* Stem and leaf plot of effect sizes from 19 studies.

**Summary Mean Effect**

The inclusion of all studies yielded a summary mean effect of $g = 0.46$, with a 95% confidence interval (CI) of $g = 0.163$ at the lower limit (LL) and $g = 0.764$ at the upper limit (UL) (see Figure 3). Tests of statistical significance indicate support for rejection of the null hypothesis, since the mean effect of predictors in included studies was most likely not zero ($Z = 3.022, p = .003$).
Figure 3. Forest plot of observed and mean effects across author and academic subject.

Heterogeneity tests indicated that the random-effects model was justified, as there was variation in the effect size among studies that was not most likely attributable to sampling error ($Q = 173.883, p = .000$). Additional discussion of this decision to employ the random-effects model over the fixed-effects model is included in Chapter 3. $\tau$, ($T = 0.617$), the standard deviation of the expected true study effect, indicates the distribution of effect sizes above and below the mean by 1.21 ($1.96 \times \tau$). The $I^2$ value indicates that nearly 90% of the observed dispersion or variance in effect sizes is likely due to actual dispersion of true effect sizes and not simple random dispersion ($I^2 = 89.648$).
Effect of Reflection by Subject Area

The six academic subjects included in this meta-analysis were math \((n = 7)\), science \((n = 5)\), social studies \((n = 3)\), English Language Arts (ELA) \((n = 2)\), foreign language, \((n = 1)\) and other \((n = 1)\). The main effects for math, science, and social studies were analyzed individually. Due to small sample sizes, ELA, foreign language, and other were not analyzed as individual subjects.

Effect of reflection on math learning. The forest plot in Figure 4, a meta-analysis of math achievement \((n = 7)\), indicated a summary mean effect of \(g = 0.387\), CI of \(g = -0.320\) (LL) to 1.095 (UL), \(p = 0.283\). Observed effects range from \(g = -1.186\) to 1.964. The \(p\) value supports the acceptance of the null hypothesis, indicating a statistically non-significant finding.

<table>
<thead>
<tr>
<th>Study name</th>
<th>Grade level</th>
<th>Subject</th>
<th>Statistics for each study</th>
<th>Hedges's (g) and 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>O’Neal</td>
<td>High school</td>
<td>Math</td>
<td>-1.186 -1.857-0.516 0.001</td>
<td>Hedges's Lower Upper</td>
</tr>
<tr>
<td>Edwards</td>
<td>High school</td>
<td>Math</td>
<td>-0.382 -0.913 0.149 0.158</td>
<td>g limit limit p-Value</td>
</tr>
<tr>
<td>Johnson</td>
<td>Middle school</td>
<td>Math</td>
<td>-0.187 -0.756 0.383 0.521</td>
<td>Hedges's Lower Upper</td>
</tr>
<tr>
<td>Mvearech &amp; Kramarsk</td>
<td>Middle school</td>
<td>Math</td>
<td>0.309 -0.050 0.657 0.092</td>
<td>g limit limit p-Value</td>
</tr>
<tr>
<td>Baliram</td>
<td>High school</td>
<td>Math</td>
<td>0.554 0.094 1.014 0.018</td>
<td>Hedges's Lower Upper</td>
</tr>
<tr>
<td>Bond &amp; Ellis</td>
<td>Middle school</td>
<td>Math</td>
<td>1.619 1.153 2.084 0.000</td>
<td>g limit limit p-Value</td>
</tr>
<tr>
<td>Ozsoy &amp; Ataman</td>
<td>Middle school</td>
<td>Math</td>
<td>1.964 1.275 2.652 0.000</td>
<td>Hedges's Lower Upper</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.387 -0.320 1.095 0.283</td>
<td>g limit limit p-Value</td>
</tr>
</tbody>
</table>

Figure 4. Forest plot of observed and mean effect by subject (math).
**Effect of reflection on science learning.** The forest plot in Figure 5, a meta-analysis of science achievement \((n = 5)\), indicated a summary mean effect of \(g = 0.527\), CI of \(g = -0.137\) (LL) to 1.191 (UL), \(p = 0.120\). Observed effects range from \(g = -0.323\) to 1.935. The \(p\) value supports the acceptance of the null hypothesis, indicating a statistically non-significant finding.

<table>
<thead>
<tr>
<th>Study name</th>
<th>Grade level</th>
<th>Subject</th>
<th>Statistics for each study</th>
<th>Hedges’s g and 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Hedges’s Lower Upper</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(g)</td>
</tr>
<tr>
<td>Yin et al.</td>
<td>Middle school</td>
<td>Science</td>
<td>-0.323</td>
<td>-0.570</td>
</tr>
<tr>
<td>Shoop</td>
<td>High school</td>
<td>Science</td>
<td>-0.082</td>
<td>-0.419</td>
</tr>
<tr>
<td>Bianchi</td>
<td>High school</td>
<td>Science</td>
<td>0.621</td>
<td>0.241</td>
</tr>
<tr>
<td>Moore</td>
<td>Middle school</td>
<td>Science</td>
<td>0.672</td>
<td>0.096</td>
</tr>
<tr>
<td>Bajar-Sales et al.</td>
<td>High school</td>
<td>Science</td>
<td>1.935</td>
<td>1.327</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.527</td>
<td>-0.137</td>
</tr>
</tbody>
</table>

*Figure 5.* Forest plot of observed and mean effect by subject (science).

**Effect of reflection on social studies learning.** The forest plot in Figure 6, a meta-analysis of social studies achievement \((n = 3)\), indicated a summary mean effect of \(g = 0.084\), CI of \(g = -0.206\) (LL) to 0.375 (UL), \(p = 0.571\). Observed effects range from \(g = -0.192\) to 0.350. The \(p\) value supports the acceptance of the null hypothesis, indicating a statistically non-significant finding.
**Figure 6.** Forest plot of observed and mean effect by subject (social studies).

**Effect of Reflection by Grade Level**

The effect of reflection on academic achievement by grade level was analyzed to determine differences between middle school \( (n = 9) \) and high school \( (n = 9) \) (see Figures 7 and 8). Only one study at the elementary level (Glaubman, Glaubman, & Ofir, 1997) was part of this meta-analysis \( (g = 1.751, p = 0.00) \).

Analysis at the middle school level indicated a mean effect of \( g = 0.57 \), CI of \( g = 0.128 \) (LL) to 1.02 (UL), \( p = 0.012 \). Observed effects range from \( g = -0.323 \) to 1.964. The \( p \) value supports the rejection of the null hypothesis, indicating a statistically significant finding. \( Tau = 0.63 \), indicating that the distribution about the mean is \( g = \pm 1.24 \). The \( I^2 \) value indicates that 91% of the observed dispersion or variance in effect sizes is likely

<table>
<thead>
<tr>
<th>Study name</th>
<th>Grade level</th>
<th>Subject</th>
<th>Statistics for each study</th>
<th>Hedges's g and 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boyer</td>
<td>High school</td>
<td>Social Studies</td>
<td>-0.192</td>
<td>-0.557 0.174 0.304</td>
</tr>
<tr>
<td>Denton</td>
<td>Middle school</td>
<td>Social Studies</td>
<td>0.051</td>
<td>-0.219 0.321 0.712</td>
</tr>
<tr>
<td>Denton &amp; Ellis</td>
<td>Middle school</td>
<td>Social Studies</td>
<td>0.350</td>
<td>0.057 0.643 0.019</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.084</td>
<td>-0.206 0.375 0.571</td>
</tr>
</tbody>
</table>

\(-2.00 -1.00 0.00 1.00 2.00\)
due to actual dispersion of true effect sizes and not simple random dispersion ($I^2 = 91.032$).

<table>
<thead>
<tr>
<th>Study name</th>
<th>Subject</th>
<th>Statistics for each study</th>
<th>Hedges's g and 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yin et al.</td>
<td>Science</td>
<td>-0.323 -0.570 -0.076 0.010</td>
<td></td>
</tr>
<tr>
<td>Johnson</td>
<td>Math</td>
<td>-0.187 -0.756 0.383 0.521</td>
<td></td>
</tr>
<tr>
<td>Denton</td>
<td>Social Studies</td>
<td>0.051 -0.219 0.321 0.712</td>
<td></td>
</tr>
<tr>
<td>Mevarech &amp; Kramarski</td>
<td>Math</td>
<td>0.309 -0.050 0.667 0.092</td>
<td></td>
</tr>
<tr>
<td>Denton &amp; Ellis</td>
<td>Social Studies</td>
<td>0.350 0.057 0.643 0.019</td>
<td></td>
</tr>
<tr>
<td>Moore</td>
<td>Science</td>
<td>0.672 0.096 1.248 0.022</td>
<td></td>
</tr>
<tr>
<td>King</td>
<td>Other</td>
<td>1.064 0.316 1.812 0.005</td>
<td></td>
</tr>
<tr>
<td>Bond &amp; Ellis</td>
<td>Math</td>
<td>1.619 1.153 2.084 0.000</td>
<td></td>
</tr>
<tr>
<td>Ozsoy &amp; Ataman</td>
<td>Math</td>
<td>1.964 1.275 2.652 0.000</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 7. Forest plot of observed and mean effects across middle school level ($n = 9$).
Outcomes at the high school level indicated a mean effect of \( g = 0.253 \), CI of \( g = -0.188 \) (LL) to 0.695 (UL), \( p = 0.261 \). Observed effects range from \( g = -1.186 \) to 1.935. The \( p \) value supports the acceptance of the null hypothesis, indicating a statistically non-significant finding.

<table>
<thead>
<tr>
<th>Study name</th>
<th>Subject</th>
<th>Statistics for each study</th>
<th>Hedges's g and 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Hedges's g</td>
<td>Lower limit</td>
</tr>
<tr>
<td>O'Neal</td>
<td>Math</td>
<td>-1.186</td>
<td>-1.857</td>
</tr>
<tr>
<td>Edwards</td>
<td>Math</td>
<td>-0.382</td>
<td>-0.913</td>
</tr>
<tr>
<td>Boyer</td>
<td>Social Studies</td>
<td>-0.192</td>
<td>-0.557</td>
</tr>
<tr>
<td>Shoop</td>
<td>Science</td>
<td>-0.082</td>
<td>-0.419</td>
</tr>
<tr>
<td>Kourilenko</td>
<td>Foreign Lang</td>
<td>0.262</td>
<td>-0.255</td>
</tr>
<tr>
<td>Baliram</td>
<td>Math</td>
<td>0.554</td>
<td>0.094</td>
</tr>
<tr>
<td>Bianchi</td>
<td>Science</td>
<td>0.621</td>
<td>0.241</td>
</tr>
<tr>
<td>Evans</td>
<td>ELA</td>
<td>0.690</td>
<td>0.375</td>
</tr>
<tr>
<td>Bajar-Sales et al.</td>
<td>Science</td>
<td>1.935</td>
<td>1.327</td>
</tr>
</tbody>
</table>

\[ g = 0.253 \]

Figure 8. Forest plot of observed and mean effects across high school level (\( n = 9 \)).

**Effect of Reflection by Publication Type**

The effect of reflection on academic achievement by publication type was analyzed to determine differences between articles published in peer-reviewed journals (\( n = 8 \)) and unpublished doctoral dissertations (\( n = 11 \)) (see Figures 9 and 10).
Analysis of journal published studies indicated a mean effect of $g = 1.038$, CI of $g = 0.408$ (LL) to 1.668 (UL), $p = 0.001$. Observed effects range from $g = -0.323$ to 1.964. The $p$ value supports the rejection of the null hypothesis, indicating a statistically significant finding. $Tau = 0.54$, indicates that the effects (95%) distribute $g = \pm 1.07$ about the mean.

Analysis of unpublished doctoral dissertation studies indicated a mean effect of $g = 0.105$, CI of $g = -0.177$ (LL) to 0.387 (UL), $p = 0.466$. Observed effects range from $g = -1.186$ to 1.190. The $p$ value supports the acceptance of the null hypothesis, indicating a statistically non-significant finding.

<table>
<thead>
<tr>
<th>Study name</th>
<th>Grade level</th>
<th>Subject</th>
<th>Statistics for each study</th>
<th>Hedges's $g$ and 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hedges's $g$</td>
<td>Lower limit</td>
</tr>
<tr>
<td>Yin et al.</td>
<td>Middle school</td>
<td>Science</td>
<td>-0.323</td>
<td>-0.570</td>
</tr>
<tr>
<td>Mevarech &amp; Kramarski</td>
<td>Middle school</td>
<td>Math</td>
<td>0.309</td>
<td>0.050</td>
</tr>
<tr>
<td>Denton &amp; Ellis</td>
<td>Middle school</td>
<td>Social Studies</td>
<td>0.350</td>
<td>0.057</td>
</tr>
<tr>
<td>King</td>
<td>Middle school</td>
<td>Other</td>
<td>1.064</td>
<td>0.316</td>
</tr>
<tr>
<td>Bond &amp; Ellis</td>
<td>Middle school</td>
<td>Math</td>
<td>1.619</td>
<td>1.153</td>
</tr>
<tr>
<td>Glabman &amp; Glabman</td>
<td>Elementary</td>
<td>ELA</td>
<td>1.751</td>
<td>0.838</td>
</tr>
<tr>
<td>Bajar-Sales et al.</td>
<td>High school</td>
<td>Science</td>
<td>1.935</td>
<td>1.327</td>
</tr>
<tr>
<td>Ozsoy &amp; Ataman</td>
<td>Middle school</td>
<td>Math</td>
<td>1.964</td>
<td>1.275</td>
</tr>
</tbody>
</table>

$g = 1.038$, CI of $g = 0.408$ (LL) to 1.668 (UL), $p = 0.001$. Observed effects range from $g = -0.323$ to 1.964. The $p$ value supports the rejection of the null hypothesis, indicating a statistically significant finding. $Tau = 0.54$, indicates that the effects (95%) distribute $g = \pm 1.07$ about the mean.

$g = 0.105$, CI of $g = -0.177$ (LL) to 0.387 (UL), $p = 0.466$. Observed effects range from $g = -1.186$ to 1.190. The $p$ value supports the acceptance of the null hypothesis, indicating a statistically non-significant finding.

Figure 9. Forest plot of observed and mean effects across journal published studies ($n = 8$).
Figure 10. Forest plot of observed and mean effects across unpublished dissertations ($n = 11$).

Publication Bias Analysis

Funnel plot analysis.

A funnel plot analysis was conducted to visually analyze the possibility of publication bias. Figure 11 shows the funnel plot generated from 19 studies across six academic subjects, and three grade levels (elementary, middle, and high school), accounting for independence of effect sizes. The pattern is widespread, and does not clearly demonstrate more studies to the lower right and upper left, which could be evidence of potential publication bias. However, the dispersion of data nodes on this funnel plot also indicates that the studies have a low level of precision, most likely due to the overall small number of studies, as well as few studies that had larger sample sizes.
While the asymmetry may also suggest true heterogeneity among studies, this researcher finds this plot inconclusive (Borenstein et al., 2009).

Figure 11. Funnel plot of potential publication bias. Evidence indicating a lack of publication bias may simply be the fact that about half of the studies included were unpublished, and half were published. However, as noted in Figures 9 and 10, the group of unpublished studies indicated low mean effect (g = 0.105) as well as statistical insignificance (p = 0.466). When taken in context, these factors do not rule out the presence or absence of publication bias. Thus, the overall summary effect may be overestimated in this meta-analysis. Additional discussion of this potential bias is found in Chapter 5.

Orwin’s fail-safe N

Orwin’s fail-safe N is a method for discerning any potential bias in the overall effect. Orwin’s fail-safe N allows a determination of how many missing studies should be
added to the meta-analysis to bring the overall effect to a level other than zero. This level can be determined by the researcher to be the smallest substantive effect in the context of the study.

The combined use of funnel plots and Orwin’s fail-safe N allows a researcher to place a meta-analysis in one of three categories: impact of bias is small and would not impact effect size estimate; impact of bias is medium, and inclusion of missing studies would leave the effect size largely unchanged; and the impact of bias is large, and robustness of effect size could change if all relevant and missing studies were located and included (Borenstein et al., 2009).

This investigator used the Orwin’s fail-safe N formula to determine that if an additional 9 studies with a mean Hedges’ $g = 0.00$ were located and included in the meta-analysis, the summary effect size would be rendered trivial ($g =< 0.20$). An additional nine studies is nearly half again the total studies included in this meta-analysis, and therefore it is possible that the summary mean effect is overestimating the true population effect.

**Dubal and Tweedie’s trim and fill**

This procedure can be used to relocate any small outlier studies from the funnel plot’s positive effect size side (the right side in Figure 11) to the negative side to simulate a more balanced funnel plot. A visual inspection of the forest plot indicates general balance between the positive and negative sides. Therefore, the trim and fill procedure is not called for. This decision is supported by the investigator’s running of a trim and fill procedure using the random-effects model, which indicated that zero studies would be adjusted via trim and fill.
Chapter 5: Discussion

The purpose of this meta-analysis was to examine the effects of reflective assessment on academic achievement by students in grades K-12. A search for literature yielded 19 research studies that were included in this meta-analysis. This chapter will summarize the hypotheses which defined the scope of this study, discuss the findings of the meta-analysis, identify limitations, and provide suggestions for related future research.

Null Hypothesis Testing

Two null hypotheses were identified relating the effects of reflective self-assessment on academic achievement. Null Hypothesis 1 stated that there would be no statistically significant difference in the effects of reflective self-assessment on academic achievement compared to instruction that did not include such metacognition. Null Hypothesis 2 stated that there would be no statistically significant differences in the effects of reflective self-assessment on academic achievement as moderated by subject, grade level, or publication type. To address these hypotheses, a summary mean effect was calculated across all subjects, grade levels, and publication types to analyze the general effect of reflective self-assessment on academic achievement compared to instruction that did not include such metacognition. Then, separate analyses were conducted to calculate mean effects by subject, grade level, and publication type.

Null hypothesis 1, overall mean effect. This investigator conducted a random-effects meta-analysis to determine the overall effect of reflective self-assessment on academic achievement compared to instruction that did not include such metacognition. This overall mean effect included six academic subject areas (math, science, social
studies, ELA, foreign language, and other) and was measured by content or skill achievement tests. An overall summary mean effect was $g = 0.46$, with a 95% confidence interval (CI) of $g = 0.163$ at the lower limit (LL) and $g = .764$ at the upper limit (UL). Tests of statistical significance indicate support for rejection of the null hypothesis, since the mean effect of predictors in included studies was most likely not zero ($Z = 3.022, p = .003$). The confidence interval indicates that the effect size is not overly precise, and the effect size for educational interventions considered to be a medium effect ($g = 0.46$). An effect size of $g = 0.46$ means that the average score in the experimental group is 0.46 standard deviation higher than average score in the control group, and is equivalent to a 18% percentile gain. In practical terms, this gain may mean that students exposed to self-reflection will on average outperform 68% of students who are not exposed to self-reflection. The wide confidence interval ($g = 0.163$ at the lower limit and $g = 0.764$ at the upper limit) indicates a wide variability in effect sizes, that the overall expected effect for self-assessment is positive, and that students exposed to such conditions could be expected to outperform 56% to 78% of peers not exposed to self-reflection.

**Null hypothesis 2, moderator effects.** This investigator conducted a random-effects meta-analysis to determine the overall effect of self-assessment on academic achievement as moderated by academic subject, grade level, and publication type.

**Effects of self-reflection by academic subject.** Six academic subjects were included in the meta-analysis: math ($n = 7$), science ($n = 5$), social studies ($n = 3$), ELA ($n = 2$), foreign language ($n = 1$), and other ($n = 1$). Main effect analyses were conducted on math, science, and social studies. ELA, foreign language, and other subject categories
were excluded from subject analyses due to insufficient number of studies in each category.

**Effect of reflection on math learning.** A meta-analysis of math achievement \( (n = 7) \), indicated a summary mean effect of \( g = 0.387 \), CI of \( g = -0.320 \) (LL) to 1.095 (UL), \( p = 0.283 \). Observed effects range from \( g = -0.1.186 \) to 1.964. The \( p \) value supports the acceptance of the null hypothesis, indicating a statistically non-significant finding.

**Effect of reflection on science learning.** A meta-analysis of science achievement \( (n = 5) \), indicated a summary mean effect of \( g = 0.527 \), CI of \( g = -0.137 \) (LL) to 1.191 (UL), \( p = 0.120 \). Observed effects range from \( g = -0.323 \) to 1.935. The \( p \) value supports the acceptance of the null hypothesis, indicating a statistically non-significant finding.

**Effect of reflection on social studies learning.** A meta-analysis of social studies achievement \( (n = 3) \), indicated a summary mean effect of \( g = 0.084 \), CI of \( g = -0.206 \) (LL) to 0.375 (UL), \( p = 0.571 \). Observed effects range from \( g = -0.192 \) to 0.350. The \( p \) value supports the acceptance of the null hypothesis, indicating a statistically non-significant finding.

In all three subject areas, lack of statistically significant findings is noted. In each subject area, a positive effect size was determined, with science indicating the highest mean effect \( (g = 0.527) \). All subject area analyses indicated a lack of precision, with 95% confidence intervals ranging from as low as -0.320 to as high as +1.191. The statistically non-significant result for each subject area moderator is interesting when considering the statistically significant overall mean effect. These combined results encourage a more cautionary rejection of Null Hypothesis 1.
Effects of self-reflection by grade level. The effect of reflection on academic achievement by grade level was analyzed to determine differences between middle school \((n = 9)\) and high school \((n = 9)\). Only one study at the elementary level (Glaubman, Glaubman, & Ofir, 1997) was part of this meta-analysis \((g = 1.751, p = 0.00)\), and thus no grade level analysis was conducted with the elementary level study.

Analysis at the middle school level indicated a mean effect of \(g = 0.57\), CI of \(g = 0.128\) (LL) to 1.012 (UL), \(p = .012\). Observed effects range from \(g = -0.323\) to 1.964. The \(p\) value supports the rejection of the null hypothesis, indicating a statistically significant finding. The wide confidence interval indicates that the effect size is not overly precise, and the effect size for educational interventions considered to be a high effect \((g = 0.57)\).

An effect size of \(g = 0.57\) means that the average score in the experimental group is 0.57 standard deviation higher that the average score in the control group, and is equivalent to a 22% percentile gain. Thus, students exposed to self-reflection will on average outperform 72% of students who are not exposed to self-reflection. The wide confidence interval \((g = 0.128\) at the lower limit and \(g = 1.012\) at the upper limit) indicates a wide variability in effect sizes, that the overall expected effect for self-assessment is positive, and that middle school students exposed to such conditions could be expected to outperform 55% to 84% of peers not exposed to self-reflection.

Outcomes at the high school level indicated a mean effect of \(g = 0.253\), CI of \(g = -0.188\) (LL) to 0.695 (UL), \(p = 0.261\). Observed effects range from \(g = -1.186\) to 1.935. The \(p\) value supports the acceptance of the null hypothesis, indicating a statistically non-significant finding.
The statistically significant finding at the middle school level, along with a high effect size ($g = 0.57$), lends support to the notion that at the middle school level, reflective self-assessment can have positive outcomes on student achievement. The large, statistically significant effect size at the middle school level, compared to the moderate, statistically insignificant effect size at the high school level, is notable, and perhaps warrants further investigation. Is there some characteristic of the middle school learner or the middle school classroom that allows reflective self-assessment to have stronger impact than at the high school level? Is there some characteristic of the high school learner or high school classroom that lessens the impact of reflective self-assessment at the high school level? Is the content at lower grade levels easier for students to understand, and therefore students at lower levels show higher overall gains?

Cheung and Slavin (2016) note that when interpreting effect size, researchers should be cautious when interpreting what constitutes large, moderate, or small effect sizes. Context, specifically grade level context, may play a role in this case, as a small effect for a middle school student may constitute a high effect for a high school student. Larger effect sizes seen at primary levels in a subject like reading, for example, may be demonstrating not the impact of an intervention, but the fact that more reading growth is typical for a fourth grader than an eleventh grader (Cheung & Slavin, 2016). Bloom, Hill, Black, and Lipsey (2008) suggest that effect sizes be interpreted not as absolutes across grade levels and subjects, but as a proportion of how much growth is typical for a student in that subject and grade level.

**Effects of self-reflection by publication type.** The effect of reflection on academic achievement by publication type was analyzed to determine differences between articles
published in peer-reviewed journals ($n = 8$) and unpublished doctoral dissertations ($n = 11$). Analysis of journal published studies indicated a mean effect of $g = 1.038$, CI of $g = 0.408$ (LL) to 1.668 (UL), $p = 0.001$. Observed effects range from $g = -0.323$ to 1.964, and the $p$ value supports the rejection of the null hypothesis, indicating a statistically significant finding. Analysis of unpublished doctoral dissertation studies indicated a mean effect of $g = 0.105$, CI of $g = -0.177$ (LL) to 0.387 (UL), $p = 0.466$. Observed effects range from $g = -1.186$ to 1.190. The $p$ value supports the acceptance of the null hypothesis, indicating a statistically non-significant finding.

Contrasting the statistically significant high effect size for published articles ($g = 1.038$), with the overall summary mean effect ($g = 0.46$), lends support for the notion that publication bias may be present. Additionally, the unpublished studies have a low mean effect ($g = 0.105$) as well as statistical insignificance ($p = 0.466$), even as they represent more than half of the studies included in the meta-analysis. The high effect size for published articles may indicate that studies that did not have higher effect sizes, statistical significance, or both, were not submitted for publication. This is sometimes known as the file drawer problem, as a researcher who completes a study that has non-significant finding(s), low effect size(s), or both, may simply file the paper away instead of submitting it for publication (Slavin, 2008).

Additional publication bias analysis using a funnel plot and Orwin’s fail-safe N indicates that 9 additional studies would be needed to render the effect size as trivial ($g \leq 0.20$). An additional nine studies is approximately half again the total studies included in this meta-analysis, and therefore it seems possible that the summary mean effect is overestimating the true population effect.
Limitations

Several limitations in this study are important to note. One general limitation was the inclusion of classroom based, quasi-experimental studies in the meta-analysis. As opposed to true experiments, quasi-experimental social science studies done in classroom settings are limited by an investigator’s lack of control over aspects beyond the treatment variable(s). The nature of school and classroom settings means that threats to internal validity (history, maturation, regression, etc.) are likely to influence results (Campbell & Stanley, 1963). While it is often impractical to conduct true experiments in school or classroom settings, the use of results from quasi-experimental studies has the advantage of being applied contextually into real-world, practical situations. Thus, results and conclusions from such studies, while more threatened by internal validity, may have the potential for practical insights into what works and does not work with students in real classrooms.

Another general limitation is what could be considered the subjective aspect of conducting meta-analyses. While meta-analysis is primarily a type of quantitative statistical analysis, the selection and interpretation of research studies involves qualitative, judgment-based point of interpretation. Of note in the current study was the investigator’s decision whether a study employed a treatment (students responding to reflective prompts), which was compatible with the construct, and therefore included in the meta-analysis. There is some variability in the types of reflective prompts given to students, and it could be argued that any such variation in prompt (i.e., I Learned versus Clear and Unclear Windows) could lead to variability of response. Additionally, several
general critiques of the meta-analytic process, as reported by Borenstein, Hedges, Higgins, and Rothstein (2009) were reviewed in Chapter 3.

A limitation specific to this study is the small number of studies on reflective self-assessment included in this meta-analysis. Glass et al. (1981) noted the difficulty of conducting a perfect literature search, because somewhere out there exists one or more studies that could or should have been included. While this researcher identified several studies that purported to address student reflection, metacognition, or related constructs, fewer studies were located which employed the specific type of formative self-reflective techniques in question. This limited the number of included studies to less than 20. While meta-analyses can be conducted with as few as two studies (Borenstein et al., 2009), including a larger number of studies could have given more statistical power to the analyses, with the potential for more conclusions to be drawn. Additionally, Field (2013 and Borenstein et al. (2009) both recommend including a minimum of 10 studies per covariate, which was not met in moderator analysis categories.

Another limitation could be the lack of research institution diversity. Of the 19 studies included in this meta-analysis, 11 originated as doctoral dissertations from the graduate school of education at one university. One could argue that this narrow scope is a potential limitation to the theoretical breadth of this meta-analysis, since most studies originated at a single institution.

However, it could be argued that since the current study looked closely at a very specific type of classroom intervention (reflective self-assessment), the location of studies in a single research institution could be a positive rather than a negative factor. Several similar studies likely benefit from each other by sharing or replicating effective
research designs, and improved understating of the construct. In a review of replication studies, Makel and Plucker (2014) found that replication studies were “less likely to be successful when there was no overlap in authorship between the original and replicating articles” (p. 5). Thus, if several studies originated from one institution, the classroom intervention and treatment could be more uniform, thus improve the results and conclusions drawn. Overall, this researcher notes that studies from this single university share commonalities in research design, which strengthens the ability of this synthesis to improve understanding of the impact of reflective self-assessment.

**Recommendations for Future Research**

The effectiveness of reflective self-assessment on student achievement is reported in this meta-analysis and in primary studies. More research on the impact of student self-reflection on academic achievement is needed. Particularly useful would be replication studies at the middle school level, which was the level that this study found to have significant results. Such studies could help to substantiate results, deepen the understanding of the effect, and lend more statistical power to future research syntheses. Additionally, studies are called for at the elementary level to alleviate a lack of studies and to provide insight into differences between elementary, middle, and high school level effects of self-reflection. There may be a development threshold to the benefits of these interventions, in that there likely exists a level of intellectual or cognitive sophistication required for students to be metacognitively aware.

Conclusions drawn in the current meta-analysis were made based on a small number of studies. To substantiate these results, more primary research is needed overall, at all levels and in all subject areas. Replication studies are needed in all subject areas,
and the paucity of studies located in ELA and other humanities subjects indicates a potential research gap. The reason for this gap may be connected to challenges in assessing student achievement in subjects other than math and science. Researchers and educators may find it simpler to effectively assess math and science skills, as opposed to more subjective scoring of ELA or social studies achievement tests. Thus, some researchers may shy away from undertaking such studies.

The current study focused on the quantitative impact of reflective self-assessment on student academic achievement. Additional research may be useful in qualitative impacts of such treatment. Prompted reflective self-assessment could be a vector for student-to-teacher communication. An increase in meaningful communication about the learning process can help teachers see student misconceptions, as well as to forge a positive collaborative learning environment overall. Evans (2009) noted student survey results, which indicated students practicing reflective assessment “felt more willing to work collaboratively with their peers and to discuss difficult emotion-laden issues” (p. 80). Additional qualitative research on these and similar non-academic impacts of reflective self-assessment could be a useful area of future research.

The current meta-analysis attempted to measure the impact of reflective self-assessment on student achievement. It did not measure to what extent the use of such assessments were used by the teacher to modify instruction in subsequent lessons. Presumably the use of such a formative assessment gives a teacher a view of what each student understood or learned as a result of a particular lesson. This knowledge of student understanding can allow a teacher to modify future instruction for an individual student, for a group of students, or for the whole class.
Ideally, a teacher can respond to each student with feedback, either general or specific. This feedback can serve to strengthen the student’s learning, by confirming correct understanding, correcting misunderstanding, and by generally strengthening student/teacher communication and connections. Hattie (2012) identifies teaching strategies emphasizing feedback as one of the top ten influences on student achievement, with a high effect size, \( d = 0.75 \), well-above what Hattie considers to be the average size for educational interventions, \( d = 0.40 \). Several research questions might address the practical classroom applications for formative self-assessment:

1) Is it the act of reflecting what provides most the effect, or is it a combination of reflection plus any secondary actions on part of the teacher that provide the most impact on achievement? What is the difference between a student self-assessment, versus a student self-assessment accompanied by a teacher response?

2) What is the ideal frequency for the use of reflective self-assessment? Daily? Three times per week? Weekly? Does overuse of the same prompt (an “I Learned… statement, for example, repeated daily) contribute to a lessening effect due to student familiarity? Boyer (2015) chose a methodology that utilized seven different reflective prompts, which may be a way to motivate students to provide more authentic and engaged responses.

**Summary**

Reflective self-assessment appears to be an effective instructional approach for students in middle school and high school across academic subject areas. The results of this meta-analysis indicate that overall, students exposed to self-reflection outperformed students not exposed to self-reflection on measures of academic achievement (\( g = 0.46, p \))
Analysis using a funnel plot and Orwin’s fail-safe N suggests potential publication bias. The mean summary effect is statistically significant and would be considered a medium effect size in educational contexts. Variation by grade level was present, with middle school students showing a statistically significant, large effect size ($g = 0.57, p = .012$). High school students provided evidence of a statistically non-significant, low effect size ($g = 0.253, p = .261$). Results by subject area were statistically non-significant. Heterogeneity tests indicated that the random-effects model was justified, as there was variation in the effect size among studies that was not most likely attributable to sampling error ($Q = 173.883, p = .000$). A test of homogeneity indicated that nearly 90% of the observed dispersion or variance in effect sizes is likely due to actual dispersion of true effect sizes and not simple random dispersion ($I^2 = 89.648$).

While the overall effect size ($g = 0.46$) is considered medium, the cost in time and resources to implement such treatments is minimal. Providing metacognitive prompts to students in the last five minutes of a lesson or class period uses very little instructional time, and costs very little. Additionally, training teachers to effectively use this technique should require a less-intensive, and therefore, less costly, professional development. A low resource-intensive classroom practice, with potentially positive effects, may be a wise choice for educators looking to improve student outcomes. Policymakers who are looking for classroom adaptations to improve student achievement should consider an increased use of reflective self-assessment.

Current and future teachers have a variety of classroom practices at their disposal, including techniques related to metacognition, formative assessment, and reflective self-assessment. Reflective self-assessment, a form of metacognition, allows a student to think
about past, current, and future learning experiences. Through a synthesis of research on the classroom use of self-assessment, the current study provides support for the continued use of self-reflection.
References


Denton, D., & Ellis, A. (2017). *The effects of reflective prompts on academic achievement by middle school social studies students.* Unpublished manuscript, School of Education, Seattle Pacific University, Seattle, WA.


Moore, C. (2010). *Mediated reflection and science achievement of fourth grade students in a highly diverse international school* (Doctoral dissertation). Retrieved from Dissertations & Theses@Seattle Pacific University. (3451005)


