January 1st, 2019

Using Metacognitive Reflection to Improve Student Learning

Eric Howe
Seattle Pacific University

Follow this and additional works at: https://digitalcommons.spu.edu/soe_etd
Part of the Art Education Commons

Recommended Citation
Howe, Eric, "Using Metacognitive Reflection to Improve Student Learning" (2019). Education Dissertations. 44.
https://digitalcommons.spu.edu/soe_etd/44

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.
Using Metacognitive Reflection to Improve Student Learning

By

ERIC HOWE

A dissertation submitted in partial fulfillment

Of the requirements for the degree of

Doctor of Education

Seattle Pacific University
Using Metacognitive Reflection to Improve Student Learning

By

ERIC HOWE

A dissertation submitted in partial fulfillment

Of the requirements for the degree of

Doctor of Education

Seattle Pacific University

2019

Approved by

Arthur Ellis, Ed.D., Professor, Chairperson

(Nyaradzo Mvududu, Ed.D., Professor & Dean, Committee Member)

(John Bond Ed.D., Associate Professor, Committee Member)

Program Authorized to Offer Degree

SCHOOL OF EDUCATION

APRIL 2019

Nyaradzo Mvududu, Ed.D., Professor & Dean, School of Education
Copyright Page

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral 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 the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to University Microfilms, 1490 Eisenhower Place, P.O. 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 ____________

Eric Howe
6-4-2018
Dedication

This work is dedicated to Maude and Mildred, though they will never read it, I know they are proud.

1 Corinthians 13:4-7

4Love is patient, love is kind. It does not envy, it does not boast, it is not proud. 5It does not dishonor others, it is not self-seeking, it is not easily angered, it keeps no record of wrongs. 6Love does not delight in evil but rejoices with the truth. 7It always protects, always trusts, always hopes, always perseveres.

New International Version (NIV)

Acknowledgements

My wife gave up countless hours of time we normally would have spent together. She believed in me, and through her example I have become a better writer.

I would like to thank my dissertation committee. Dr. Ellis has provided support and encouragement throughout my doctoral studies. Dr. Mvududu helped me understand statistics and was always available to provide guidance. Dr. Bond added his insight to my research and reminded me to always park uphill when writing. With the support of professors like these, including the rest of the SPU faculty, the learning though difficult, was joyful.

My journey in this program was aided by my principal and other staff members at my building who asked how my studies were progressing and helped whenever I asked.

Finally, my mother taught me to read, encouraged me to pursue learning, and made sure I had opportunities to develop my gifts. She is a testimony to God’s faithfulness.
# Table of Contents

List of Figures ........................................................................................................................................ vi
List of Tables .......................................................................................................................................... vii
List of Appendices ................................................................................................................................. viii

Chapter 1: Introduction .......................................................................................................................... 2
  Problem.............................................................................................................................................. 2
  Significance......................................................................................................................................... 2
  Definitions.......................................................................................................................................... 3
  Research Questions............................................................................................................................. 3
  Hypothesis........................................................................................................................................... 4
  Structure of Dissertation ..................................................................................................................... 4

Chapter 2: Literature Review .................................................................................................................. 6
  Introduction.......................................................................................................................................... 6
  Historical Context ............................................................................................................................... 6
    Stage Theory...................................................................................................................................... 6
    Thought........................................................................................................................................... 7
    Social Development ......................................................................................................................... 7
  Theoretical Constructs ......................................................................................................................... 8
    Metacognition.................................................................................................................................. 8
    Metacognitive Knowledge ............................................................................................................... 10
    Metacognitive Experience .............................................................................................................. 10
    Metacognitive Goals ....................................................................................................................... 11
    Metacognitive Actions ..................................................................................................................... 11
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age of Students</td>
<td>11</td>
</tr>
<tr>
<td>Measurement of Metacognition</td>
<td>12</td>
</tr>
<tr>
<td>Metacognitive Reflection</td>
<td>14</td>
</tr>
<tr>
<td>Related Constructs</td>
<td>15</td>
</tr>
<tr>
<td>Self-regulation</td>
<td>15</td>
</tr>
<tr>
<td>Self-regulating Learning</td>
<td>16</td>
</tr>
<tr>
<td>Additional Constructs</td>
<td>16</td>
</tr>
<tr>
<td>Feedback</td>
<td>16</td>
</tr>
<tr>
<td>Teacher Feedback</td>
<td>17</td>
</tr>
<tr>
<td>Reflection in Art</td>
<td>18</td>
</tr>
<tr>
<td>Academic Language</td>
<td>19</td>
</tr>
<tr>
<td>Academic Language in the edTPA</td>
<td>20</td>
</tr>
<tr>
<td>Empirical Studies of Metacognition</td>
<td>21</td>
</tr>
<tr>
<td>Metacognition and Math Journals</td>
<td>21</td>
</tr>
<tr>
<td>How Metacognition Was Taught</td>
<td>22</td>
</tr>
<tr>
<td>Metacognition, Academic Achievement, and Intelligence</td>
<td>24</td>
</tr>
<tr>
<td>Teacher’s Thoughts</td>
<td>25</td>
</tr>
<tr>
<td>Metacognition and Confidence</td>
<td>26</td>
</tr>
<tr>
<td>Metacognition and Academic Achievement</td>
<td>27</td>
</tr>
<tr>
<td>Dissertation Synthesis</td>
<td>28</td>
</tr>
<tr>
<td>How to Prompt Metacognition</td>
<td>29</td>
</tr>
<tr>
<td>Learning Science</td>
<td>31</td>
</tr>
<tr>
<td>Empirical Studies Influence on Current Study</td>
<td>33</td>
</tr>
</tbody>
</table>
Chapter 3: Method ................................................................. 36

Introduction ........................................................................... 36

Research Design ..................................................................... 36

Participants and Sampling Process ........................................ 38

Demographics ....................................................................... 39

Assignment of Condition ....................................................... 40

Sample Size .......................................................................... 40

Description of Samples ......................................................... 40

Spring 2018 Sample .............................................................. 41

Autumn 2018 Sample ........................................................... 42

Winter 2019 Sample ............................................................. 43

Protection of Participants ....................................................... 44

Measures .............................................................................. 44

Reliability ............................................................................. 45

Validity .................................................................................. 46

Procedure ............................................................................. 46

Statistical Analysis ............................................................... 48

Descriptive .......................................................................... 50

Inferential ............................................................................. 50

Chapter 4: Results ................................................................. 51

Introduction .......................................................................... 51

Hypothesis ............................................................................ 51

Spring 2018 ......................................................................... 51
Descriptive Statistics ................................................................. 52
Inferential Statistics ................................................................. 54
Autumn 2018 ................................................................................ 53
Descriptive Statistics ................................................................. 54
Inferential Statistics ................................................................. 56
Winter 2019 ................................................................................ 55
Descriptive Statistics ................................................................. 56
Inferential Statistics ................................................................. 58
Effect sizes .................................................................................... 57
Summary ......................................................................................... 58
Chapter 5: Discussion .................................................................... 61
Introduction ....................................................................................... 62
Spring 2018 ..................................................................................... 62
Autumn 2018 ................................................................................... 63
Winter 2019 ..................................................................................... 64
Synthesis ......................................................................................... 66
Limitations ....................................................................................... 67
Intact Groups .................................................................................... 67
Pretest Sensitization ........................................................................ 68
Teacher Generated Test .................................................................... 68
Time of Day ....................................................................................... 68
Intelligence ....................................................................................... 69
Reading Ability ............................................................................... 69
Length ................................................................. 70
Attrition .............................................................. 71
Feedback ............................................................... 71
Measurement of Reflection ........................................ 72
Depth of Reflection .................................................. 72
Length of Term ....................................................... 73
Snow Days ............................................................. 73
Implications ............................................................ 73
Theory ................................................................. 74
Research .............................................................. 74
Practice ............................................................... 74
Further Research ................................................... 74
Conclusion ............................................................ 76
References ............................................................ 78
Appendices ............................................................ 93
List of Figures

Figure 1. Mean Score by Condition Spring 2018 ......................................................... 62
Figure 2. Mean Score by Condition Autumn 2018 ....................................................... 63
Figure 3. Mean Score by Condition Winter 2019 .......................................................... 64
List of Tables

Table 1. Quasi-Experimental Design .......................................................... 37
Table 2. Spring 2018 Sample ......................................................................... 41
Table 3. Autumn 2018 Sample ...................................................................... 42
Table 4. Winter 2019 Sample ........................................................................ 43
Table 5. Spring 2018 Descriptive Statistics ..................................................... 52
Table 6. Autumn 2018 Descriptive Statistics ................................................... 54
Table 7. Winter 2019 Descriptive Statistics ...................................................... 56
Table 8. Pretest to Posttest Effect Sizes in \( d \) ............................................... 58
Table 9. Pretest to Retention Test Effect Sizes in \( d \) ........................................ 58
Table 10. Characteristics of Winter 2019 Sample ........................................... 69
Abstract

The study of art, especially perspective, involves the use of specialized vocabulary words. Vocabulary words can be difficult to comprehend, but when students learn to use the specialized vocabulary or academic language of a subject, the learner is better able to think about the content. While academic language is only a part of a visual art curriculum, students need support from the teacher to learn it. Metacognitive reflection (MR) offers a method to increase student learning of academic language in art specifically, and other subjects in general. Teacher feedback naturally occurs in response to students’ reflections and gives the learner direction and motivation to continue learning. This quasi-experimental study used a repeated measures design with a sample of intact middle school visual art classes to determine the influence of MR and teacher feedback on students’ ability to learn and retain academic language related to perspective drawing as measured by a multiple-choice test. This study was conducted three separate times, to improve validity. While the MR treatment groups attained and maintained greater mean gains overall, post-hoc tests revealed that differences between groups in two of three studies were not statistically significant. The groups who engaged in reflection with feedback added a weighted mean gain of $d = .37$ to their posttest score beyond that of the comparison groups. This finding provides moderate evidence for the efficacy of practicing reflection with feedback in favor of conventional teaching alone.

Keywords: metacognition, academic vocabulary, reflection, teacher feedback, visual arts
Chapter 1: Introduction

Problem

The study of art, especially perspective, involves the use of specialized vocabulary words (Montague, 2013). Students in art do not always enjoy or respond well when asked to learn the more formal parts of the curriculum (Pennisi, 2013). This includes vocabulary words, which can be difficult to comprehend (Jucks & Paus, 2012), because they convey context specific concepts (Uccelli, Galloway, Barr, Meneses, & Dobbs, 2015). When students have learned to use the specialized vocabulary or academic language of a subject, communication between teacher and student improves (Lahey, 2017).

Knowing academic language also helps the learner think about the content (Nagy, Townsend, Lesaux, & Schmitt, 2012). While academic language is only a part of a visual art curriculum (National Coalition for Core Arts Standards, 2014), students need support from the teacher to learn it (Lahey, 2017). Metacognitive reflection (MR) offers a method to increase student learning of academic language related to perspective drawing, and other subjects in general (Bond, Denton, & Ellis, 2015; Jucks & Paus, 2012).

Significance

Oddly, while reflection was almost universally called for in art education theory, few empirical studies have examined the efficacy of this technique when applied to the art classroom. A Boolean search of five leading peer reviewed art education journals revealed 21 articles that included the words “reflective” and “assessment” in the title, or body of the text. Of these, only a handful used the words together in the sense Bond,
Denton, and Ellis (2015) studied. None of these studies reported statistical information that could be further examined.

Definitions

In the study of reflection, metacognition, and related constructs, many words and the concepts they represent were not clearly defined, or appropriately applied (Dinsmore, Alexander, & Louglin, 2008; Schunk, 2008). In this study there was consistent use of the following definitions:

*Metacognition:* thinking about thinking (Flavell, 1985).

*Metacognitive Reflection:* thinking about what strategies worked well during learning and adjusting based on this information to guide future learning (Georghiades, 2004a).

The precise definitions of academic vocabulary and academic language (Baumann & Graves, 2010; Uccelli et al., 2015) remained under debate. For this study:

*Academic Vocabulary:* special words used to describe academic ideas (Cunningham & Moore, 1993).

*Academic Language:* applying academic vocabulary to talk about what is being learned in school (Uccelli et al., 2015).

Research Questions

This study was guided by the following questions:

1. To what extent does MR influence students’ initial ability to learn academic language related to perspective drawing?

2. To what extent does teacher feedback to the MR influence students’ initial ability to learn academic language related to perspective drawing?
3. To what extent does MR influence students’ ability to retain academic language related to perspective drawing?

4. To what extent does teacher feedback to the MR influence students’ ability to retain academic language related to perspective drawing?

**Hypothesis**

The above questions lead to this hypothesis:

H<sub>0</sub>: There is no statistically significant difference based on group (three levels: reflection with feedback, reflection, comparison) on students’ ability to learn and retain academic language related to perspective drawing as measured by mean score differences on a multiple-choice test.

H<sub>a</sub>: There is a statistically significant difference based on group (three levels: reflection with feedback, reflection, comparison) on students’ initial ability to learn and retain academic language related to perspective drawing as measured by mean score differences on a multiple-choice test.

**Structure of Dissertation**

This remainder of this dissertation was divided into four chapters: Literature Review, Method, Results, and Discussion. A summary of their content follows.

Chapter two includes more comprehensive information about metacognition including its history and key developers. The chapter presents empirical data taken from studies to demonstrate the effectiveness of MR in improving learning outcomes and explains how these studies influence the current study. There is also an examination of secondary constructs such as feedback, teacher feedback, reflection in art, academic language, and academic language in the edTPA.
Chapter three describes the methods used in the study. Methods include a description of the design of the study, information about participants in the study and how they were selected, procedures for teaching and assessing the learning, and the data analysis used to address the results of the assessments.

Chapter four presents the results of the study. Data from each of three iterations of the study is presented in turn including the relevant descriptive and inferential statistics. Effect sizes pretest to posttest and pretest to retention test are also displayed in chart form for ease of comparison within and across the three related studies.

Chapter five contains the author’s reflection on the results organized by study and question. An overall synthesis of trends across studies is presented. Limitations of the study, and discussion of the practical implications of the study follow. The study concludes with suggestions for future research in the area of MR, and some final thoughts.
Chapter 2: Literature Review

Introduction

This chapter includes additional information about metacognition including the history and key developers. There is also an examination of secondary constructs such as feedback, teacher feedback, reflection in art, academic language, and academic language in the edTPA. Next, there is a review of empirical data taken from studies to demonstrate the effectiveness of MR in improving learning outcomes. The chapter concludes with an explanation of how the literature has influenced the current study.

Historical Context

Among the many theoreticians who have contributed to the study of metacognition, three stood out as foundational: Piaget, James, and Vygotsky.

Stage Theory. Jean Piaget, a Swiss psychologist, studied the development of the reasoning process in children. He concluded, for example, that 10-year old children formed social groups, and autonomously organized rules and values, yet the four-year-old child did not, because cognitive structures developed over time (Piaget, 1967, p. 65).

Piaget noted four turning points in a child’s development, when a child could (a) communicate verbally; (b) think about, communicate with, and understand others; (c) turn their thoughts inward; (d) think about their thoughts and thought process (Piaget, 1928).

Piaget proposed a development theory consisting of stages, based on the four turning points. In the final stages of development, a child could reflect on his or her own thinking. Piaget called this type of thinking, “thought raised to the second power” (Piaget, 1967, p. 63) more commonly called metacognition (Flavell, 1985).
**Thought.** John Dewey called William James the “greatest of American psychologists” (Dewey, 1910, p. 506). James described how an action started in the mind, and when acted on, formed a habit no longer requiring thought or will to carry out (James, 1992). For this reason, he encouraged teachers to promote “good habits” in their students (p. 750). This included mental habits so “our higher powers of mind will be set free for their own proper work” (James, 1992, p. 146).

James explained that cognition “is a function of consciousness” (James, 1885, p. 26). This conscious thinking involved a stream of thought accessible to the one doing the thinking (James, 1884, p. 4). Both the mind of the thinker and outside influences could direct attention to a perceived object so that it “…most completely occupies the mind…” and “…the mind attends to it with maximum power…” (James, 1899, p. 34). James ultimately viewed the question of who was doing the thinking as a topic for future research but concluded for now “the thoughts themselves are the thinkers” (James, 1992, p. 209).

A conscious person could remember the order of thoughts and how they related to each other (James, 1992). The mind also knew what it had learned in the past even when it could not recall exact details, because there was a difference between storage and retrieval. For instance, when one tried to remember a forgotten name and wrong names were proposed, then the person knew immediately they were incorrect (James, 1884). The ability to access one’s thoughts and memories as described by James allowed people to think reflectively and formed the basis of self-regulation (Fox & Riconscente, 2008).

**Social Development.** Lev Vygotsky described a threefold process by which a child learns to self-regulate. This process occurred during adolescence when the subject’s
control moved from (a) adult directed through words; (b) to directing others’ attention with words; (c) to using one’s own words internally to control one’s own attention (Vygotsky, 1978). Vygotsky also distinguished between learning and development. Learning was more basic and could happen without a psychological change in the learner. Development was marked by increased psychological ability (Vygotsky, 1978).

To assess a child’s level of development Vygotsky formulated the zone of proximal development (ZPD) which was, “the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers” (Vygotsky, 1978, p. 86). Assessment was accomplished through interaction with the child.

Collaboration was not a mutual effort between teacher and student to solve a problem, but an interaction between them leading to a solved problem. Because the child's developing psychological abilities included cognitive and social functioning, Vygotsky (1986) suggested assessment could be accomplished through the teacher beginning to solve a problem and seeing if the child could finish solving it. When assessment was formative and took place through social interaction, this method could be a more effective measure of potential than I.Q. tests which relied heavily on fixed cultural, socioeconomic, and educational assumptions (Lidz & Gindis, 2003).

**Theoretical Constructs**

**Metacognition.** Building on the work of these foundational theoreticians, John Flavell and Ann Brown developed the theory of metacognition through research on children’s use of learning strategies. Brown (1994) expressed metacognition as the
process through which learners “have insight into their own strengths and weaknesses and access to their own repertoires of strategies for learning” (p.9). Brown noted two main domains of metacognition: knowing about one’s thinking and controlling thinking. She also listed two main issues that confused the term: no easy way to separate cognitive and metacognitive processes, and the many roots of the theory (Brown, 1987).

There were four routes early psychological studies of metacognition had taken (Brown, 1987). First, verbal reports as employed in the study by Myers and Paris (1978) with their attendant unreliability (Brown, 1978). Next, studies of executive control like those conducted by Wellman, Ritter, and Flavell (1975). Followed by, attempts to understand (e.g. Paris & Myers, 1981) and describe self-regulation (e.g. Reeve, & Brown, 1984), when according to Piaget the mature learner was able to control their own thinking (Piaget, 1967). Finally, examining other-regulation, which occurred during the course of social interaction, of the sort described by Vygotsky (1978), and studied by Palinscar and Brown (1984).

After Brown’s untimely death in 1999, Flavell continued to research metacognition extensively (e.g. Flavell, 1979; Flavell, 1985; Flavell, 1999; Flavell, 2000; Flavell, Green, & Flavell, 2000; Wellman, Ritter, & Flavell, 1975). Flavell stated uncertainty about whether he or Brown coined the term metacognition, noting that while the term originated with one of them the idea was not original to either of them (Shaughnessy, 2008).

Brown’s two domains of metacognition were later described as Knowledge of Cognition (KoC) and Regulation of Cognition (RoC) (Bannister-Tyrrell & Clary, 2017). KoC included knowing about what one knew or declarative knowledge, knowing how to
do something or *procedural knowledge*, and knowing the appropriate way and time to use declarative and procedural knowledge or *conditional knowledge* (Moshman, 2018). RoC included application of metacognitive skills to regulate learning. These skills included planning how to learn, monitoring learning in the moment, and evaluation of learning (Bannister-Tyrrell & Clary, 2017).

Using different terminology to describe similar concepts, Flavell differentiated among metacognitive knowledge, experiences, goals, and actions (Flavell, 1979).

**Metacognitive knowledge.** Papaleontiou-Louca (2008) explained that metacognitive knowledge included those beliefs stored in long-term memory. These could be manifest in both knowing what and knowing how. Knowing one forgot things and needed a reminder (what) and then wrote a list (how) was an example (Papaleontiou-Louca, 2008). Within the category of metacognitive knowledge were beliefs about people, such as knowing someone learned best by hearing. When a person needed to be able to recall a location (goal) she might have realized she could memorize the gist of something easily, while exact recall of details would be more difficult, so she repeated the address back to herself to aid memory (Papaleontiou-Louca, 2008).

**Metacognitive experiences.** Metacognitive experiences may have been less easily articulated by the knower. These experiences involved emotions brought on while learning, like a feeling of doing well at something (Flavell, 1979) or a feeling of frustration when unable to grasp a concept (Papaleontiou-Louca, 2008). The desire to understand might have activated a strategy such as talking through the steps mentally. During this process metacognitive knowledge and skills worked together. Papaleontiou-Louca (2008) stated, “…it seems likely that metacognitive knowledge, metacognitive
experience and metacognitive skills, are constantly informing and eliciting one another during the course of a cognitive task” (p. 14).

**Metacognitive goals.** Metacognitive goals were one reason a person engaged in a task (Flavell, 1979). Schunk (1994) explained how students set diverse types of goals based on their personality and past experiences. Learning centered students attempted to increase their skill and knowledge while ego or performance-centric students strove to perform at an elevated level. The first type of students valued learning and set their own goals while the second type enjoyed feeling competent and compared their performance to established standards (Meece, 1994). Thus, a student’s outlook affected the difficulty level of set goals and the persistence with which goals were pursued.

**Metacognitive actions.** Metacognitive actions were the strategies people used to attain their goals. Strategies could be developed over time by the learner (Flavell, 1979). For example, a student noticed the bold sections in the textbook corresponded to test questions and spent time learning these sections to prepare for the test. Strategies could also be taught (Dignath & Büttner, 2008). In this instance the teacher might have directed students’ attention to the bold areas of the text and encouraged them to memorize these parts. In the future students could remember and apply this strategy to help themselves learn the main points in a new subject.

**Age of students.** The exact age a student must be in order to engage in metacognition has been the subject of debate (Dignath & Büttner, 2018). Piaget concluded that students must be older than ages “eleven or twelve” in order to engage in “abstract intellectual operations” (Piaget, 1967, p. 6). Since then researchers have sought to find just how young students could be and still benefit from metacognitive practices
Brown (1997) concluded that young children do not know much about their own mental capacity or the way learning strategies could benefit them and therefore did not engage in much metacognition. They did, however, naturally use some strategies. Brown (1997) shared an example of three-year-old subjects remembering what cup an item was under by putting their hand on it or moving the cup away from the other two.

Dignath and Büttner (2008) conducted a meta-analysis which included 49 primary school level studies. Based on these more recent studies they concluded that “the development of children’s metacognition goes on during schooling from 5 to 16 years” in increasing complexity over time (Dignath & Büttner, 2008, p. 235). They clarified that while young students benefited from learning strategies, older students benefited from opportunities to refine their learning strategies (Dignath & Büttner, 2008). This partially explained why in a more recent study by Dignath and Büttner (2018), MR, a more advanced application of learning strategies, was noted as only beneficial when used at the secondary level.

**Measurement of metacognition.** Metacognition involved mental activity such as when a person after the fact thought back to remember where they learned something or made a judgment about how well they learned it or made predictions of their ability to learn new information in the future based on past experience (Proust, 2007). Accurate measurement of metacognitive use was challenging because one could not physically see what was happening in the learner’s mind (Panadero, Klug, & Järvelä, 2016).

Researchers distinguished a measure of learning by its timing. Online measures occurred while the learning was taking place. Offline measures were applied after the
fact. Online measures of metacognitive process, while potentially disruptive (Schellings & Van Hout-Wolters, 2011) were considered more accurate (Bannert & Mengelkamp, 2008; Dent & Koenka, 2016; Ohtani and Hisasaka, 2018), as offline measures were poor predictors of learning outcomes (Schellings & Van Hout-Wolters, 2011).

Common measures of metacognition included structured interviews, self-report questionnaires, monitoring student behavior during learning, monitoring student speech while learning (Dent & Koenka, 2016), asking students to watch video of themselves learning and report using the video as a cue, and computer generated predictions of what strategies were being used based on answer patterns (Schellings & Van Hout-Wolters, 2011). While scale-based, self-report measures involving questionnaires were most common (Schellings & Van Hout-Wolters, 2011), any measure that relied on verbalizing was limited, because there were non-linguistic mental processes taking place during metacognition, that simply could not be expressed verbally, and therefore not communicated to others (Proust, 2007). Additionally, learners’ perceptions of their own metacognition were found to be inaccurate (McCardle & Hadwin, 2015).

While recognizing the limitations of self-report methods (Schellings & Van Hout-Wolters, 2011), when applied properly they were often the single most practical way of measuring the highly personal process of metacognition (McCardle & Hadwin, 2015). Surveys could be administered at the conclusion of a learning session in the classroom, while the learner remembered what they were thinking (Ohtani & Hisasaka, 2018). An increase in accuracy, but also cost, could be attained through using a combination of measures (Dinsmore et al., 2008).
**Metacognitive reflection.** One metacognitive action that was developed over time and used by students to attain learning goals was MR. Metacognitive reflection was defined as thinking about learning or “critical revisiting of the learning process in the sense of noting important points of the procedures followed, acknowledging mistakes made on the way, identifying relationships and tracing connections between initial understanding and learning outcome” (Georghiades, 2004a, p. 371).

Though beneficial, students did not always engage in metacognitive thinking (Wismath, Orr, & Good, 2014), even when provided with a variety of well-designed prompts and activities (Kwon & Jonassen, 2011). A conducive classroom environment was necessary for students to engage in meaningful reflection (Black, & Wiliam, 2009). Students must have trusted the teacher (Georghiades, 2004a) and been free from fear of judgment from other students or the instructor (Slinger-Friedman & Patterson, 2016).

Even with ideal conditions, students needed the guidance of a teacher as they engaged in reflection. Learners’ perceptions of their use of metacognition was often inaccurate (McCardle & Hadwin, 2015). They may have felt they were regularly reflecting in deep ways on their learning, while in reality their reflections were relatively shallow and infrequent. Additionally, there was a potential for students to be misled by their reflections. When students found a subject easy to learn and conflated this with thinking they would be able to easily remember the subject in the future, they may not have devoted enough effort to review (Proust, 2007). Finally, students must have applied the results of their reflection to future learning in order to complete the process and this was not guaranteed to happen (Tarricone, 2011).
To mitigate these pitfalls, teachers could model MR for their students (Ellis, Denton, & Bond, 2014). Zimmerman (2013) explained that when a student carefully watched a skilled person such as a teacher or more advanced student perform a task and subsequently observed positive benefits as a result of correct task completion, the student could become highly motivated to continue with their own learning. Additionally, when a person modeling showed how they self-corrected this helped the observing student in the future when they encountered similar situations (Zimmerman, 2013). Think Aloud was an example of one such strategy where a teacher talked through their thinking as they solved problems in front of students (Ellis et al., 2014).

Though many teachers seemed to understand the benefits of MR they did little to promote it (Dignath & Büttner, 2018). If the educator did not intentionally plan time for reflective habits to be cultivated, the other parts of the curriculum squeezed this out (Zuckerman, 2003). Fortunately, MR activities could be simple enough for teachers to easily implement in the face of competing priorities (Bannister-Tyrrell & Clary, 2017).

Related Constructs

Brown (1987) noted that metacognition was a theory with multiple roots that covered a wide scope of mental activity. She expressed hope that this “many headed monster” (p. 105) could be more fully described through methodical research. Two of the very closely related areas of study were self-regulation and self-regulated learning.

Self-regulation. Self-regulation (SR) dealt with those areas where an individual “has control over his own learning, steering and directing cognitive and motivation processes to achieve the learning goal” (Boekaerts & Cascallar, 2006, p. 200). This control implied an intentional, conscious process, including a goal structure (Boekaerts &
Cascallar, 2006). Therefore, theories of self-regulation have focused on why people learn, what people were trying to achieve through learning, and how people used strategies to improve their learning (Reeve, Ryan, Deci, & Jang, 2008).

**Self-regulated learning.** Self-regulated learning (SRL) had roots in the works of Piaget (1967), Vygotsky (1978), and more recently in Bandura’s (1993) linkage of SRL to neo-behaviorism. Studies of SRL had typically focused on academic learning (Dinsmore et al., 2008). Because metacognition was both a regulatory mechanism employed during the evaluation phase of SRL (Metcalfe, 2008) and an overarching construct encompassing some (Pintrich, 2000) or many (Zimmerman, 2000) of the regulatory process used in SRL, it was used at times interchangeably with SRL (Schunk, 2008).

After reviewing 255 studies, Dinsmore et al. (2008) listed three problems in many studies including (a) no definition of the constructs; (b) wrong construct identified through excessive key word tagging; (c) multiple constructs studied at once. These problems caused a conceptual haze. For clarity, apart from the brief mention of SR and SRL above, this study was delimited to metacognition and MR only.

**Additional Constructs**

**Feedback.** In addition to metacognition and MR, another strategy that held promise for improved learning outcomes was feedback. Feedback could come from teachers, peers, or the learner themselves (Panadero, 2017), and was critical in guiding and sustaining learning (Hattie & Timperley, 2007). Butler and Winne (1995) acknowledged the distinction between internal and external feedback and noted that regardless of the source, if it got the learner to engage in metacognition it could promote
further learning. Self-generated feedback has been studied (e.g. Lam, 2015) and found to be an important factor in learning but was outside the scope of the current study. Similarly, peer feedback while helpful in promoting metacognition (Benton, 2013), was not examined in the current study which was delimited to teacher feedback.

**Teacher feedback.** While teacher feedback seemed to be a common feature of many classrooms, it was often misapplied by well-meaning teachers when they praised a student without addressing the task (Hattie & Timperley, 2007). Schunk, Pintrich, and Meece, (2008) noted four types of productive feedback teachers could use (a) performance feedback, which could be positive or corrective and most helpful if it carried the idea that students can and are improving; (b) motivational feedback, such as seeing one’s peers succeed and be complimented helped someone persevere until they too succeeded; (c) attributional feedback, for example, telling students their extra work preparing for the test paid off, helped students see success as related to effort rather than ability; (d) strategy feedback, such as reminding students they followed the order of operations and got it right, encouraged additional use of the effective strategy.

When done well, feedback helped the learner understand where to head next in their learning (Hattie & Clarke, 2019). This gave the learner direction and motivation to continue. In a recent meta-analysis, Hattie & Clarke (2019) reported feedback as having an effect of $d = 0.73$ on student achievement. Reading and responding to student reflections on the day’s learning helped the teacher stayed focused on task type feedback, with an emphasis on future learning. Data from these statements was then used to guide the next steps of instruction (Black & Wiliam, 2009).
Feedback could also prevent faulty models from being adopted. Students who did not know they misunderstood a concept might not have realized they had misunderstood until they were presented with information that challenged their understanding. Presenting students with feedback on their correct and incorrect answers was one way to help them sort out which areas they had learned and still needed to learn (Kwon & Jonassen, 2011).

The exact timing of the delivery of the feedback was controversial. Perhaps the best approach was to use a mixture of immediate and delayed feedback (Shute, 2008). What was clear was that students generally needed between three to five exposures to new material in order to learn it (Hattie & Clarke, 2019). A teacher’s feedback could have provided some of these needed exposures.

**Reflection in art.** Reflection had long been a part of art education. As early as 1992, Winner and Simmons writing for Harvard’s Arts PROPEL project asked art teachers to encourage art students to reflect on their work. Standards published by the National Coalition for Core Arts Standards (2018) called on students to respond to artwork by analyzing, interpreting, and evaluating art work. This response could have presumably been achieved through MR.

Many state and local school districts also advocated for students to reflect on their work. For instance, the Maryland State Department of Education (2018) published a standard titled *Use of Student Self-Reflection in Assessment Tasks*. Included on their Fine Arts Education website was a short form for students to use to reflect on their work. Another example was the San Diego Unified School District’s (2018) *Self-Reflection Assessments*. Oddly, while reflection was almost universally called for in art education
theory, few empirical studies had examined the efficacy of reflection when applied to the art classroom, providing impetus for the current study.

**Academic language.** The current study included teaching students academic language related to perspective drawing. Academic language could be difficult to comprehend (Jucks & Paus, 2012), because it conveyed context specific concepts (Uccelli et al., 2015). The exact definitions of academic vocabulary (Baumann & Graves, 2010) and academic language (Uccelli et al., 2015) and how they were best taught (Krashen, 2012) remained controversial. However, there was some consensus that in a broad sense, academic language consisted of specialized vocabulary and its application in a subject (Uccelli et al., 2015).

Academic language skill was increasingly recognized as critical to student learning (Lawrence, Corosson, Paré-Blagoev, & Snow, 2015; Uccelli et al., 2015). When students had learned to use the academic language of a subject, communication between teacher and student improved (Lahey, 2017). Knowing academic language also helped the learner think about the content (Nagy et al., 2012).

Properly understanding the technical language of a subject was difficult (Jucks & Paus, 2012). Students needed support from the teacher to learn and apply subject specific vocabulary (Lahey, 2017). If a technical term had elements from colloquial language, the learner may have incorrectly assumed they understand the word’s application. Context clues could have similarly lulled the reader into a false sense of understanding (Jucks & Paus, 2012).

Metacognition reflection was one method that could support student learning (Bond et al., 2015). During reflection a learner might have wondered if they really
comprehended a word, realize they didn’t and asked the teacher for help, or looked it up in a dictionary, and subsequently arrived at a more accurate understanding (Jucks & Paus, 2012). Beyond simply reciting a definition, students used these words to convey context specific concepts (Uccelli et al., 2015). Meaningful learning included practicing the language (Uccelli et al., 2015) in various ways, including discussion (Lawrence et al., 2015). While this approach privileged formal language, thoughtful teachers took care not to devalue the language skills learners brought to the classroom from their communities (MacSwan, 2018).

**Academic language in the edTPA.** Because of the pivotal role of academic langue in learning, a section of the edTPA evaluated teacher candidates on their ability to teach academic language (Lahey, 2017). The edTPA was developed by the American Association of Colleges for Teacher Education (AACTE) and the Stanford Center for Assessment, Learning, and Equity (SCALE) (Kissau, Hart, & Algozzine, 2017). The goal was to create a formative assessment (Ledwell & Oyler, 2016), which could provide recommendations for improvement of the perceived poor quality of teacher preparation programs (Dover & Schultz, 2016). Public attention had been focused on improvement in education since 1983 when A Nation at Risk (National Commission on Excellence in Education, 1983) was published.

The edTPA was used as a licensing requirement in 40 states (Cohen, Hutt, Berlin, Mathews, McGraw, & Gottlieb, 2018). While many educators took issue with the edTPA being used as sole measure of teacher quality, many saw the value of the assessment (Seymour, Burns, & Henry, 2018) and it was likely to be used for quite some time.
Beyond the educational benefits of teachers being familiar with academic language teaching methods, knowing them was a requirement for licensure.

**Empirical Studies of Metacognition**

The following studies provided evidence of the effectiveness of MR in improving learning outcomes. The findings of these studies, methods used, and authors’ reflections on the efficacy of the methods and results have informed the design and execution of the current study.

**Metacognition and math journals.** Baliram and Ellis (2019) conducted a study in a high school geometry classroom. Five intact classes were randomly assigned to a treatment or comparison condition. A pretest was administered to help control for preexisting group differences. This was followed by an intervention consisting of MR, posttest, and retention test. The test was developed and published by a textbook company.

This study was informed by Hattie’s (2012) work and included a teacher feedback component. Other researchers (e.g. Bianchi, 2007) have pointed out the possible differential effect of teachers reacting to student reflections. If one group benefited from improved instruction based on their expressed needs and another did not, this could have confounded interpretation of results. In order to prevent this, a third party, in this case one of the researchers, read the student responses and provided feedback. While this may have avoided biased responses from the teacher to individual students, the teacher was aware of general trends in feedback and did act on these. Therefore, this method may have only partially controlled for the differential effect noted earlier.

The author acknowledged that intact classes may have impacted results. For example, the sample of 75 participants was slightly below the number indicated by a
power analysis. Nevertheless, the results did achieve statistical significance with the
treatment group outscoring the comparison group on the posttest ($F(1, 73) = 7.27, p = .009, \eta_p^2 = .09$) (Baliram & Ellis, 2019).

In addition to the tests of geometry content, there was also a survey administered
at the end of the study to gauge student feelings of preparedness for the assessment, and
overall satisfaction with their experience. The literature predicted groups engaged in
reflection and receiving feedback to be more prepared and supported than those who
were not. Inexplicably, the groups who did not receive treatment answered more
favorably to all four questions, but this difference was not statistically significant.
Perhaps the teacher subconsciously compensated the comparison groups lack of treatment
by being more attentive to that group’s needs. Though there were limitations, this study
was thoughtfully conducted and was representative of what could be realistically done in
educational settings (Gall, Gall, & Borg, 2007).

**How metacognition was taught.** In a recent study (Dignath & Büttner, 2018) of
teacher promotion of learning strategies including metacognition, cognitive strategies,
and motivation, 12 primary and 16 secondary teachers were videotaped while teaching
and these tapes later coded using an instrument designed for the purpose. This instrument
allowed teachers to be rated on a standardized scale and their teaching methods
compared. In addition, nine of the secondary teachers volunteered to be interviewed to
gain more information on how they perceived their own practice. These interviews were
also coded for analysis. The coders of the videos and interviews were trained, and their
results checked against each other to achieve high levels of reliability.
After analyzing the results, little direct promotion of learning strategies was observed. At the primary level teachers did occasionally promote cognitive strategies, but metacognition and motivation were very infrequently addressed. Secondary teachers spent more time promoting metacognition. A quarter of them spent more than five minutes of a standardized 45-minute lesson directly teaching metacognitive strategies.

Primary teachers indirectly promoted learning strategies more frequently than secondary teachers ($F(1,26) = 16.63; p < .01; d = 1.56$) with cognitive strategies most frequently supported by learning tasks. Secondary teachers did little to indirectly promote any learning strategies, in fact the researchers noted the teacher-directed classrooms they observed did not appear to offer student many chances to construct knowledge.

Teachers’ perceptions at the secondary level as revealed by the interviews did not align with video evidence. While teachers were generally in favor of teaching the strategies and reported that they had been promoting them, there was little evidence in the videos. Correlations between observed and reported teaching of strategies were $r = -.40$, ($p = .27$) for cognitive, $r = .41$, ($p = .31$) for metacognitive, and $r = .02$, ($p = .97$) for motivational.

To explain the disconnect between positive teacher beliefs about learning strategies and low implementation, Dignath and Büttner hypothesized, the low levels of promotion observed were due to a lack of training in how to teach the strategies (Dignath & Büttner, 2018). Dignath and Büttner (2018) called for such training to be made part of the teacher preparation curriculum. They also highlighted the importance of teacher perceptions in shaping practice.
While this study was carefully designed, the sample size was small, and those teachers sampled were volunteers. In addition, only nine of the secondary teachers and none of the primary teachers took part in the interviews. This limited the generalizability of the results to teachers at large because it was possible that only teachers with a positive attitude about learning strategies took part in the study.

**Metacognition, academic achievement, and intelligence.** Ohtani and Hisasaka (2018) conducted a meta-analysis of 118 articles that reported correlations between metacognition and academic achievement and included a measure of intelligence. They did not attempt to discriminate between metacognitive knowledge and metacognitive activities, because they assumed that in order to activate the activity, one must possess the knowledge. The authors did distinguish between online and offline measures of metacognition; favoring online measures as slightly more accurate.

After combining the effect sizes of the articles, Ohtani and Hisasaka (2018) reported a moderate correlation between metacognition and academic achievement when controlling for intelligence ($r = .28$, 95% CI $[0.24, 0.31]$, $p < .001$). They concluded that intelligence was a confounding variable. Individuals with higher intelligence tended to process information rapidly, which might have freed up extra mental capacity for metacognition (Ohtani & Hisasaka, 2018).

A limitation of this study was the authors’ choice to exclude students and adults with disabilities. During the 2015-16 school year, 13% of all students age 3-21 enrolled in U.S. public schools received special education services (McFarland et al., 2018). Excluding students with disabilities meant a significant segment of the population was not included in the study.
**Teachers’ thoughts.** Bannister-Tyrrell and Clary (2017) conducted a qualitative study by surveying four teacher candidates and nine experienced public-school teachers of English in Australia. While initially concerned that simple definitions of metacognition had underrepresented its importance in learning, they found that the teachers as a group had a fairly sophisticated understanding of what metacognition was and could articulate why it was a needed skill. However, these same teachers did not often actively teach strategies or otherwise engage students in metacognitive activities. Many expressed that because metacognition was not explicitly written into the curriculum, they were either uncertain of how to promote it, or if they even should, given other priorities. As they thought of ways to add reflective practices into the curriculum one teacher warned against simply handing out forms for students to fill out because then students would not truly think about their learning (Bannister-Tyrrell & Clary, 2017).

The goal of a qualitative study such as the aforementioned, was to provide a thick description of practice. Unlike quantitative work there was no hypothesis to prove or discard (Creswell & Poth, 2018). Steps such as utilizing multiple observers (Gall et al., 2007), and triangulation (Carter, Bryant-Lukosius, DiCenso, Blythe, & Neville, 2014) could assist in minimizing error and serve to increase confidence in results. Because Bannister-Tyrrell and Clary (2017) did not provide a description of these measures, the conclusions that could be drawn from this work were limited. Nevertheless, the questions asked remained important because teacher perceptions shaped their approach to teaching (Black & Wiliam, 2009).

**Metacognition and confidence.** Weight (2017) studied 171 elementary and secondary teachers and support staff. In this correlational study, staff members who used
metacognitive instructional strategies reported greater confidence in their ability to work with students who experienced anxiety ($\chi^2 (1, N = 171) = 20.93, p < .05$) on a self-report, Likert-type survey (Weight, 2017).

The sample of surveyed teachers was large and representative of both primary and secondary teachers with equal distribution of a wide range of years of experience. There was not an explanation of how the sample pool was formed. If the training was elective, this might have influenced the type of teachers who participated in the survey.

In addition to the survey, a small group of teachers were also interviewed to gather qualitative insights on their use of metacognitive strategies and the extent to which these strategies increased their confidence in working with students who experienced anxiety. The results from this interview helped confirm the quantitative correlations and shed additional light on the benefits of using metacognitive strategies in the classroom.

As the author noted, surveys were limited by the honesty of the participants (Weight, 2017). Even when respondents were presumably as forthright as possible, there was a known lack of fit between teacher report of metacognitive promotion and actual practice (Dignath & Büttner, 2018) which called into question the accuracy of the results of the self-report measure.

As the author also noted, the study was correlational and non-experimental in nature thus, could not establish causation (Weight, 2017). In a recent synthesis of 1200 meta-analyses of studies of influences on achievement in higher education, John Hattie (2015) found collective teacher efficacy to have the greatest effect. Regardless of the direction of the cause, teachers who believed they had the ability to help students learn, often did (Hattie, 2015).
**Metacognition and academic achievement.** Dent and Koenka (2016) conducted a meta-analysis to examine the correlation between metacognitive process use and academic achievement and correlation between cognitive strategy use and academic achievement. Academic achievement was measured by standardized measures. Only studies that included enough data to allow correlations between one or both of the independent variables and achievement to be calculated were included.

For the first correlation they located 61 studies of elementary and secondary students containing 490 correlations between academic performance and student use of metacognition (Dent & Koenka, 2016). When averaged, academic performance was moderately correlated with metacognition ($r = .20, 95 \% \text{ CI}[.16, .24], p < .001$) (Dent & Koenka, 2016).

For the second correlation they located 57 studies of elementary and secondary students containing 343 correlations between academic performance and student use of cognitive strategies. When averaged, academic performance was mildly correlated with cognitive strategies ($r = .11, 95 \% \text{ CI}[.08, .14], p < .001$) (Dent & Koenka, 2016).

The authors identified and evaluated five moderating variables from the studies they reviewed. These included what metacognitive process was used, what cognitive strategy was used, which academic subject it was applied in, the grade level, and how the use of strategy and resulting achievement were measured. Because many studies had grouped planning, goal setting, self-monitoring, self-control, and self-evaluation into one category, the researchers looked at these separately. They found planning to have the strongest correlation with achievement when considered independently.
Of all subjects, success in social studies had the strongest correlation with metacognitive processes. Of all grades, achievement at middle school had the strongest correlation with metacognitive processes. Studies that used online measures of metacognition reported the highest correlation with academic achievement. Possibly because offline measures were less accurate and may have missed metacognitive activity that was taking place.

The correlations found between cognitive strategy use and academic achievement were strongest in science classes. Strategies that promoted deep processing of material over simple memorization were also more strongly correlated with academic achievement. Like metacognitive processes, cognitive strategy and success were correlated with middle school more than elementary, but most strongly correlated with high school.

The researchers reminded the reader that correlation did not imply causation and acknowledged that while there was no implied direction for the effects in this study, they found theoretical support to believe the relationship between metacognitive process use, cognitive strategy use, and academic achievement were reciprocal. The researchers concluded that “while cognitive strategies help students learn, metacognitive processes ensure that they have done so” (Dent & Koenka, 2016, p. 459).

**Dissertation synthesis.** Bond, Denton, and Ellis (2015) examined the impact on student learning as a result of reflective self-assessment as documented in 10 doctoral dissertations. These dissertations reported results from a broad array of classes including math, science, world languages, English Language arts, social studies and geography. In each study, students were asked to participate in reflective activities towards the end of
the period. Teacher feedback was a part of the intervention in six of the studies. A positive effect size using Cohen’s $d$ was documented for posttest scores in seven of the studies, while three studies showed a negative effect size. The resulting weighted mean effect size was 0.28 for the posttest, with a range of -0.34 to 0.69 (Bond et al., 2015).

Bond et al. (2015) clearly defined their methodology including the criteria for inclusion of studies, information about each study, and how effect sizes were calculated and combined. A limitation of this synthesis was that all studies were conducted at one institution, generally representing the public schools of one geographical area, with an exception. To strengthen these findings studies including students from other types of schools and locations could be added.

**How to prompt metacognition.** In a quasi-experimental study of 70 college students, Bannert and Mengelkamp (2008) looked for a difference in learning outcomes between three groups who had completed the same computer-based course. There was a group that used a think-aloud method, a group that used a reflect when prompted method, and a group that did not engage in reflection. In the think-aloud condition students were asked to talk about their thoughts as they solved problems. This verbal record was recorded and coded for themes. The authors noted the previously mentioned limitation of this method: some mental processes cannot be put into words.

In the second condition students were asked to stop and reflect on their learning when prompted, the prompts were strategically located at the end of a section to minimize disruption to the flow of learning. The researchers hypothesized that this group would make the most learning gains. In the final condition students completed the same learning but were not asked to engage in reflection in order to serve as a control group.
After learning the material all groups completed a multiple-choice test to measure knowledge of the subject, they also completed a questionnaire to measure perceived disorientation and disruption during learning. Transfer of learning was also measured by asking students to apply what they had learned to new situations. ANOVA did not show any statistically significant difference in knowledge outcomes between groups, however the reflect when prompted group did have higher ability to transfer learning compared to the control group, \( t(1,44) = 3.64, p < .05, d = 0.55 \).

Based on the questionnaires, the researchers concluded that the control group perceived the fewest disruptions to learning. Further, the think aloud procedure did provide a slight advantage over stop and reflect in that it was less distracting. In addition, the students who used think aloud reported higher feelings of strategic learning than the control group (Bannert & Mengelkamp, 2008).

Because this study was quasi-experimental, there was a major threat to internal validity. While the assignment of condition was random by group, the sample was not randomly selected, nor were individuals randomly assigned to groups, this meant differences between groups could have been due to preexisting conditions and not the treatment (Gall et al., 2007). The authors did attempt to address this limitation by comparing the groups based on prior knowledge, metacognitive knowledge, verbal intelligence, and motivation to ensure meaningful comparisons could be made between them.

**Learning science.** Georghiades (2004b) conducted an experiment with students in Year Five. The average age of these students was 11. After placing 60 students evenly into two groups, one group received metacognitive instruction as part of the regular
classroom activities, the other group did not. With this exception both groups received the same instruction on concepts in the Current Electricity unit. Scores from the previous year’s science exam and a general thinking ability exam were included in the analysis to ensure a valid comparison between groups could be made. Following four 80-minute lessons, both groups were assessed three times on their understanding of scientific concepts related to the lessons. The same assessment was used each time. The test was given a week, two months, and eight months after the unit concluded. The groups were initially close in mean scores, however over time the experimental group retained more information as evidenced on the final administration of the exam ($p = .048$) (Georghiades, 2004b).

The statistical test used in this study provided a reason for readers to interpret the results with caution. The researcher relied on three t-tests to analyze the data. The use of multiple t-tests inflated the chance of Type I error (Field, 2013). A more conservative approach would have been to use ANOVA with Bonferroni adjustment (Tabachnick & Fidell, 2007).

Additionally, the researcher adjusted instruction based on reading student diaries. Other researchers (e.g. Bianchi, 2007) have pointed out the possible differential effect of teachers reacting to student reflections. If one group benefited from improved instruction based on their expressed needs and another did not, this could have confounded interpretation of results. The researcher did not explicitly state if only the treatment group benefited from this adjustment, providing another reason to interpret results of this study with caution. Although there were flaws in the study, it was the type of situated inquiry
that relied on methods beyond student self-report and was needed to add to our knowledge of the effects of metacognition (Dinsmore et al., 2008).

**Gains from reflection.** In a study of 33 college students enrolled in a computer programing course, Kwon and Jonassen (2011) examined the effect of student reflection on students’ ability to relearn areas of the curriculum they had initially not understood. To begin, all students took a pretest, immediately followed by a chance to see their correct and incorrect answers. Students were asked to reflect on why they thought they had either correctly or incorrectly answered each question. Students later completed a three-part posttest.

Students were classified into four groups based on results of the pretest and the number of times they reflected on their answers. Researchers then classified students as having high prior knowledge and either high or low frequency reflection and low prior knowledge with either high or low frequency reflection. Researchers then looked at what kinds of reflections students had generated. They found three main kinds of reflections: those that confirmed prior knowledge, those that corrected it, and reflections of puzzlement.

A MANOVA was then conducted using the two factors: frequency and prior knowledge, high and low, as independent variables and the type of reflection generated as the dependent variable. Those students with high prior knowledge also reflected more (Wilks’ \( \lambda = .58, F(3, 27) = 6.64, p = .002, \eta^2 = .425 \)) (Kwon & Jonassen, 2011).

Those students who initially had low knowledge of the subject but engaged in high levels of reflection were able to perform as well on the posttest as the high knowledge students who did not reflect much. The researchers attributed this gain on the
part of the low knowledge students to their ability to learn from their errors via reflection (Kwon & Jonassen, 2011).

While the number of subjects in this study was small, the questions were important. Replicating this type of fine-grained study could strengthen the findings and help answer Brown’s (1987) call for methodical research in metacognition.

**Empirical Studies Influence on Current Study**

Students needed some prior knowledge of a topic in order to attain maximum benefit from reflection on their learning (Kwon & Jonassen, 2011). Additionally, students who were new at applying MR might have been less able to learn information when asked to reflect while learning due to the need to split processing power between two tasks (Ohtani & Hisasaka, 2018). Learning was also hindered if students felt interrupted by the prompts to reflect (Bannert & Mengelkamp, 2008). For these reasons, in the current study the reflection took place immediately following the learning, soon enough that students could still remember what they felt while learning, but not during learning to prevent being distracted by the reflection.

The implementation of metacognitive practice in the classroom was a two-edged sword, it must have been simple enough that teachers could easily implement it in the face of competing priorities, but if it was reduced to a simple checklist activity its efficacy was lost (Bannister-Tyrrell & Clary, 2017). In the current study, the investigator in the role of classroom teacher needed a simple activity that did not require much time to teach or implement, but also had a good chance of causing students to engage in authentic reflection. The reflective assessments Ellis and Denton (2010) published met these
criteria. The way the assessments were presented, as described in the method section was
designed to further increase their effectiveness.

Bianchi’s (2007) notation of the possible differential effect of teachers reacting to
student reflections as a weakness in earlier studies called the inclusion of feedback in the
study current study into question. To avoid this, one could have chosen to not read the
student reflections and remove the feedback aspect from the study entirely. However,
there was a way to provide feedback, but avoid teacher adjustment of lessons. Feedback
could be provided by a third party, as in Baliram and Ellis’ (2019) study. While this
approach eliminated the differential effect of the teacher adjusting instruction based on
feedback, it was rejected in the current study because it also seemed to reduce the
likelihood that a student would respond to the feedback because students might not have
trusted an unknown third party to provide guidance (Hattie & Clarke, 2019).

In the current study the investigator acting in the role of classroom teacher
adjusted instruction for all classes based on feedback received from the reflection groups.
The investigator assumed that groups who had been taught using the same methods
would have similar questions and the adjustments would benefit all groups, although
perhaps not equally. This adjustment based on student questions may have inflated the
comparison group outcomes, possibly obscuring some of the effects of the intervention.

The ability to synthesize results based on multiple studies as done by Bond,
Denton, and Ellis (2015), Hattie, (2015) and Dignath & Büttner, (2008) was intended to
increase the readers’ confidence in the data presented and the conclusions that could be
drawn from it. These types of meta-analysis relied on effect sizes. An effect size provided
a way to express the practical significance of a study (Ellis, 2010) and compare studies
that used different methods of data analysis. After conducting numerous meta-analyses Hattie (2012) cited anything over .40 as a worthwhile effect size for an academic intervention. The synthesis approach (Bond et al., 2015) to data analysis informed the weighted mean effect sizes calculated in the current study.

Gall et al. (2007) noted the threat to internal validity presented by intact groups. Without random assignment of subjects to conditions, any effect detected in the study might have been due to preexisting differences between the groups instead of the treatment. Studies by Georgiades (2004b) and Baliram and Ellis (2019) provided examples of the use of pretest results to mitigate this threat.

In the current study, students who were skilled in learning academic language may have performed well regardless of condition. While obtaining general information on the students’ academic ability as Georgiades (2004b) did was beyond the scope of the current study, the investigator did consider this as a limitation and included a pretest to maximize internal validity.

A retention test of science concepts (e.g. Georgiades, 2004b) provided evidence of a long-term benefit to learning resulting from metacognitive strategy use. To test for this long term-benefit in learning academic language related to perspective drawing, a retention test was included in the current study. The investigator expected a student in any condition to improve from pretest to posttest due to learning from the classroom teaching and practice, but the ability to retain or increase said score after the posttest might be aided by reflection during the initial learning. The retention test allowed this added benefit to be documented.
Chapter 3: Method

Introduction

This chapter describes the methods used in the study. Methods include a description of the design of the study, information about participants in the study, and how they were selected. There is an explanation of the procedures developed for teaching and assessing the learning. The descriptive and inferential data analysis processes used to investigate the results of the assessments are also reported in detail.

Research Design

This study was designed to test the following hypothesis:

$H_0$: There is no statistically significant difference based on group (three levels: reflection with feedback, reflection, comparison) on students’ ability to learn and retain academic language related to perspective drawing as measured by mean score differences on a multiple-choice test.

$H_a$: There is a statistically significant difference based on group (three levels: reflection with feedback, reflection, comparison) on students’ initial ability to learn and retain academic language related to perspective drawing as measured by mean score differences on a multiple-choice test.

The research design used in this study was quasi-experimental using intact classes taught by the investigator (see Table 1).
Table 1

*Quasi-Experimental Design*

<table>
<thead>
<tr>
<th>Group</th>
<th>Pretest</th>
<th>Intervention</th>
<th>Posttest</th>
<th>Retention Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reflection with Feedback</td>
<td>O</td>
<td>X₁</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>Reflection</td>
<td>O</td>
<td>X₂</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>Comparison</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>

The sample was a convenience sample consisting of students enrolled in the investigator’s semester length middle school visual arts classes. To overcome the reduced internal validity of intact groups, three iterations of the study: Spring 2018, Autumn 2018, and Winter 2019, were conducted with different groups over several terms. This replication strengthened the conclusions that could be drawn from the findings.

The intervention portion of the studies took place during a ten-day perspective unit, beginning with a pretest on academic language related to perspective drawing being administered to all classes on day one. Students in all conditions were then instructed using a variety of methods. At the end of the unit, all groups completed the same questions as a posttest. Three weeks after the posttest the same exam was administered as a retention test. Scores on these tests were compared based on the groups’ condition to determine the effects of reflection, feedback, and time of test, in a repeated measures ANOVA (Field, 2013).

Because there was a wide range of ages and abilities represented in each class, students with prior art knowledge might have performed better on the assessments due to
their prior knowledge and not due to their assigned condition, or for other reasons not addressed by the study. The pretest was an attempt to mitigate this threat to internal validity. However, with the pretest came the possibility of what Gall et al. (2007) referred to as pretest sensitization. This occurred when students’ exposure to material on the pretest affects their response on later tests. Because these effects were most prominent on Likert type tests (Gall et al., 2007) which the test was not, and the same test was administered to each group in the same way, the threat to internal validity was mitigated to the extent possible.

**Participants and Sampling Process**

**Demographics.** The students sampled were enrolled in the investigator’s middle school visual arts class. This school was located in a city in King County and was part of a K-12 public school district that served approximately 17,000 students. According to the most recently published demographic data at the time of the study, 48.8% of students in the district were female while 51.2% were male. The district reported 1.2% of students enrolled were American Indian/Alaskan Native, 8.9% Asian, 7.3% Black/African American, 29.7% Hispanic/Latino of any race(s), 4% Native Hawaiian/Pacific Islander, 39.2% White, and 9.7% Two or More Races. Approximately 18.9% of these students were English Language Learners, 12% received special education support, and 51.8% of the students qualified for free or reduced lunch. The district had an Unexcused Absence Rate of .69%.

The middle school in which the studies were conducted served approximately 790 students in grades six through eight. Of these students, 51.4% were female while 48.6% were male. The school records indicated .5% of students enrolled were American
Indian/Alaskan Native, 8.1% Asian, 10% Black/African American, 26.6% Hispanic/Latino of any race(s), 5.2% Native Hawaiian/Pacific Islander, 40% White, and 9.2% Two or More Races. Approximately 14.8% of these students were English Language Learners, 12.5% received special education support, and 59.4% of the students qualified for free or reduced lunch. The school had an Unexcused Absence Rate of 1.81%.

**Assignment of condition.** To remain objective, the investigator flipped a coin to decide which condition each class would receive. The assignment of condition took place before the unit commenced. The first coin flip determined MR intervention or comparison. For classes assigned to the MR condition, the second flip assigned reflective assessment with or without teacher feedback.

**Sample size.** To ensure the number of participants in each study was large enough for the statistical test to detect an effect if it existed (Gall et al., 2007), an a priori power analysis was conducted using G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007). This program calculated required sample size based on investigator determined power level, significance level, and the population effect size expected to be found at a set probability. For this analysis the population effect size was set to 0.3 (Ellis, 2010) (see Appendix A). Output from the analysis indicated a total sample size of 75 required with \( p < .05 \) and a power level of .8 (Lakens, 2013).

**Description of samples.** Five art sections were included in each study. Students ranged in age from 11-15. The demographics of each class largely mirrored the overall school demographics with one exception, noted in Chapter 5. A majority of students enrolled in Art One were in sixth grade and a majority of Art Two students were in
seventh or eighth grade. The classes averaged 25 students, of whom 20.6 participated in the study, on average.

Students in these classes had widely varying levels of past art instruction and skill. A small percentage of students reported receiving regular art instruction in elementary school, while many received sporadic, or none. There was no systematic method of tracking students’ past art experience at the middle school, potentially an eighth grader could have been in their first art class while a seventh grader in the same class could have successfully completed an art class in sixth grade and had some elementary art experiences. The sample of five art classes that included students of various ages and experience levels taught by one teacher, while not ideal for external validity, was the best available for this study. Using intact groups was common in studies involving public school students (Gall et al., 2007).

**Spring 2018 sample.** In the first study conducted in spring 2018, three classes were assigned to reflective assessment (see Table 2).
### Table 2

**Spring 2018 Sample**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Section</th>
<th>Class Period</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison</td>
<td>Art 1</td>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td>Reflection with Feedback</td>
<td>Art 1</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>Reflection with Feedback</td>
<td>Art 2</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>Comparison</td>
<td>Art 2</td>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>Reflection only</td>
<td>Art 1</td>
<td>6</td>
<td>21</td>
</tr>
</tbody>
</table>

This sample included 102 of the 119 students assigned to the investigator in the role of classroom teacher at the time of the study. Of these, 43 students were in the comparison group, 21 in the reflection only group, and 38 in reflection with feedback group. The 12 students who missed the pretest and were subsequently dropped from the study were evenly distributed among all five classes. A further four students missed three or more lessons and were dropped from the study. Three of these students had been in the reflection with feedback group and one had been in the reflection only group. A final student in the reflection with feedback group refused to take the posttest and was also dropped from the study.

**Autumn 2018 Sample.** In the second study conducted in autumn 2018, three classes were assigned to reflective assessment (see Table 3).
This sample included 94 of the 130 students assigned to the investigator in the role of classroom teacher at the time of the study. Of these, 34 students were in the comparison group, 23 in the reflection only group, and 37 in reflection with feedback group. The 17 students who missed the pretest and were subsequently dropped from the study were not evenly distributed among the three groups, 7 students had been in the comparison group, 2 in the reflection only group, and 8 in the reflection with feedback group. A further 19 students missed three or more lessons and were dropped from the study. Ten students had been in the comparison group, 3 in the reflection only group, and 6 in the reflection with feedback group.

**Winter 2019 sample.** In the third study conducted in winter 2019 three classes were assigned to reflective assessment (see Table 4).
This sample included 112 of the 131 students assigned to the investigator in the role of classroom teacher at the beginning of the study. Of these, 45 students were in the comparison group, 42 in the reflection only group, and 25 in the reflection with feedback group. The five students who missed the pretest and were subsequently dropped from the study were evenly distributed among the three groups. Four students had elected to take the semester length art class for a second term. Because they participated in the Autumn 2018 study, they were provided alternate assignments and not included in the Winter 2019 study. A further seven students missed three or more lessons and were dropped from the study. Of these, one student had been in the comparison group, five in the reflection only group, and one in the reflection with feedback group. In addition, three students transferred to other classes during the study and were also dropped.

**Protection of participants.** This study involved typical classroom instruction and assessment procedures, which did not require informed consent from participants (see
Appendix B). At the advice of the Director of Doctoral Programs in which the investigator was enrolled, a letter was sent home to students’ parents or guardians informing them about the study. The investigator protected the privacy of participants’ data by only reporting scores that could not be linked to individual students. Additionally, raw data was kept in secure locations and destroyed at appropriate intervals as prescribed by Washington State Administrative Code.

Students did not receive compensation for participating in this study. However, the benefits of the study to students in the treatment group could have been a better long-term retention of the academic language related to perspective drawing. All students likely benefited from targeted instruction made possible by the investigator in the role of classroom teacher reacting to student written reflections. Participating in this study posed no risks to students or the investigator in either their role as investigator or classroom teacher.

Measures

The first study relied on a teacher generated thirty-question multiple choice test of academic language related to perspective drawing. This test was developed by the investigator in the role of classroom teacher as part of the regular curriculum and was similar to tests administered by other art teachers in the district. While Gall et al. (2007) noted the limitations of teacher generated tests, after searching, no suitable standardized measure of the academic language relating to perspective was located. Thus, additional measures were undertaken to ensure the test was appropriate for use in research.

Reliability. The test was examined for reliability by generating split-half reliabilities using posttest scores. If the two halves were highly correlated, the measure
was reliable (Field, 2013). A value above .7 indicated that the instrument was consistently measuring the same factor (Vogt & Johnson, 2011). Spearman's rho correlations between the halves of the Spring 2018 test administration were .81 indicating a reliable measure.

Pretest scores may have been more accurate for calculating this metric in that they were not affected by the treatment (Glass, McGaw, & Smith, 1981). However, the investigator chose not to base this test on the performance of students who generally had minimal exposure to the topic. Instead, scores from students who had some knowledge of the material were used.

**Validity.** Construct validity was assessed by comparison to similar measures in published art curricula and inclusion of academic language listed in state and national visual art standards. Content validity was attained through a review by a group of art teachers teaching similar ages.

During a discussion with the Dissertation Chair following the Spring 2018 test administration, a potential ceiling effect was noted in the results. This effect was indicated by score distributions on the posttest with negative skewness (see Appendix C) (Ho & Yu, 2015). A ceiling effect might have prevented proper data analysis (French, Sycamore, McGlashan, Blanchard, & Holmes, 2018). To reduce this effect, the investigator added an additional six questions of greater difficulty bringing the total to 36 (see Appendix D).

One way to determine the difficulty of a test was to divide the raw mean score by the maximum possible score, if the value was greater than .5 the test was said to be “easy” (Ho & Yu, 2015). The Spring 2018 administration of the test resulted in a value of
.67 on the posttest. For the second and third study following the additional questions, values were .62 and .64 respectively, indicating a slight decrease in easiness.

The same measures to assure validity and reliability used with the original version were conducted on the revised test. Spearman's rho correlations between the split-halves of the Autumn and Winter test administration were .80 and .71 respectively, indicating a reliable measure.

**Procedure**

For each iteration, at the beginning of the ten-day perspective unit, a pretest on academic language related to perspective drawing was administered to all classes on day one. Students in all conditions were then instructed over the course of the following eight, 56-minute class periods using a variety of methods including teacher modeling, note-taking, guided practice, independent practice and group discussion. At the end of the unit on day ten, following a review, all groups completed the same questions as a posttest. Three weeks after the posttest, the same exam was administered as a retention test.

The investigator in the role of classroom teacher used the same instructional methods with all classes, except the classes assigned to the comparison group did not complete reflective assessments, but instead spent the final five minutes of class on instructional days in guided or independent practice. For the classes assigned to reflective assessment, students engaged in a four to five-minute reflective activity. These took place on instructional days, toward the end of class, for a total of nine reflective sessions.

During these sessions, students were asked to complete a short reflective assessment of the day’s learning such as an I Learned statement, a Key Idea Identification, or a Clear and Unclear Windows (Ellis, & Denton, 2010) (see Appendix
The MR prompt used was varied from day to day, so students would not lose interest (Georghiades, 2004b). The assessments were also presented in a dynamic format to keep students engaged in authentic assessment and reduce the chances students simply viewed this as one more activity to check off the list (Bannister-Tyrrell & Clary, 2017). Some days students were asked to write their reflections a sticky note and post this on the board for the investigator in the role of classroom teacher to read (see Appendix F). Other times the reflection was shared verbally between student and teacher.

In the reflection with feedback condition, the investigator in the role of classroom teacher individually responded to each student’s reflection with a short note or verbal comment related to what they wrote as soon as possible (Slinger-Friedman & Patterson, 2016). Because exact timing of the delivery of the feedback was controversial (Shute, 2008), some delayed feedback was also provided. When an obvious theme in student responses emerged, the investigator in the role of classroom teacher communicated this to the entire class (Hattie & Clarke, 2019), often as a way of introducing the following day’s lesson. On the day before testing, the reflection with feedback group was encouraged to share their Unclear Windows so the investigator in the role of classroom teacher could provide feedback on how to remedy any lingering misunderstanding.

By reading and responding to specific statements about the day’s learning, the intervention remained focused on the most beneficial task-type feedback (Hattie & Timperley, 2007). Data from these statements was used to guide the next steps of instruction (Black & Wiliam, 2009). Reflection only classes did not receive feedback on
their reflective assessments apart from the investigator in the role of classroom teacher thanking them for completing it.

Students who were absent for the pretest were offered a chance at the beginning of the following class session to complete the assessment before instruction began. Students who were absent, or unable to complete the pretest during this time were not included in the study. These students’ attendance in class during the lessons and then subsequent completion of the pretest would have skewed the results. Students who missed significant class time, in this case three or more lessons out of the ten-day unit, were also dropped from the study.

**Statistical Analysis**

Many years ago, Pillemer (1991) called for the merit of two tailed hypotheses to be considered in educational research. Recently, Field (2013) concluded that two tailed hypothesis and two-tailed statistical tests were almost always more appropriate in research. For these reasons the hypothesis in this study was written in two-tailed form. Because ANOVA had a lower chance of Type I error than multiple t-tests (Field, 2013), allowed post-hoc testing with a Bonferroni adjustment (Tabachnick & Fidell, 2007), and could properly test two-tailed hypothesis (Cho & Abe, 2013), a repeated measures ANOVA was used in this study. This test determined if the difference in mean values of the three groups was more likely due to chance or other causes. Repeated measure meant the same entities provided data at multiple time points (Field, 2013). Because the same entities were measured multiple times, the two sources of variance were the effects of the treatment or manipulation, and individual differences in performance. ANOVA generated an F-ratio using this assumption (Field, 2013). If the test was statistically significant, one
could reasonably conclude that differences between groups existed and reject the null hypothesis.

**Descriptive.** Mean scores and standard deviations for each group at each test time were calculated. Data was then checked for the usual assumptions of the general linear model including skewness and kurtosis (Vogt & Johnson, 2011). To conduct this test, scores of skewness and kurtosis were converted to z-scores by dividing by their standard error and comparing their absolute value to 1.96. Scores greater than 1.96 were statistically significant at the \( p < .05 \) level (Field, 2013).

To conduct the repeated measures ANOVA, student scores on the academic language test were entered into SPSS Version 25 software. Mauchly’s test was used to see if the assumption of sphericity was violated, if so the Greenhouse-Geisser values were interpreted (Field, 2013). Any missing scores on post or retention test were replaced with a mean substitution.

**Inferential.** A repeated measures ANOVA was used to analyze scores based on whether they came from a group engaged in MR or not, and presence or absence of teacher feedback on the students’ reflections. As such, there was one within subjects’ factor: time of test, with three levels: pretest, posttest, and retention test. There was one between subject factor: group, with three levels: reflection with feedback, reflection only, and comparison. The level of statistical significance for this analysis was set at \( p < .05 \).

A post-hoc test with a Bonferroni adjustment post-hoc test was conducted. Post-hoc comparisons allowed the investigator to determine the direction and magnitude of differences based on group (Field, 2013). A Bonferroni adjustment was used to reduce
chances of a Type I error or detecting an effect when there was not one (Tabachnick & Fidell, 2007).

In addition to tests of statistical significance the investigator calculated effect sizes pretest to posttest and pretest to retention test for each study using Cohen’s $d$. An effect size provided a way to express the practical significance of a study (Ellis, 2010). Pretest to posttest comparisons showed which group had higher initial gains while, pretest to retention test comparisons showed which group better retained these gains (Little, 1960).

Effects by condition were also combined to compare overall results. Borenstein, Hedges, Higgins, and Rothstein (2009) proposed using a fixed-effect model to calculate mean weighted effect sizes if two conditions were met: studies were very similar, and the object was to calculate effect sizes for the population represented in the studies only, not to generalize to other groups. The current study satisfied these conditions. The method used in this study to calculate weighted mean effect size involved multiplying each studies’ effect size by the sample size of that study, adding these together and then dividing by the combined sample size of all three studies (Ellis, 2010).
Chapter 4: Results

Introduction

This chapter reports the results of three related studies. Data from each iteration of the study is presented in turn including the relevant descriptive and inferential information the hypothesis. Weighted mean effect sizes pretest to posttest and pretest to retention test are displayed in chart form for ease of comparison within and across the three repetitions of the intervention.

Hypothesis

The data presented provided evidence for evaluating the following hypothesis:

H₀: There is no statistically significant difference based on group (three levels: reflection with feedback, reflection, comparison) on students’ ability to learn and retain academic language related to perspective drawing as measured by mean score differences on a multiple-choice test.

Hₐ: There is a statistically significant difference based on group (three levels: reflection with feedback, reflection, comparison) on students’ initial ability to learn and retain academic language related to perspective drawing as measured by mean score differences on a multiple-choice test.

Spring 2018

Descriptive statistics. In the Spring 2018 study, all groups made gains between each test (see Table 5), except for the reflection with feedback group which plateaued between post and retention test.
Table 5

*Spring 2018 Descriptive Statistics*

<table>
<thead>
<tr>
<th>Time</th>
<th>Condition</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td>Comparison</td>
<td>13.36</td>
<td>3.86</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Reflection</td>
<td>12.43</td>
<td>3.61</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Reflection with Feedback</td>
<td>14.61</td>
<td>5.56</td>
<td>38</td>
</tr>
<tr>
<td>Posttest</td>
<td>Comparison</td>
<td>18.44</td>
<td>6.71</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Reflection</td>
<td>19.10</td>
<td>5.33</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Reflection with Feedback</td>
<td>22.26</td>
<td>4.78</td>
<td>38</td>
</tr>
<tr>
<td>Retention</td>
<td>Comparison</td>
<td>20.66</td>
<td>6.16</td>
<td>43</td>
</tr>
<tr>
<td>Test</td>
<td>Reflection</td>
<td>22.42</td>
<td>4.34</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Reflection with Feedback</td>
<td>22.21</td>
<td>5.14</td>
<td>38</td>
</tr>
</tbody>
</table>

Students’ scores on the academic language test were entered into SPSS Version 25 software. Data was checked for the usual assumptions of the general linear model, including skewness and kurtosis (Vogt & Johnson, 2011). There was statistically significant: positive kurtosis in the pretest scores, negative skewness in the posttest, and both negative skewness and positive kurtosis in the retention test (see Appendix C). The investigator concluded the non-normality was based on real data and not errors. The non-normality in pretest data was likely caused by many students with limited knowledge of
test items prior to the unit. The non-normality in posttest and retention test data was likely caused by a combination of many students making gains due to learning and a somewhat easy test. These trends appeared in later studies which helped to confirm this reasoning. Because non-normality could be offset by the sample size, and this study involved 102 subjects, the investigator proceeded with statistical testing (Field, 2013).

**Inferential statistics.** A repeated measures ANOVA was used to analyze scores based on whether they came from a group engaged in MR or not, and presence or absence of teacher feedback on the students’ reflections. Mauchly’s test confirmed the assumption of Sphericity was not violated ($p = .27$).

There was a statistically significant within-subject interaction effect between time of test and condition ($F(4, 198) = 2.66, p = .03$). However, a Bonferroni adjustment revealed group score differences were not statistically significant when compared across condition (see Appendix G).

**Autumn 2018**

**Descriptive statistics.** In the second study conducted in autumn 2018, all groups made gains between each test (see Table 6).
Table 6

*Autumn 2018 Descriptive Statistics*

<table>
<thead>
<tr>
<th>Time</th>
<th>Condition</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td>Comparison</td>
<td>15.68</td>
<td>4.80</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Reflection</td>
<td>18.65</td>
<td>7.09</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Reflection with Feedback</td>
<td>16.62</td>
<td>3.90</td>
<td>37</td>
</tr>
<tr>
<td>Posttest</td>
<td>Comparison</td>
<td>21.09</td>
<td>7.26</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Reflection</td>
<td>24.00</td>
<td>8.09</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Reflection with Feedback</td>
<td>22.27</td>
<td>6.26</td>
<td>37</td>
</tr>
<tr>
<td>Retention</td>
<td>Comparison</td>
<td>24.09</td>
<td>7.76</td>
<td>34</td>
</tr>
<tr>
<td>Test</td>
<td>Reflection</td>
<td>26.35</td>
<td>8.30</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Reflection with Feedback</td>
<td>24.58</td>
<td>6.01</td>
<td>37</td>
</tr>
</tbody>
</table>

The reflection only group started with the highest pretest mean scores and maintained this lead for the following two assessments.

Students’ scores on the academic language test were entered into SPSS Version 25 software. Data was checked for the usual assumptions of the general linear model, including skewness and kurtosis (Vogt & Johnson, 2011). There was statistically significant positive skewness and kurtosis in the pretest scores (see Appendix H). While there was mild negative skewness and kurtosis in the post and retention tests a simple test
confirmed these to be not statistically significant (Field, 2013). The investigator concluded the pretest non-normality was indicative of subjects having limited knowledge of test items prior to the unit and proceeded with statistical testing (Field, 2013).

**Inferential statistics.** A repeated measures ANOVA was used to analyze scores based on whether they came from a group engaged in MR or not, and presence or absence of teacher feedback on the students’ reflections. Mauchly’s test revealed the assumption of Sphericity had been violated \( \chi^2(2) = 13.96 \ (p < .001) \) so the Greenhouse-Geisser values were interpreted. While all groups made gains between each test, a Bonferroni adjustment revealed group score differences were not statistically significant when compared across condition (see Appendix I).

**Winter 2019**

**Descriptive statistics.** In the third study conducted in winter 2019, all groups made gains between each test, except for the reflection group which plateaued between posttest and retention test (see Table 7).
Table 7

Winter 2019 Descriptive Statistics

<table>
<thead>
<tr>
<th>Time</th>
<th>Condition</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td>Comparison</td>
<td>16.96</td>
<td>4.25</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Reflection</td>
<td>15.21</td>
<td>4.52</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Reflection with Feedback</td>
<td>13.44</td>
<td>3.65</td>
<td>25</td>
</tr>
<tr>
<td>Posttest</td>
<td>Comparison</td>
<td>24.59</td>
<td>5.37</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Reflection</td>
<td>23.50</td>
<td>5.70</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Reflection with Feedback</td>
<td>21.60</td>
<td>5.50</td>
<td>25</td>
</tr>
<tr>
<td>Retention</td>
<td>Comparison</td>
<td>27.72</td>
<td>5.28</td>
<td>45</td>
</tr>
<tr>
<td>Test</td>
<td>Reflection</td>
<td>23.68</td>
<td>6.06</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Reflection with Feedback</td>
<td>23.52</td>
<td>5.57</td>
<td>25</td>
</tr>
</tbody>
</table>

The comparison group started out with and maintained the highest mean scores over the entire study. In contrast, the reflection and reflection with feedback groups made greater gains than the comparison group pretest to posttest. On the retention test, the reflection with feedback group made the largest gain, followed by the comparison group.

Students’ scores on the academic language test were entered into SPSS Version 25 software. Data was checked for the usual assumptions of the general linear model, including skewness and kurtosis (Vogt & Johnson, 2011). There was statistically
significant positive skewness in the pretest scores (see Appendix J). While there was mild skewness and kurtosis in the posttest and retention test a simple test confirmed these to be not statistically significant (Field, 2013). The investigator concluded the pretest non-normality was indicative of subjects having limited knowledge of test items prior to the unit and proceeded with statistical testing (Field, 2013).

**Inferential statistics.** A repeated measures ANOVA was used to analyze scores based on whether they came from a group engaged in MR or not, and presence or absence of teacher feedback on the students’ reflections. Mauchly’s test confirmed the assumption of Sphericity was not violated ($p = .30$). There was a statistically significant between-subject effect based on condition ($F(2, 109) = 7.21, p < .001$). A Bonferroni adjustment revealed group score differences between the comparison group and the reflection group were statistically significant ($p = .03$). Score differences between the comparison group and the reflection with feedback group were also statistically significant ($p = .002$) (see Appendix K).

**Effect Sizes**

To synthesize the results and look for trends in the data within and across all three studies the investigator calculated effect sizes. Pretest to posttest effect sizes showed which group had higher initial gains (see Table 8). These effects were also pooled as weighted mean effect sizes to compare overall results.
Table 8

Pretest to Posttest Effect Sizes in $d$

<table>
<thead>
<tr>
<th>Condition</th>
<th>Spring 2018</th>
<th>Autumn 2018</th>
<th>Winter 2019</th>
<th>Weighted Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison</td>
<td>.94</td>
<td>.89</td>
<td>1.57</td>
<td>1.15</td>
</tr>
<tr>
<td>Reflection</td>
<td>1.46</td>
<td>.72</td>
<td>1.61</td>
<td>1.29</td>
</tr>
<tr>
<td>Reflection with Feedback</td>
<td>1.62</td>
<td>1.11</td>
<td>1.75</td>
<td>1.52</td>
</tr>
</tbody>
</table>

Pretest to retention test effect sizes show which group better retained these gains (see Table 9). These effects were also pooled as weighted mean effect sizes to compare overall results.

Table 9

Pretest to Retention Test Effect Sizes in $d$

<table>
<thead>
<tr>
<th>Condition</th>
<th>Spring 2018</th>
<th>Autumn 2018</th>
<th>Winter 2019</th>
<th>Weighted Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison</td>
<td>1.35</td>
<td>1.31</td>
<td>2.24</td>
<td>1.66</td>
</tr>
<tr>
<td>Reflection</td>
<td>2.42</td>
<td>1.01</td>
<td>1.58</td>
<td>1.69</td>
</tr>
<tr>
<td>Reflection with Feedback</td>
<td>1.55</td>
<td>1.62</td>
<td>2.14</td>
<td>1.79</td>
</tr>
</tbody>
</table>

Summary

For each iteration of the study, mean scores and standard deviation for each group at each test time were calculated. Data was then entered into SPSS Version 25 software and checked for the usual assumptions of the general linear model including skewness.
and kurtosis. Although the data did show deviations from normality in each study, a trend emerged. Students generally performed poorly on the pretest resulting in positive skewness and kurtosis in the pretest scores. After learning the material most groups made large gains and these were now grouped on the other end of the curve, as evidenced by both negative skewness and positive kurtosis in the posttest, and retention test.

Because this trend was most likely a reflection of accurate data and not due to error, the investigator did not attempt to transform the data. Additionally, these deviations from normalcy were not extreme. Mauchly’s test revealed that in two cases the assumption of Sphericity was not violated, the one time it was, the Greenhouse-Geisser values were interpreted.

A repeated measures ANOVA was used to analyze scores based on whether they came from a group engaged in MR or not, and presence or absence of teacher feedback on the students’ reflections. In one study, this test revealed a statistically significant within-subject interaction effect between condition and time of test. However, in this case and in the Autumn 2018 study, a Bonferroni adjustment revealed score differences between groups were not statistically significant at the $p < .05$ level. In the Winter 2019 study there was a statistically significant between-subject effect based on condition ($F(2, 109) = 7.21, p < .001$). A Bonferroni adjustment revealed group score differences between the comparison group and the reflection group were statistically significant ($p = .03$). Score differences between the comparison group and the reflection with feedback group were also statistically significant ($p = .002$).

Thus, in one of three studies, there was a statistically significant difference based on group (three levels: reflection with feedback, reflection, comparison) on students’
initial ability to learn and retain academic language related to perspective drawing as measured by mean score differences on a multiple-choice test. For this reason, there was moderate evidence to reject the null hypothesis.

In addition to tests of statistical significance, the investigator calculated effect sizes pretest to posttest and pretest to retention test for each study using Cohen’s $d$. Effects by condition were also combined as weighted mean effects to compare overall results. These weighted mean effects favored the reflection with feedback group, followed by the reflection only group, for both pretest to posttest and pretest to retention test.
Chapter 5: Discussion

Introduction

The major purposes of this study were (1) to measure the influence of MR on students’ initial ability to learn academic language related to perspective drawing, (2) to measure the influence of teacher feedback on students’ initial ability to learn academic language related to perspective drawing, (3) to measure the influence of MR on students’ ability to retain academic language related to perspective drawing, (4) to measure the influence of teacher feedback on students’ ability to retain academic language related to perspective drawing, and (5) to suggest ways to improve use of MR in the classroom based on this study. This chapter is organized to address these topics.

First, results from the three studies are discussed in relation to the first four purposes, followed by a section containing discussion of trends across the three studies, as well as limitations of the study. The fifth purpose is addressed in a section that includes implications for theory, research, and practice. The chapter ends with suggestions for future research in the area of MR and a conclusion.

Spring 2018

In the Spring 2018 study, all conditions made gains. While results from the post-hoc statistical test did not reach \( p < .05 \), the groups that engaged in reflection did earn higher mean scores (see Figure 1).
As suggested by the literature, the reflection with feedback group made the greatest gains pretest to posttest. The reflection group also outperformed the comparison group, starting lowest overall and then surpassing the reflection with feedback group on the retention test. The reflection with feedback group plateaued between posttest and retention test. Engaging in reflection may have caused this group to maximize gains early in the study and achieve their full potential by the posttest. The other two groups, possibly due to continued use of the academic language related to perspective drawing in later units that built on the first unit, continued to learn to apply the academic language, but at a reduced pace.

**Autumn 2018**

In the Autumn 2018 study, all conditions made gains. While results from the post-hoc statistical test did not reach *p* < .05, the groups that engaged in reflection (see Figure 2) did retain their gains in learning at higher rates than the comparison group.
Figure 2. Mean Score by Condition Autumn 2018

Unexpectedly, the comparison group made the greatest gains post to retention test, almost surpassing the reflection with feedback group in mean score on the retention test. This sample suffered from a high attrition rate of 36 subjects compared to 16 and 19 from Spring and Winter respectively. At the time of this study, students reported high rates of illness. This might have partially accounted for the high absence-based attrition as well as the slight decrease in learning in general seen in this study compared to the other two as evidenced by effect sizes both post and retention (see Tables 7 and 8).

Winter 2019

In the Winter 2019 study, all conditions made gains (see Figure 3).
Based on effect sizes (see Tables 7 and 8), all conditions in this study made the greatest gains compared to any other conditions in previous studies with one exception. This could have been partially due to the investigator in the role of classroom teacher improving the delivery of lessons.

Unexpectedly, the comparison group began and continued to outscore either intervention group throughout the study and scored the highest mean score on all tests. A possible reason for the relatively high performance of the comparison group as seen in effect sizes (see Tables 7 and 8) was the addition of a reading intervention program at the middle school where the study took place, as discussed further in the limitations section of this chapter.

Results from the post-hoc statistical tests revealed group score differences between the comparison group and the reflection group were statistically significant \((p = .03)\). Score differences between the comparison group and the reflection with feedback group were also statistically significant \((p = .002)\). However, the score differences

Figure 3. Mean Score by Condition Winter 2019
between the reflection and the reflection with feedback group were not statically significant. As the graph shows, the groups that engaged in reflection attained slightly greater gains in mean scores pretest to posttest (see Figure 3). The comparison group gained 7.62 mean points, the reflection group gained 8.29, and the reflection with feedback group gained 8.16. On the retention test, the reflection with feedback group gained the greatest mean points beyond posttest scores, followed by the comparison group and the reflection only group.

**Synthesis**

To synthesize the results and look for trends in the data across all three studies the investigator calculated weighted mean effects (see Table 6). An effect size provided a way to express the practical significance of a study (Ellis, 2010). While scores from two of the three studies were not statistically significant based on condition, effect sizes favored the intervention overall to help students learn and apply the academic language of perspective drawing.

John Hattie (2012) cited anything over $d = 0.40$ as a worthwhile effect size for an academic intervention. Pooling the effects of all three studies generated weighted mean effects pretest to posttest of: comparison group ($d = 1.15$), reflection only ($d = 1.27$), and reflection with feedback ($d = 1.52$). This provided mild evidence for the efficacy of adding reflection over teaching alone and moderate evidence for the efficacy of adding reflection with feedback over teaching alone in initial learning.

Pooling the effects of all three studies generated weighted mean effects posttest to retention test of: comparison group ($d = 1.66$), reflection only ($d = 1.69$), and reflection with feedback ($d = 1.79$). This provided minimal evidence for the efficacy of adding
reflection over teaching alone and mild evidence for the efficacy of adding reflection with feedback over teaching alone in retaining learning.

The groups who engaged in reflection with feedback added an average \( d = 0.37 \) to their initial scores and an average \( d = 0.13 \) to their retention scores above the comparison groups. These findings were strengthened by comparable results in three similar studies. When an educational intervention can add these moderate effect sizes to the learning and retention for minimal cost, it is generally worth pursuing. However, as these calculations were based on quasi-experimental studies, other factors could have contributed to these results.

**Limitations**

This study had a number of limitations. Some were outside the investigator’s control; some were due to intentional choices. These limitations were listed to help the reader draw more accurate conclusions and to keep the results in perspective in relation to other settings.

**Intact groups.** Because this study was quasi-experimental, there was a major threat to internal validity. While the assignment of condition was random by group, the sample was not randomly selected, nor were individuals randomly assigned to groups, this meant differences between groups could be due to preexisting conditions and not the treatment (Gall et al., 2007). The pretest was an attempt to mitigate this threat to internal validity.

In addition, the comparison between five intact art classes composed of students of various ages and experience levels taught by one investigator in the role of classroom
teacher, while not ideal for external validity, was the best available in this study. Using intact groups was common in studies involving public school students (Gall et al., 2007).

**Pretest sensitization.** As mentioned previously, the pretest was an attempt to control for preexisting differences in ability. However, with the pretest came the possibility of what Gall et al. (2007) referred to as pretest sensitization, because the same test was administered to each group in the same way, the threat to internal validity was mitigated to the extent possible.

**Teacher generated test.** As noted earlier, this study relied on a teacher generated test. While Gall et al., (2007) noted the limitations of these instruments, no suitable standardized measure of the academic language related to perspective drawing was located for use in this study, so additional measures were undertaken to ensure the test was reliable. In the future, a standardized test might be developed to help the results of a similar study to have enhanced external validity.

**Time of day.** Timing of and number of tests students take in a given day, as well as frequency and duration of breaks, affected assessment results (Sievertsen, Gino, & Piovesan, 2016). Use of intact classes meant that time of day was not considered in the current study. Nor were the number of other tests given that day, nor break information, because gathering this type of information from a class of students who have six different classes taught by 44 teachers was beyond the scope of this study. Future studies might be conducted in environments that allow for these factors to be examined. The current study was strengthened by three iterations which had the conditions distributed across the day in a variety of formats. This variety spread across three studies helped attenuate the effect of the previously mentioned factors, especially when viewing weighted mean trends.
**Intelligence.** Intelligence was a confounding variable in studies of achievement and metacognition (Ohtani & Hisasaka, 2018). While Georghiades (2004b) was able to obtain general information on the students’ academic ability in his study by accessing archival information, this type of data was not available to the researcher and administering a general aptitude test was beyond the scope of the current study.

**Reading ability.** The need for academic data was underscored during the Winter 2019 study. As previously noted, a possible reason for the relatively high performance of the comparison group as seen in effect sizes (see Tables 7 and 8) was the addition of a reading intervention program at the middle school where the study took place which caused discrepancies between groups that were not present in the first two studies.

Two new reading intervention classes were formed at the start of the term, just prior to the Winter 2019 study. Scores from a standardized reading test were used to identify students who would benefit from a reading intervention, these students were then placed with a language arts teacher who would provide targeted interventions. This schedule change had the effect of grouping struggling readers together to attend the reading intervention.

By default, these same students would potentially attend elective classes together as a group. While this possibility was not formally evaluated due to the reading scores being unavailable for analysis at the time of the study, there was data to suggest this grouping took place. This included number of students served by an Individual Education Program (IEP) or 504 plan, and number of students identified as English Language Learners (ELL) (see Table 10).
In general, students served by IEPs for a specific learning disability experienced greater difficulty in reading than their peers who were not served by IEP plans (Corcoran & Chard, 2019). Students identified as ELL also had greater difficulty comprehending what they read in English than their peers who were not identified as ELL (Praveen & Rajan, 2013). It follows that these students also experienced greater difficulty learning and retaining academic language than students in general. This suggested that the reflection group had the most challenges in learning and retaining academic language. This was one explanation for the lower gains this group made from posttest to retention test (see Figure 3). The inability to conclusively analyze the covariate of reading ability was a limitation of the study.

**Length.** The current study was based on ten, 56-minute class periods, comprised of one day of pretest administration, eight days of instructional lessons and one day of review and posttest administration. During the instructional and review days, students in the reflective conditions engaged in one four to five-minute reflective activity, for a total
of nine reflective sessions. Georghiades (2004b) implemented five or six, six to seven-minute metacognitive activities during four 80-minute lessons, for a total of between 20 to 30 reflective sessions, this length and frequency may be required for a long-term benefit gained by using metacognitive strategies. In studies where a significant effect was detected the interval was often longer than two weeks (e.g. Rabin, & Nutter-Upham, 2010).

Attrition. During the 2017-18 school year, the Washington State Unexcused Absence Rate was 0.80% (OSPI, 2017). The District in which the study took place had an Unexcused Absence Rate for the same year of 0.69%, while the Unexcused Absence Rate at school in which the study took place was 1.81%, almost three times the district average. As noted earlier, absence caused 16 students to be dropped from the Spring 2018 study, 36 from the Autumn 2018 study and 12 from the Winter 2019 study. These dropped students represented 13.5%, 27.7%, and 25.0% respectively of each total enrollment for the investigator in the role of classroom teacher.

Low attendance at school could have many causes, but anxiety was certainly a contributing factor (Ingul & Nordahl, 2013). Weight’s (2017) study provided evidence that teachers and their students benefit from MR. The chronically absent students who were dropped from the current study may have been the ones who would have benefited most from the perceived environmental improvements of MR (Black & Wiliam, 2009).

Feedback. Bianchi (2007) noted the possible differential effect of teachers reacting to student reflections as a weakness in earlier studies. If one group benefited from improved instruction based on their expressed needs and another did not, this could confound interpretation of results. To avoid this one could have chosen to not read the
student reflections and remove the feedback aspect from the study entirely, or feedback could have been provided by a third party, as in Baliram, and Ellis’ (2019) study.

In this study the investigator in the role of classroom teacher adjusted instruction for all classes based on feedback received from the reflection groups. This may have inflated the comparison group scores, possibly obscuring some of the effects of the intervention.

**Measurement of reflection.** Measurement of metacognition was difficult. The most accurate way, monitoring of student behavior during reflection (Dent & Koenka, 2016) was not feasible in the current study as there was only one investigator and an average of 25 students in each classroom. While the use of student questionnaires may have provided information about the depth of student reflection, this was outside the scope of the current study. As a result, the investigator took the written student reflections at face value as evidence of reflection. Not knowing the depth of student reflection was a limitation in this study.

**Depth of reflection.** This study relied on a relatively simple reflective activity. A more in-depth reflection activity such as journaling (McDonald & Dominguez, 2009) may have led to greater gains and avoided the danger of reflection becoming rote (Bannister-Tyrrell, & Clary, 2017). Journaling was ultimately rejected as too time consuming. Instead, the reflection prompt varied from day to day, so students did not lose interest (Georghiades, 2004b). The assessments were also presented in a novel format to keep students engaged.

However, students might not have engaged in metacognitive thinking (Wismath, Orr, & Good, 2014), even though provided with a variety of prompts (Kwon & Jonassen, 2011). Additional studies where students learn and apply the academic language related
to perspective drawing and other areas in art, might benefit from a greater variety and depth of reflective activities, as well as the teacher-student relationship as a moderating variable.

**Length of term.** A conducive classroom environment for reflection (Black, & Wiliam, 2009) included student trust of the teacher (Georghiades, 2004a; Hattie & Clarke, 2019), the current studies were conducted in the context of a semester length class, at times, towards the beginning of the term. There may have been insufficient time for students to develop trust of the investigator in the role of classroom teacher. Additionally, students must have been free from fear of judgment from other students or the instructor (Slinger-Friedman & Patterson, 2016). Perhaps the manner in which the instructor provided feedback was not optimal. An anonymous method, while less responsive, may have promoted greater feelings of student security.

**Snow days.** During the Winter 2019 study, an unusual weather-related event caused school to be canceled on four days. Three additional days were reduced in length. The investigator in the role of classroom teacher had initially planned to administer the posttest nine days after the pretest. Due to missed instructional time, the posttest was postponed until all the material had been taught. The delay between pretest and posttest was increased by a weeklong mid-winter break that occurred towards the end of the unit. This resulted in the unit spanning four weeks instead of the planned two weeks.

All these delays also made for a somewhat tumultuous classroom environment. Students seemed distracted by the change in routine and possibility of additional snow days. The effect on student learning was unknown, but possibly significant.
Implications

Notwithstanding the above limitations, this study had implications for metacognitive theory, research, and practice.

Theory. This study demonstrated a moderate effect for MR with feedback over the comparison group, but not compared to the MR group. While MR groups were predicted to outscore comparison groups based on theory, the MR groups who received feedback were predicted to score highest. This did not happen. Studies including fine-grained comparisons between types of prompts, when and how feedback is provided, and length of intervention, will allow theory on these topics to be further developed.

Research. The studies in this dissertation were relatively easy to implement. The data collected and methods of teaching did not differ greatly from regular practice. The only significant differences were the extra time in computing split-half reliabilities of the test, and the ANOVA, statistical work the investigator in the role of classroom teacher would not normally undertake. The calculation of effect sizes had however been used by the teacher for some time in response to Hattie’s (2012) call to “know thy impact.” While Hattie (2012) concluded that teachers should focus on teaching and researchers on researching, a collaborative approach should be considered to add to the empirical data in this area.

Practice. Teachers should be trained in MR. It was not difficult to implement and can have gains that outweigh its minimal costs. More importantly MR prepared students to be critical thinkers now and for their future learning (Zuckerman, 2003).

The implementation of metacognitive practice in the classroom was a two-edged sword. The reflection activity must be simple enough that teachers could easily
implement it in the face of competing priorities, but if it was reduced to a simple checklist activity its efficacy was lost (Bannister-Tyrrell & Clary, 2017). The reflective prompts such as an I Learned statement, a Key Idea Identification, or a Clear and Unclear Windows (Ellis & Denton, 2010) were simple enough to be quickly taught to students. When the MR prompt was varied from day to day and presented in a dynamic format this encouraged students to engage in authentic assessment. The variety reduced the chances students simply viewed this as one more activity to check off the list (Bannister-Tyrrell & Clary, 2017).

Further Research

While research on metacognition was broad and well developed, some specific applications had not been fully studied, and not all studies were as rigorously conducted. Schunk (2008) recommended five practices to improve research in the areas of metacognition and related areas. First, researchers promote clarity by explicitly defining what they are researching. Second, researchers ground their work when they provide links from their study to foundational theories. Third, it is helpful to explain how the assessments used relate to the defined theories. Fourth, too many studies correlate metacognition with academic success; there is a need for more cause/effect research. Fifth, when the methods which foster a process, such as self-regulation, are listed, and tested, i.e. goal setting and self-evaluation, the study’s utility increases (Schunk, 2008).

Bannert and Mengelkamp (2008) called for improved measures of metacognition. They hypothesized that while questionnaires measure quantity, and think aloud methods measure quality, both have their limitations. Other researchers (e.g. Dent & Koenka, 2016; McCardle & Hadwin, 2015, Ohtani & Hisasaka, 2018; and Schellings & Van Hout-
Wolters, 2011) have made similar observations and noted the need for further research in this area.

There were multiple reports of teachers valuing metacognitive strategies, but not spending much time promoting them (e.g. Bannister-Tyrrell & Clary, 2017; Dignath & Büttner, 2018). Studies which ask the types of questions Bannister-Tyrrell and Clary (2017) posed to more teachers to see if the reasons for this disconnect can be understood and addressed, are called for.

Studies such as the current one, provided evidence of the short-term benefits of MR. There is reason to believe that metacognition has long-term effects as well (Georghiades, 2004a). This will require longitudinal studies of the type called for by Dignath and Büttner (2008) and Panadero (2017).

The link between feedback and improved academic outcomes has been studied in depth (e.g. Hattie & Clarke, 2019; Hattie & Timperley, 2007; Schunk, Pintrich, & Meece, 2008), but some areas, including the best timing of the delivery of the feedback are still being researched (Shute, 2008). Baliram, and Ellis (2019) called for additional study in this area.

**Conclusion**

The empirical evidence provided by this study should be interpreted with some caution based on aforementioned limitations. A strength of this study was that it did not rely on self-report which was often the case in these types of study (Dinsmore et al., 2008). Also, because it was conducted in a school classroom, it had a certain “real world” authenticity. Thus, the study avoided Zimmerman’s (2011) criticism that many of these types of studies have been done outside the classroom context and after the fact.
Metacognition reflection is a not magic solution to every problem in education. Teachers would do well to remember interventions come with a cost and the best interventions do not always work for all students (Black & Wiliam, 2009). Ellis and Bond (2016) cautioned that many educational innovations, even those with sound theoretical foundations are subject to failure when attempted in the classroom. They also reminded us that using numbers as the sole basis for judging the effectiveness of an intervention in a democratic society, while getting at the academic side of things, runs the risk of overlooking other beneficial aspects of school life such as social interaction (Ellis & Bond, 2016).
References


http://dx.doi.org/10.1037/0003-066X.34.10.906


http://doi:10.1080/016502500383421


(MABC-2) suggest that non-parametric scoring methods are required. *PLoS ONE*, 13(5), 1–22.


https://doi.org/10.1177/0022487117696279


https://doi.org/10.3389/fpsyg.2013.00863


Washington, D.C.


10.1080/00313831.2015.1066436


https://doi.org/10.1080/10862968109547390


https://doi.org/10.3102/0013189X020009013


---

87


doi:10.1080/00461520.2013.794676

doi:10.1017/CBO9780511840975
Appendix A

Power Analysis

![Diagram of G*Power 3.1.9.2 interface with power analysis results]

- **Test family**: F tests
- **Statistical test**: ANOVA: Repeated measures, between factors
- **Type of power analysis**: A priori: Compute required sample size - given α, power, and effect size

**Input Parameters**
- Effect size f: 0.30
- α err prob: 0.05
- Power (1-β err prob): 0.8
- Number of groups: 3
- Number of measurements: 3
- Corr among rep measures: 0.5

**Output Parameters**
- Noncentrality parameter λ: 10.1250000
- Critical F: 5.1230624
- Numerator df: 2.0000000
- Denominator df: 72.0000000
- Total sample size: 75
- Actual power: 0.8032375

---

**Critical F**: 3.12391

---
Appendix B

IRB Approval Letter

April 15, 2019

Subject: IRB Approval – IRB # 181906010 (Exempt Review)

To: Eric Howe

Your research project “Using Metacognitive Reflection to Improve Student Learning” has been approved. This study was approved under exempt review as it meets the criteria listed in the SPU IRB User Guidelines (2012, p. 5).

Your approval is in effect until what time any methods of the study change substantively. When that occurs, you will need to renew your IRB application. Your study has been assigned IRB number: IRB # 181906010.

To complete your documents please add the IRB # to your study’s written recruitment material and invitation to participate in the research project.

Best wishes in the completion of your research.

Sincerely,

[Signature]

John B. Bond, Ed.D.
SOE IRB Coordinator
Professor of Educational Leadership

Cc: Dr. Arthur Ellis
### Appendix C

#### Descriptive Statistics Spring 2018

<table>
<thead>
<tr>
<th></th>
<th>$N$</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statistic</td>
<td>Statistic</td>
<td>Std. Error</td>
</tr>
<tr>
<td>Pretest</td>
<td>102</td>
<td>0.077</td>
<td>0.239</td>
</tr>
<tr>
<td>Posttest</td>
<td>102</td>
<td>-0.635</td>
<td>0.239</td>
</tr>
<tr>
<td>Retention Test</td>
<td>102</td>
<td>-0.999</td>
<td>0.239</td>
</tr>
</tbody>
</table>
Appendix D

Sample of 36-question Perspective Test

1. What is perspective?

Mark only one oval.

- A way of seeing.
- A type of drawing.
- A viewpoint.
- All of the above.
- Other: __________

3. Where is the horizon line in this photo?

Mark only one oval.

- There is no horizon line.
- By the cat’s ears.
- By the cat’s tail.
- By the pinecone.
4. Where is/are the vanishing point(s) in this photo?

Mark only one oval.

☐ By the trees in the middle.
☐ By the roof, high up.
☐ There is no vanishing point.
☐ On the bricks and roof area.

5. Which picture has one vanishing point?

Mark only one oval.

☐ Option 1
☐ Option 2
6. Which picture has two vanishing points?
Mark only one oval.

- Option 1

8. In one point perspective, the vanishing point is?
Mark only one oval.

- Never seen.
- In the middle.
- On the horizon line.
- Used for shading.

9. In two point perspective, both of the vanishing points are?
Mark only one oval.

- Never seen.
- In the middle.
- On the horizon line.
- Used for shading.

10. In three point perspective, two of the vanishing points are?
Mark only one oval.

- Never seen.
- In the middle.
- On the horizon line.
- Used for shading.
12. How many vanishing points?

Mark only one oval.

☐ This picture has no vanishing points.
☐ Three
☐ Two.
☐ One.

17. In birds eye view the horizon line is?

Mark only one oval.

☐ In the middle.
☐ At the very bottom.
☐ Near the bottom
☐ Up high.

18. In worms eye view the horizon line is?

Mark only one oval.

☐ In the middle.
☐ At the very bottom.
☐ Near the bottom
☐ Up high.
21. What is the viewpoint?

Mark only one oval.

- Overhead.
- Birds eye.
- Worms eye.
- Normal.

22. What is the viewpoint?

Mark only one oval.

- Overhead.
- Birds eye.
- Worms eye.
- Normal.
25. To draw the people smaller as they get farther away you use?

Mark only one oval.

- A pencil.
- Foreshortening.
- Texture.
- Planetary forces.

26. The hills fade in the distance showing?

Mark only one oval.

- Pollution.
- Atmospheric perspective.
- Earth's curvature.
- Fading light.
28. What forms an orthogonal line in this photo?

Mark only one oval.
- The edges of the tables.
- The recycle bin.
- The sink.
- There is no orthogonal line.

30. If you were drawing the corner of a building in normal view you would use?
Mark only one oval.
- One point perspective.
- Two point perspective.
- Three point perspective.
- None, this is not a good perspective drawing.

35. If you were drawing a building in overhead view you would put the horizon line?
Mark only one oval.
- At the very bottom.
- A little way up the page.
- In the middle.
- At the top.

36. Perspective drawing causes people to see?
Mark only one oval.
- Three dimensions on a two dimensional surface.
- Two dimensions on a three dimensional surface.
- One dimension on a three dimensional surface.
- Two dimensions on a one dimensional surface.
Appendix E

Reflective Prompt Slides

Now you are going to create an I learned statement. An I learned statement explains what you learned.

- Today I learned ________________________.

Now you are going to create Key Idea Identification. A Key Idea Identification tells the main point of the lesson.

- I think the Key Idea for today's lesson is ________________________.

Now you are going to create Clear and Unclear Windows. Clear windows are things you understand. Unclear windows are things you are still learning or have questions about.

<table>
<thead>
<tr>
<th>Clear Windows</th>
<th>Unclear Windows</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix F

Sample of Student Responses

Now you are going to create Clear and Unclear Windows.
- Clear windows are things you understand.
- Unclear windows are things you are still learning or have questions about.

<table>
<thead>
<tr>
<th>Clear Windows</th>
<th>Unclear Windows</th>
</tr>
</thead>
</table>

- [Notes on the board]
Appendix G

Post-Hoc Spring 2018

Multiple Comparisons

Bonferroni Adjustment

<table>
<thead>
<tr>
<th>(I) Condition</th>
<th>(J) Condition</th>
<th>Mean Difference (I-J)</th>
<th>Std. Error</th>
<th>Sig.</th>
<th>95% Confidence Interval</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison</td>
<td>Reflection</td>
<td>-.506</td>
<td>1.161</td>
<td>1.000</td>
<td>-3.334</td>
<td>2.322</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reflection with Feedback</td>
<td>-2.218</td>
<td>.971</td>
<td>.074</td>
<td>-4.583</td>
<td>0.147</td>
<td></td>
</tr>
<tr>
<td>Reflection</td>
<td>Comparison</td>
<td>0.506</td>
<td>1.161</td>
<td>1.000</td>
<td>-2.322</td>
<td>3.334</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reflection with Feedback</td>
<td>-1.711</td>
<td>1.186</td>
<td>.457</td>
<td>-4.600</td>
<td>1.177</td>
<td></td>
</tr>
<tr>
<td>Reflection with Feedback</td>
<td>Comparison</td>
<td>2.218</td>
<td>0.971</td>
<td>.074</td>
<td>-0.147</td>
<td>4.583</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reflection</td>
<td>1.711</td>
<td>1.186</td>
<td>.457</td>
<td>-1.177</td>
<td>4.600</td>
<td></td>
</tr>
</tbody>
</table>

Based on observed means.
Appendix H

Descriptive Statistics Autumn 2018

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Skewness Statistic</th>
<th>Skewness Std. Error</th>
<th>Kurtosis Statistic</th>
<th>Kurtosis Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td>94</td>
<td>0.702</td>
<td>0.249</td>
<td>1.12</td>
<td>0.493</td>
</tr>
<tr>
<td>Posttest</td>
<td>94</td>
<td>-0.205</td>
<td>0.249</td>
<td>-0.414</td>
<td>0.493</td>
</tr>
<tr>
<td>Retention test</td>
<td>94</td>
<td>-0.233</td>
<td>0.249</td>
<td>-0.641</td>
<td>0.493</td>
</tr>
</tbody>
</table>
Appendix I

Post-Hoc Autumn 2018

*Multiple Comparisons*

Bonferroni Adjustment

<table>
<thead>
<tr>
<th>(I) Condition</th>
<th>(J) Condition</th>
<th>Mean Difference (I-J)</th>
<th>Std. Error</th>
<th>Sig.</th>
<th>95% Confidence Interval Lower Bound</th>
<th>95% Confidence Interval Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison</td>
<td>Reflection</td>
<td>-2.716</td>
<td>1.468</td>
<td>0.203</td>
<td>-6.296</td>
<td>0.865</td>
</tr>
<tr>
<td></td>
<td>Reflection with Feedback</td>
<td>-0.873</td>
<td>1.293</td>
<td>1.000</td>
<td>-4.024</td>
<td>2.277</td>
</tr>
<tr>
<td>Reflection</td>
<td>Comparison</td>
<td>2.716</td>
<td>1.468</td>
<td>0.203</td>
<td>-0.865</td>
<td>6.296</td>
</tr>
<tr>
<td></td>
<td>Reflection with Feedback</td>
<td>1.842</td>
<td>1.444</td>
<td>0.616</td>
<td>-1.679</td>
<td>5.364</td>
</tr>
<tr>
<td>Reflection with Feedback</td>
<td>Comparison</td>
<td>0.873</td>
<td>1.292</td>
<td>1.000</td>
<td>-2.277</td>
<td>4.024</td>
</tr>
<tr>
<td></td>
<td>Reflection</td>
<td>-1.842</td>
<td>1.444</td>
<td>0.616</td>
<td>-5.364</td>
<td>1.679</td>
</tr>
</tbody>
</table>

Based on observed means.
Appendix J

Descriptive Statistics Winter 2019

<table>
<thead>
<tr>
<th>Test Type</th>
<th>N</th>
<th>Skewness Statistic</th>
<th>Skewness Std. Error</th>
<th>Kurtosis Statistic</th>
<th>Kurtosis Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td>112</td>
<td>0.456</td>
<td>0.228</td>
<td>0.872</td>
<td>0.453</td>
</tr>
<tr>
<td>Posttest</td>
<td>112</td>
<td>-0.262</td>
<td>0.228</td>
<td>-0.332</td>
<td>0.453</td>
</tr>
<tr>
<td>Retention Test</td>
<td>112</td>
<td>-0.259</td>
<td>0.228</td>
<td>-0.008</td>
<td>0.453</td>
</tr>
</tbody>
</table>
Appendix K

Post-Hoc Winter 2019

Multiple Comparisons

Bonferroni Adjustment

<table>
<thead>
<tr>
<th>(I) Condition</th>
<th>(J) Condition</th>
<th>Mean Difference (I-J)</th>
<th>Std. Error</th>
<th>Sig.</th>
<th>95% Confidence Interval Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison</td>
<td>Reflection</td>
<td>2.287*</td>
<td>0.859</td>
<td>.027</td>
<td>0.197</td>
<td>4.376</td>
</tr>
<tr>
<td></td>
<td>Reflection with Feedback</td>
<td>3.562*</td>
<td>0.999</td>
<td>.002</td>
<td>1.135</td>
<td>5.993</td>
</tr>
<tr>
<td>Reflection</td>
<td>Comparison</td>
<td>-2.287*</td>
<td>0.859</td>
<td>.027</td>
<td>-4.376</td>
<td>-.197</td>
</tr>
<tr>
<td></td>
<td>Reflection with Feedback</td>
<td>1.277</td>
<td>1.012</td>
<td>.628</td>
<td>-1.182</td>
<td>3.737</td>
</tr>
<tr>
<td>Reflection with Feedback</td>
<td>Comparison</td>
<td>-3.564*</td>
<td>0.999</td>
<td>.002</td>
<td>-5.993</td>
<td>-1.135</td>
</tr>
<tr>
<td></td>
<td>Reflection</td>
<td>-1.277</td>
<td>1.012</td>
<td>.628</td>
<td>-3.737</td>
<td>1.182</td>
</tr>
</tbody>
</table>

Based on observed means.
The mean difference is significant at the .05 level.