ABSTRACT

JONES, JASON PATRICK. Effects of Iterative Course Changes on Student Performance, Confidence, and Metacognitive Monitoring Accuracy within an Undergraduate Introductory Geoscience Course. (Under the direction of Dr. David McConnell.)

In recent years, there has been an impetus within the geoscience community towards developing and adapting research-validated, student-centered learning activities for geoscience courses. This type of initiative is essential if we are to work towards maximizing student learning. However, while these teaching strategies and course design elements are useful for the instructor, they may not make the learning process itself explicit to the student. The two studies presented in this work represent efforts to contribute to both instructional reform and student awareness of the learning process. The first study considers the effects of utilizing short videos to move lessons on basic geoscience content outside of the classroom. As a result of this “flipping,” instructors can use newly-gained time in-class to leverage benefits of active learning strategies to improve student learning. Results demonstrated that during summative assessments (midterm exams), students demonstrated at least equal levels of mastery on concepts that were no longer presented upon in-class. Additionally, many concepts that were augmented with additional activities elicited significant gains in student performance and confidence in their knowledge.

Metacognitive monitoring accuracy is an important educational skill that is predictive of student performance. The second study investigated the relationship between students’ judgments of their performance and their actual performance during summative exams in an introductory geoscience course. The study collected student confidence data for every question of three midterm exams and compared this confidence to performance via an
empirically-derived measure of the disparity between students’ perceptions of their performance and their actual performance called calibration accuracy. Additionally, students were asked to make a judgment of total performance on their exam after answering all questions (a “postdiction”). These “local” (item-level) and “global” (postdiction) measures of accuracy were analyzed in response to incremental changes in the course over two semesters.

After adding metacognitive prompts into required course homework assignments to little cumulative effect, the platform of formative assessment was experimentally altered in the second study semester. The researcher-designed quizzing system (CLASS) collected student performance and confidence data and automatically presented students with feedback regarding their monitoring accuracy related to geology content. Students utilizing the CLASS system performed significantly better than their predecessors and were generally more accurate in their approximations.
Effects of Iterative Course Changes on Student Performance, Confidence, and Metacognitive Monitoring Accuracy within an Undergraduate Introductory Geoscience Course

by

Jason Patrick Jones

A thesis submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the degree of Master of Science

Marine, Earth, and Atmospheric Sciences

Raleigh, North Carolina

2017

APPROVED BY:

_______________________________
Dr. David A. McConnell
Committee Chair

_______________________________
_______________________________
Dr. John L. Nietfeld            Dr. Karen S. McNeal
DEDICATION

For the family who gave me a name, a love that is nameless, the family we’ve named, and the family we’ve yet to name.
BIOGRAPHY

Jason was born in Salina, KS and remained a resident of Kansas for the majority of his upbringing. After a life filled with dueling interests of science and music (and a professional music career), he completed his bachelor’s degree in geology from the University of Kansas in 2015. Also while at KU, he concurrently completed an intensive STEM-based teacher preparation program, earning a secondary earth and space science teaching certificate and licensure. Eager to combine the investigative rigors of science with a love of education and promoting the science, Jason conducted undergraduate research in introductory geology courses at KU and participated in introductory course redesign. Whilst presenting the results of his undergraduate research at the Geological Society of America meeting in Vancouver, BC in 2014, Jason fatefully met Dr. McConnell and the rest of the Geoscience Learning Process Research Group. After a move to North Carolina with his wife Katheryn and nine-month-old son Desmond, he began to pursue his Master of Science degree.
ACKNOWLEDGMENTS

First of all, I would like to acknowledge my family, both nuclear and extended, without whom none of this would be possible. I want to thank my sweet son Desmond, who put up with long days and nights without me being able to “play my room,” and my beautiful wife, KC, whose unfailing support always gets me through the tough situations and makes the good situations even better. Finally, I want to thank my parents who graciously accepted my plans to move both their child and grandchild across the country.

Next I would like to acknowledge my advisor Dr. David McConnell, for being an amazing mentor and always supporting and nurturing my ideas and providing me with the opportunity to do something I love to do in an environment I am lucky to do it in. I am very much looking forward to continuing this relationship moving on into doctoral work. I would also like to acknowledge the Geoscience Learning Process Research Group, both current and alumni members, who not only helped me transition to life in Raleigh, but who have become great friends. Doug Czajka, LeeAnna Chapman, Mike Pelch, and Jennifer Wiggen, over the past two years you have made this rewarding experience at NCSU that much better and I appreciate your friendship and having the opportunity to learn from your individual personalities and skills.

Finally, I would like to acknowledge my committee for their mentorship and guidance during this time. Dr. John Nietfeld for his support and mentorship in furthering my scholarship of education psychology and Dr. Karen McNeal for her support and mentorship in our mutual pursuit of furthering the discipline of geoscience education research.

iv
TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................ vi

LIST OF FIGURES ................................................................................................... vii

CHAPTER 1: INTRODUCTION .................................................................................... 1

CHAPTER 2: EFFECTS OF CLASSROOM “FLIPPING” ON CONTENT MASTERY AND STUDENT CONFIDENCE .................................................................................. 3

  2.1 Background ........................................................................................................ 3
  2.2 Methods ........................................................................................................... 10
  2.3 Statistical methods ......................................................................................... 16
  2.4 Results ............................................................................................................ 19
  2.5 Discussion ..................................................................................................... 24

CHAPTER 3: EVALUATING THE EFFECT OF ITERATIVE COURSE CHANGES ON STUDENT METACOGNITIVE MONITORING ACCURACY AND ACHIEVEMENT ........................................................................................................ 31

  3.1 Background .................................................................................................... 31
  3.2 Methods .......................................................................................................... 37
  3.3 Statistical methods ....................................................................................... 49
  3.4 Results .......................................................................................................... 52
  3.5 Discussion .................................................................................................... 64

CHAPTER 4: CONCLUSIONS ................................................................................. 72

REFERENCES ......................................................................................................... 73
LIST OF TABLES

CHAPTER 2

Table 2.1 Participant gender, academic rank, and declared major............................................. 11
Table 2.2 Exam question categories and descriptions................................................................. 16
Table 2.3 Mean and standard deviations for question performance and confidence .............. 20
Table 2.4 Means and standard deviations of category change scores .................................. 24

CHAPTER 3

Table 3.1 Gender, academic rank, and declared major for students in each course .......... 39
Table 3.2 Total sample size and related subsets for each study semester .......................... 50
Table 3.3 Performance variable mean and standard deviation across semesters ............... 53
Table 3.4 Global accuracy variables mean and standard dev. across study semesters ....... 56
Table 3.5 Pearson correlations for local accuracy variables ...................................................... 58
Table 3.6 Local accuracy variables mean and standard deviation from 2014-2016 .......... 63
LIST OF FIGURES

CHAPTER 2

Figure 2.1 Example of data collection methods for student confidence used in the study ... 14

Figure 2.2 Mean percentage of correct responses for questions from Category 1 and Category 2 for the Fall 2014 and Fall 2016 semesters .............................................................. 21

Figure 2.3 Mean percentage of correct responses for questions from Category 3 and Category 4 for the Fall 2014 and Fall 2016 semesters .............................................................. 23

Figure 2.4 Mean percent change for each question category between treatment and control semesters (Fall 2016 – Fall 2014). Symbol (*) indicates significant change from zero at p < 0.05 level. .............................................................. 24

CHAPTER 3

Figure 3.1 Mean performance vs. mean calibration for the baseline Fall 2014 semester ..... 36

Figure 3.2 Example question and calibration calculation for local accuracy judgments ..... 41

Figure 3.3 Example global task prediction within CLASS quiz attempts ....................... 44

Figure 3.4 Sample question and item-level confidence measure within CLASS ............... 45

Figure 3.5 Global postdiction screen for a quiz within CLASS ........................................ 46

Figure 3.6 Sample results screen automatically presented to students after taking a quiz in CLASS .............................................................. 46

Figure 3.7 Global results communicated to students in CLASS ........................................ 47

Figure 3.8 Graphical representation of student quiz results in CLASS ............................ 48

Figure 3.9 Mean accuracy across all students for each study semester ............................ 54
Figure 3.10 Mean calibration between each exam across each target semester .................. 55

Figure 3.11 Mean signed accuracy for global postdictions during three exams in Fall 2015 and Fall 2016 ................................................................................................................. 56

Figure 3.12 Mean performance vs. mean calibration for all exams for the Fall 2015 semester ................................................................................................................................. 59

Figure 3.13 Mean calibration by gender for Fall 2015 semester ........................................... 61

Figure 3.14 Mean performance vs. mean calibration for all exams for Fall 2016 ............... 61

Figure 3.15 Mean local accuracy (calibration) by gender for each exam of Fall 2016 ....... 63
CHAPTER 1: INTRODUCTION

The importance of science, technology, engineering, and math (STEM) education has been the focus of several governmental calls for action within the US during the past two decades (e.g., NRC, 1996; PCAST, 2012). However, at the college level, fewer than 40% of students who enter college as STEM majors graduate with STEM degrees (PCAST, 2012). The report by the President’s Council of Advisors on Science and Technology (PCAST, 2012) noted that US colleges and universities would need to significantly increase the number of graduates in STEM fields in the following decade to meet the predicted growth in demand by 2022.

It is apparent from work in discipline-based education research that efforts to achieve this growth can be approached via two channels: 1) Investigating best practices related to effective course design and pedagogy; and, 2) Characterizing and supporting the student learning process within the discipline (Singer et al., 2012). This work presents one study towards each goal. The research investigates effects of course alterations to an introductory geoscience course on student performance, confidence, and an important and predictive student trait called metacognitive awareness. The first part of the study (Chapter 2) focuses upon the potential of partially “flipped” course to provide additional time for student-centered activities to increase student learning in an introductory geology course. The second part (Chapter 3) features an in-depth characterization of introductory geology students’ metacognitive monitoring accuracy measured during summative exams. It is hoped that information gained from these studies will be extended and generalized to provide
recommendations for practice in better supporting learners within introductory geoscience courses.
CHAPTER 2: EFFECTS OF CLASSROOM “FLIPPING” ON CONTENT MASTERY AND STUDENT CONFIDENCE

2.1 Background

This study examines two aspects of the student experience in face-to-face university introductory science courses. The first is pre-class preparation where students gather background information regarding course content and second is the interaction with instruction during class meetings. The traditional approach to providing students with the requisite background knowledge for subsequent class meetings has typically been in the form of course reading assignments (Burchfield & Sappington, 2000). Unfortunately, many students do not read the textbook as preparation for the course meetings (Burchfield & Sappington, 2000). Podolefsky and Finkelstein (2006) found that only 13% of students in various introductory physics classes reported that they “often” (>80% of the time) read the book before class. In contrast, most students appear to consider their textbook as a reference source to support preparation for exams rather than as a resource to be used as a primer for in-class learning (Podolefsky & Finkelstein, 2006). This process of students ignoring pre-class assignments may be reinforced by instructors following a traditional didactic model of teaching, as many of the concepts and terms intended to be communicated in the pre-class reading assignments may be subsequently presented in class (Crouch & Mazur, 2001). The result is an inefficient system where the parameters of student learning are defined almost exclusively by what occurs inside the classroom.
Efforts to improve the student experience both in and out of the classroom have fallen into a category of pedagogies titled “active learning” (Singer et al., 2012). Active learning places the emphasis of instruction on students and their experiences in the classroom as opposed to having them be passive recipients of information (Mazur, 1997). Expanding consideration to all of active learning and STEM disciplines, a meta-analysis of 225 studies comparing active learning to the traditional didactic model computed average gains in student learning measured by summative exams on the order of approximately 0.5 standard deviations and a 1.5x reduction of the DFW (drop, fail or withdraw) rate (Freeman et al., 2014).

Given the evidence of success for active learning to improve student performance in STEM courses, the next issue facing instructors was how to find time to incorporate these strategies into their courses without sacrificing content coverage (Crouch & Mazur, 2001). This dilemma contributed to the development of the “inverted” or “flipped” learning model (Lage et al., 2000; Bishop and Verleger, 2013; Gross et al., 2015). This process places the responsibility of learning introductory concepts to students outside of the class and is facilitated by for-credit quizzes on outside of course content (e.g., Lage et al., 2000; Crouch & Mazur, 2001; Gross et al., 2015). This ensures that students have a stake in interacting with the background content, and, by moving some content outside of the classroom experience, provides additional time during class for students to tackle more difficult or application-based activities in the presence of the instructor and peers (Evans, 2011; Strayer, 2012; Tucker, 2012; Gajjar, 2013). Additionally, the practice of “flipping” may not only allow the instructor to go into more depth on a topic, but it may also enable them to employ active
learning strategies and formative assessments that may previously have been difficult to implement due to class time constraints (Freeman et al., 2011).

Investigation into the effectiveness of flipped environments is subject to wide variability in study designs and reported results. A common experimental design investigating flipped classrooms compares a newly-adopted flipped approach to the course (including pervasive active learning elements) against a control semester that followed a more-traditional didactic approach, minus the active learning activities (Schultz et al., 2014; Strayer, 2012; Tucker, 2012; Tune et al., 2013). Given the independent effects of active learning on improving student performance in STEM courses (Freeman et al., 2014), observed benefits cannot be solely attributed to the flipped class design. Other studies have attempted to control variables via the comparison of courses utilizing active learning and those using active learning with flipped components. Jensen et al. (2015) compared two sections of an introductory biology course with both utilizing an active learning approach. In the study, students in the flipped condition were required to complete quizzes delivered in the on-line course management system (CMS) prior to each class meeting (3 per week; 39 in total) consisting of a mixture of background readings, demonstration videos, and probing questions (Jensen et al., 2015). Students then applied their knowledge of the content in novel situations via in-class group activities. In contrast, students in the non-flipped approach completed the same exploratory activities in-person via group work, but finished the lesson via a required CMS quiz that was completed after class and consisted of the same information contained in the flipped condition. Results found no significant differences between sections varying only in pre-class consumption of background material (Jensen et
al., 2015). Gross et al. (2015) compared five years of data from an upper-level biochemistry course (three iterations of traditional instruction and two years of flipped instruction with active learning activities). In two semesters containing varying amounts of in-class time, researchers recorded video lectures, broke them into 5-20 minute segments, and make them optional prior to course meetings. During class, students participated in active problem-solving and team-based learning. Upon analysis, researchers found significant gains for all students, but specifically higher gains from the flipped format for lower-performing students and females (Gross et al., 2015). These results suggest that the primary benefit for flipped learning may be found within the active learning the format may serve to facilitate in the classroom.

One may ask, is all flipping created equal? One version of a flipped class format introduces some course content prior to lecture using instructional videos and related assignments (Moravec et al., 2010). Students can also use these videos for content review and they provide a valuable means for learners of different abilities as students are able to pause, repeat, and re-watch the videos as necessary (Schultz et al., 2014). The use of video has been shown to increase students’ attention and engagement (e.g. Green, et al., 2003; Jha et al., 2002; Zhang et al., 2006), increase motivation and self-efficacy (Bennett & Glover, 2008) and improve understanding (e.g. Choi & Johnson, 2005; Seeling & Reisslein, 2005; Zhang et al., 2006). Access to video lessons in flipped classes can provide students with autonomy and control in the study process and may avoid expertise reversal effects as students with more conceptual understanding can skip over scaffolds intended for more novice learners (Kalyuga et al., 2003). Stelzer et al. (2009) described how the introduction of
video-based pre-class multimedia modules (MMs) in introductory physics classes resulted in an improvement of student performance on an assessment immediately after completion of the module, and again on a test two weeks later. Performance of a group of student volunteers was measured on 30+ question exams composed mostly of conceptual multiple choice questions. The students exposed to the MMs performed approximately one grade level better than similar student populations that reviewed equivalent information in textbook format (Stelzer et al., 2009). The MMs consisted of a series of brief “acts” combined in a narrated 15-minute-long video with some embedded multiple choice questions. Further, when student results from questions in the MM-supported classes were compared to those from prior years that relied on textbook reading alone, the MM-supported students scored higher (average 57% correct answers vs. 49%; Chen et al., 2010). The degree of improvement was not consistent among concepts, suggesting that some MMs were more effective than others (Chen et al., 2010).

Videos employ visual, audio, temporal, and other representational symbol systems to transmit information that may be difficult to convey through a textbook or lecture format (Tantrarungoroj, 2008). Some of this information may be sequences in motion (e.g., movement along a fault or how Coriolis force modifies wind direction as a function of latitude), or perspectives that are difficult to observe in real life, whether due to their temporal, spatial, or remote nature (e.g., diving under the Antarctic Ice or motions of tectonic plates and their effects; Wetzel et al., 1994). Towards this end, the work of Mayer in aligning the characteristics of students’ cognitive abilities with multimedia design has generated three primary assumptions of multimedia learning (Mayer, 2002 for a review): 1) Students possess
two separate channels for processing audio and visual information (Baddeley, 1999); 2) Presenting information efficiently via both audio and visual channels can increase the net amount of information processed (in effect increasing students’ working memory capacity; Chandler & Sweller, 1991; Mayer, 2002); and, 3) Learners are “active processors who seek to make sense of multimedia presentations” and not passive “tape recorders” of presented information (Mayer, 2002, p. 36). Multimedia that is designed with these assumptions in mind have generated four effects that have been shown to increase learning in students consuming multimedia content (Mayer, 2003). They are: a) multimedia effect - displaying words and pictures simultaneously are more effective than words alone; b) coherence effect - eliminating extraneous details in media improves learning; c) spatial contiguity effect - images and descriptive words are more effective when in close proximity; and, d) personalization effect - more conversational language is more effective than formal language (Mayer, 2003). Though these elements of multimedia design have been shown to be effective both on paper or in videos, these effects can be understood to have the greatest potential when paired with the added expansions of processing ability that video provides (Mayer, 2002).

To utilize the benefits of effective multimedia design and apply them to the introductory geoscience classroom, instructors of the target course (MEA 101) began developing videos to be used in the Fall of 2014 (Dixon & McConnell, 2014). These GeoScience Videos are short, content-related videos designed to convey foundational concepts of introductory geoscience. The videos are typically 5-7 minutes long and follow a standard format consisting of a mix of learning objectives, basic content topics, brief text
coupled with images (e.g., maps, diagrams, models, geologic features) and/or embedded video clips (e.g., demonstrations), a formative assessment, a summary reflection activity, and conversational narration throughout (which is also available as closed captions). In addition to being embedded in the required course homework assignments, videos were also available publically on YouTube for students to review (Dixon & McConnell, 2015; Dixon, 2016).

This study sought to investigate the effects on student performance and confidence of partially-flipping an introductory geoscience course via the introduction of GeoScience Videos prior to class meetings. Over four study semesters, course pre-class reading assignments were gradually replaced with video-based assignments. This change resulted in more time becoming available in class to address challenging concepts, and to utilize additional active learning activities to support learning. Specifically, we sought to determine how the increase in pre-class video content across the study semesters affected:

1. Student performance and mean level of confidence related to concepts now communicated outside of the classroom.
   a. It is hypothesized that there would be no significant differences in student performance or confidence on summative exam questions based on content that was now delivered outside of class due to the efficiency (both cognitively and procedurally) of multimedia learning.

2. Student performance and mean confidence related to concepts that were covered more extensively in class as a result of the flipped format.
   a. It is hypothesized that students would perform better on such concepts due to the positive effects of active learning strategies in science courses.
We seek to generalize any findings and make recommendations for practice to instructors who are considering pursuing a flipped model of instruction.

2.2 Methods

To investigate these research questions and to further characterize the effects of adopting a flipped model within an introductory geoscience course, we compared student performance results and confidence measures across four consecutive Fall semesters of a section of a physical geology course (MEA 101) at North Carolina State University (NCSU). Efforts were made to control for many situational factors affecting each cohort in an attempt to mitigate confounding variables. Each target class shared the same instructor, was offered during the same academic semester (Fall) and time of day (early afternoon), and measured student performance and confidence via equivalent exam questions. The format (readings vs. video) or course homework assignments and the platform for the delivery of online quizzes were altered during the study period, but course content covered in pre-class (homework) assignments and during class meetings were essentially equivalent across study semesters.

Participants and setting: Participant cohorts consisted of four convenience samples of undergraduate students from NCSU enrolled in a class section of an introductory physical geology course and spread across four Fall semesters (2013, 2014, 2015, and 2016). The class met two times a week for 75 minutes each. Students self-selected into the course via enrollment and class sizes ranged from 77 to 94 students. Students varied in academic rank and in age, with the majority of students being either freshman or sophomore (>75% in each semester) and non-STEM majors (approx. 75% in each semester; Table 2.1). As is
historically the case for this course, each semester’s sample population contained a male majority (Table 2.1).

**Table 2.1** Participant gender, academic rank, and declared major

<table>
<thead>
<tr>
<th></th>
<th>Fall 2013 (n=93)</th>
<th>Fall 2014 (n=94)</th>
<th>Fall 2015 (n=77)</th>
<th>Fall 2016 (n=91)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>count</td>
<td>%</td>
<td>count</td>
<td>%</td>
</tr>
<tr>
<td>Male</td>
<td>67</td>
<td>71.28</td>
<td>70</td>
<td>74.47</td>
</tr>
<tr>
<td>Female</td>
<td>26</td>
<td>27.66</td>
<td>24</td>
<td>25.53</td>
</tr>
<tr>
<td>Freshman</td>
<td>63</td>
<td>67.02</td>
<td>52</td>
<td>55.32</td>
</tr>
<tr>
<td>Sophomore</td>
<td>14</td>
<td>14.89</td>
<td>22</td>
<td>23.40</td>
</tr>
<tr>
<td>Junior</td>
<td>7</td>
<td>7.45</td>
<td>9</td>
<td>9.57</td>
</tr>
<tr>
<td>Senior</td>
<td>7</td>
<td>7.45</td>
<td>11</td>
<td>11.70</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>2.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STEM</td>
<td>27</td>
<td>28.72</td>
<td>23</td>
<td>24.47</td>
</tr>
<tr>
<td>non-STEM</td>
<td>66</td>
<td>70.21</td>
<td>71</td>
<td>75.53</td>
</tr>
</tbody>
</table>

**Course alterations:** *Pre-class Learning Journal Activities* – Students were expected to complete an online homework assignment known as a learning journal before attending class each day (except for days with exams). Scores on the learning journal activities accounted for approximately 20% of the course grade during each semester. In Fall 2013 students would read assigned pages (~ four pages in length) from the textbook to provide them with requisite background knowledge and basic content understanding prior to course meetings. Students would then answer a number (~3-5) of content related questions in a quiz on the online course management system (CMS, Moodle) to assess their learning. The quiz was a combination of question formats (multiple choice, true/false, short answer). At the beginning of the corresponding physical meeting of the class, students would be asked a number (approx. 2-4) of related content questions which they would answer via a classroom
response system (“clickers”). The results would then be displayed and discussed during the first few minutes of the class period.

During the Fall 2014 semester, eight short (~6 minute), content-related videos were added into the course to directly replace some of the required reading assignments. These videos were embedded into required learning journal assignments administered via the course CMS. The videos made it possible to introduce more content than readings alone and quiz questions were adjusted accordingly but the character of the questions remained consistent. As more videos were produced during the following year, this number increased to fifteen videos for the Fall 2015 course. This number of videos was maintained for the Fall 2016 semester.

*In-class activities* – The incorporation of videos resulted in students being tasked with learning more content outside of class. Consequently, during class, several physical geology concepts could be addressed in more depth than during previous class lessons and/or more active learning activities were added to the in-class meetings. While many of these activities were piloted in Fall 2015, a full suite of added in-class activities considered for this study were in place for the Fall 2016 semester. Consequently, the Fall 2016 semester serves as the true treatment semester for comparisons of student performance, confidence, and analysis of in-class activities.

**Measures: Student performance** – Student performance was measured via three summative midterm exams. Each exam consisted of either twenty-eight (Exams 1 and 2) or twenty-nine (Exam 3) dichotomously scored (correct or incorrect), multiple choice questions (85 in total) that ranged in difficulty from lower-order recognition questions akin to the
knowledge level of Bloom’s taxonomy (Bloom et al., 1956) to relatively higher-order questions that either asked students to apply concepts in novel contexts or assessed multiple concepts simultaneously (akin to Bloom’s application level). Mean student performance for each question was calculated by summing the number of students who selected the correct answer for each question and dividing this number by the total number of students who attempted the question for each semester. Kuder-Richardson Formula 20 values for each of the exams across the target semesters were 0.6 or greater, indicating reliability and internal consistency across students on each item of each exam.

*Student confidence* – Beginning in the Fall 2014 semester, measures of student confidence were collected via individual question confidence judgments collected via inserting a continuous line under each question of the three midterm exams. The origin and terminus of the line were labelled “Not at all confident in my answer (0%)” and “Very confident in my answer (100%),” respectively. Students were then instructed to draw an intersecting line that best represented the level of confidence in their answer choice. The distance from the origin of the line to the student’s intersecting line was measured and converted to a percentage representing the student’s confidence for each exam question (Figure 2.1). This process was repeated for every student and every exam across the three semesters for which confidence measures were collected (Fall 2014-2016). Student confidence values were averaged for each question to generate a mean confidence elicited by that question across all students for each semester.
Exam question analysis – We categorized exam questions in an effort to answer the primary research questions outlined in this study and to analyze the effects of the flipped components on student performance and confidence during the target semesters of the course. Each exam question featured in the three summative course exams across all four target semesters of the course was binned into one of five categories depending on how the content related to the topic was delivered in each semester in Fall 2014 vs Fall 2016 (e.g., reading vs. video, static in-class activities vs. augmented in-class activities, etc.).

- **Category 1:** Exam questions represented topics that were first introduced to students by video in a pre-class assignment and were also excluded from the in-class presentation of content during 2016. An example of this category would be questions related to the classification of igneous rocks. In Fall 2014 the difference between volcanic and plutonic igneous rocks was introduced with a reading assignment and assessed with two quiz questions. This concept was subsequently reviewed in class along with a discussion of the silica content of igneous rocks and how geologists use texture and silica content (as represented by the color of minerals) to classify igneous rocks. In contrast, in Fall 2016, all aspects of igneous rock classification were communicated to students in a learning journal video (Naming Igneous Rocks, [https://youtu.be/Zbz4e-9pjY4](https://youtu.be/Zbz4e-9pjY4)) and assessed with six quiz questions (same two questions as Fall 2014 plus four others). These concepts
were not further discussed in class beyond some review questions at the start of the next lesson. This subsequently resulted in more class time that was devoted to discussion and explanation of partial melting processes and the rock cycle.

- **Category 2:** These exam questions assessed video-related content that was also covered in-class, but with the same level of detail as in previous semesters (i.e., constant in-class materials). Examples of this type include questions regarding metamorphism and ice age climate cycles.

- **Category 3:** This content was introduced to students in a required video and the instructor removed the related content from the in-class lesson. The instructor took advantage of the additional time made available during the resulting lessons to augment the in-class presentation during the Fall 2016 semester. For example, the class presentation of the scientific process included an additional activity that asked students to read a media account of groundwater withdrawal and work in small groups to correctly identify examples of observations, predictions and hypotheses.

- **Category 4:** This content was not included in video resources, yet was augmented in lecture during Fall 2016 as a result of removing coverage of other foundational concepts to pre-class videos. There were relatively few ($n=4$) examples of this exam question type but included questions related to hot spots, a topic that saw an added activity and the inclusion of multiple ConcepTests (conceptual multiple choice questions) as a result of newly-afforded time to cover the concept in-class.

- **Category 5:** The final category represents exam questions that remained unchanged in both course coverage and content delivery throughout the four-semester period. As we
are directly comparing Fall 2014 and Fall 2016 semesters, we note that content that was introduced via videos during the first semester videos were available (Fall 2014) were included in this category in addition to content that was not communicated to students by a video and was not part of the pre-class assignments at any point in the study. Category question quantities and descriptions are provided in Table 2.2.

Table 2.2 Exam question categories and descriptions

<table>
<thead>
<tr>
<th>Category</th>
<th>n</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13</td>
<td>Question content removed from in-class presentation; first exposure to content moved to video</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>Question content first exposed to students in video and constant in-class coverage</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>Question content first exposed to students in video AND in-class augmentation</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>Question content augmented during in-class presentation</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
<td>No changes (i.e., constant pre-class assignment and/or constant in-class coverage)</td>
</tr>
</tbody>
</table>

2.3 Statistical methods

Quantitative data relating to student performance and confidence across the study semesters were analyzed using IBM Statistical Package for the Social Sciences (SPSS). All inferential statistics were run at an alpha level of .05. Effect size considerations for t-tests ($d$) follow recommendations from Cohen (1988), with sizes being defined as “small,” ($d = .2 - .49$) "medium,” ($d = .5 - .79$) and "large” ($d > .8$). Effect sizes for Mann-Whitney U and Wilcoxon signed rank tests ($r$) were calculated by dividing the test $z$-score by the square root of the sample size for each test ($n$; Fritz, Morris, & Richler, 2012). Effect sizes for $r$ were considered “small” (.1), “medium” (.3), and “large (.5), also outlined in Cohen (1988).
**Student performance:** Initial statistical analysis of the question performance data across the four target semesters revealed that each distribution severely violated the assumption of normality (Shapiro-Wilk test for normality, \( p < 0.001 \) for each). Analysis of each distribution, however, revealed that each were similar in shape so to answer research questions pertaining to student performance across all questions in each of the target semesters, non-parametric Mann-Whitney U tests (two-tailed) were used to compare means of average student performance on questions included across the three course exams for each study semester.

**Student confidence:** After calculating descriptive statistics for data pertaining to student confidence for each question included across the Fall 2014, Fall 2015, and Fall 2016 semesters, distributions failed to reject the null hypothesis for Shapiro-Wilk tests for normality and Levene’s test for the assumption of equal variances so independent samples t-tests were performed to compare global mean confidence values across each semester.

**Question category analysis:** The paired data related to student performance and confidence for each question was also divided into the five categories outlined above related to the changes made in the course between the target semesters. First, to determine whether there were any significant changes between the true (no video) baseline Fall 2013 semester and the subsequent Fall 2014 semester (8 videos pre-class but no changes to other course activities), mean performance on each question within each category in 2014 was subtracted from the mean performance on the same question in 2013 to generate a distribution of change scores for each category. Four of these five distributions failed to reject the null hypothesis of normality for the Shapiro-Wilk test, so one-sample t-tests were performed on these data to
determine if the mean was statistically significantly different from zero. One distribution (Category 1 Performance) was non-normal, so its data were subjected to a Wilcoxon signed ranks test to confirm or refute equivalency.

To determine the net change for each of the question-related variables across semesters during which the in-class activities were altered (Fall 2016 vs Fall 2014), student performance and confidence for each question in Fall 2016 was subtracted from the recorded values for the same question in Fall 2014 to generate a value representing the change in variables between the two semesters. These delta values, grouped by question category, were used in statistical analyses. Six of the ten distributions [5 categories x 2 variables each (question performance, confidence)] failed to reject the null hypotheses for the Shapiro-Wilk test for normality (i.e., were normally distributed) and Levene’s test for the assumption of equal variances with no significant outliers so to determine if the change between semesters was significantly different, one-sample t-tests were performed on each distribution of comparison values.

The distribution for change in confidence in Category One questions was not normal and one outlier was identified via the outlier labelling rule and g-factor described in Hoaglin and Inglewicz (1987). First level winsorization (7.7% winsorized) was used for this outlier (Dixon & Massey Jr., 1969) because the parameters for trimming were not met according to Hawkins (1980). As a result of winsorization, Category One confidence mean and standard deviation decreased from -1.15 (4.91) to -2.12 (2.55). Post-winsorization, data failed to reject the null hypothesis of the Shapiro-Wilk normality test. Thus, winsorized values are used for the analysis of this category. The same procedure was followed for the distribution for two
other measures: a) Category 3 Performance, with one outlier being replaced via first level winsorization (4.6% winsorized) and its mean and standard deviation decreasing slightly from 3.03 (10.11) to 2.82 (9.64); and, b) Category 5 Performance, also one outlier being replaced via first level winsorization (3.1% winsorized), increasing mean and lowering standard deviation from 2.61 (9.66) to 3.36 (7.43).

Finally, the distribution for Category 5 Confidence contained three outliers that were replaced via first level winsorization (9.3% winsorized) slightly increasing mean and decreasing standard deviation from 0.74 (4.88) to 0.78 (4.09). Despite winsorization, the distribution rejected the null hypothesis for the Shapiro-Wilk test for normality, so a non-parametric Wilcoxon signed rank test was used on winsorized data to determine whether there was a significant change in student confidence for these questions between the Fall 2016 and Fall 2014 semesters.

2.4 Results

Comparing across the four semesters of the course analyzed in this study, there were many variables that remained consistent and many that saw improvements as a result of the changes in content delivery and in-class activities within each iteration. Mean performance and confidence as measured across all questions for each semester are provided in Table 2.3. Considering each set of question results in total across the four semesters, there were no statistically significant differences between overall student performance between the semesters as revealed by Mann-Whitney U tests, although mean performance in Fall 2016 was the highest of the study semesters. Comparing student confidence across each semester as measured by mean confidence for each question indicated by students who answered that
question, values were largely similar, but the questions in Fall 2016 elicited a statistically significant increase in confidence as compared to Fall 2015 ($t(168) = 2.03, p < 0.05$) with a small effect size ($d = 0.31$).

Beyond the similarity in student exam performance between Fall 2014 and Fall 2013, there was also no statistical difference between the performance on the questions from each of the five categories of questions between these semesters ($p > 0.05$). This establishes equivalency between true control (2013) and first implementation of pre-class video (2014) semesters but with no significant alteration to in-class activities. Remaining results pertaining to category analysis between the Fall 2014 and Fall 2016 semesters are detailed in relation to each of the study’s primary research questions.

**Table 2.3** Mean and standard deviations for question performance and confidence ($n=85$)

<table>
<thead>
<tr>
<th></th>
<th>Fall 2013</th>
<th>Fall 2014</th>
<th>Fall 2015</th>
<th>Fall 2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perf.</td>
<td>77.48</td>
<td>77.06</td>
<td>77.51</td>
<td>80.97</td>
</tr>
<tr>
<td></td>
<td>16.51</td>
<td>16.02</td>
<td>18.05</td>
<td>15.95</td>
</tr>
<tr>
<td>Conf.</td>
<td>-</td>
<td>76.86</td>
<td>75.25</td>
<td>77.86</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>7.41</td>
<td>8.20</td>
<td>8.54</td>
</tr>
</tbody>
</table>

*How did the increase of video-related content delivered outside of class affect student performance and mean level of confidence on related content?*

This particular research question is best addressed by considering the results of the one-sample t-tests conducted for the change in performance and confidence in researcher-identified Category 1 (content in videos only) and Category 2 (content in video and no change in class lessons). For questions assessing concepts that were no longer addressed during in-class activities or lectures (Category 1), there was no significant difference in student performance ($t(12) = 1.17, p = 0.26$), although the mean of the difference was
positive (+2.1 %). This is in contrast to student confidence on these same questions, which showed a statistically significant decrease ($t(12) = -2.98, p < 0.05$) of similar magnitude (-2.1%). For Category 2 questions relating to concepts introduced to students via a video in pre-class work with constant in-class activities between the two semesters, there was a significant increase in student performance ($t(13) = 2.30, p < 0.05$), with students performing 5.2% better on these questions as compared to Fall 2014. Category 2 confidence showed a small increase (+1.3%) that was not statistically significant.

![Bar chart showing mean percentage of correct responses for questions from Category 1 and Category 2 for the Fall 2014 and Fall 2016 semesters.]

*Figure 2.2* Mean percentage of correct responses for questions from Category 1 and Category 2 for the Fall 2014 and Fall 2016 semesters

*How did the increased use of video-related content affect student performance and mean confidence related to concepts that were covered more extensively in class as a result of the flipped format?*
Exam topics that were covered more extensively in-class as a result of moving foundational content to video represented questions in Category 3 and Category 4. Questions that were related to concepts first introduced via video and subsequently augmented during in-class meetings were not significantly different in Fall 2016 than Fall 2014 ($t(21) = 1.37, p = 0.18$), although students did perform better on average ($M=2.82, SD=10.10$). Confidence for Category 3 questions also increased ($M=1.58, SD=4.29$), but not at a statistically significant level ($t(21) = 1.73, p =0.097$). Category 4 questions featured concepts that were augmented in-class but were not introduced to students via a video assignment. There was a significant increase in student performance ($t(3) = 3.81, p < 0.05$) despite a small sample size ($n = 4$) with students scoring on average 20% better on these questions. Measured student confidence also significantly increased for these questions as mean confidence increased by nearly 6% ($M=5.76, SD= 1.99; t(3) = 5.77, p < 0.05$).
How did students respond to questions on which there was no change in presentation?

Finally, in response to questions in Category 5, which had no changes in either pre-class content delivery or in-class activities, students performed significantly better in Fall 2016 than Fall 2014 ($t(31) = 2.56$, $p < 0.05$), increasing their performance on these 32 questions by an average of 3.36% ($SD = 7.43$). Confidence for questions in Category 5 was unchanged between the semesters with a Wilcoxon signed rank test revealing a difference insignificant from zero ($W = 359$, $p = 0.08$). Table 2.4 reports means and standard deviations of each category’s change between semesters. Figure 2.4 displays change scores for each category.

**Figure 2.3** Mean percentage of correct responses for questions from Category 3 and Category 4 for the Fall 2014 and Fall 2016 semesters.
Table 2.4 Means and standard deviations of category change scores

<table>
<thead>
<tr>
<th></th>
<th>Category 1</th>
<th>Category 2</th>
<th>Category 3</th>
<th>Category 4</th>
<th>Category 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>F2014-F2013</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance M</td>
<td>-1.82</td>
<td>-2.47</td>
<td>2.41</td>
<td>-3.92</td>
<td>-2.01</td>
</tr>
<tr>
<td>SD</td>
<td>7.18</td>
<td>9.47</td>
<td>7.89</td>
<td>10.66</td>
<td>7.29</td>
</tr>
<tr>
<td><strong>F2016-F2014</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance M</td>
<td>2.12</td>
<td>5.20*</td>
<td>2.82</td>
<td>20.25*</td>
<td>3.36*</td>
</tr>
<tr>
<td>SD</td>
<td>6.52</td>
<td>8.44</td>
<td>9.64</td>
<td>10.64</td>
<td>7.43</td>
</tr>
<tr>
<td>Confidence M</td>
<td>-2.12*</td>
<td>1.29</td>
<td>1.58</td>
<td>5.77*</td>
<td>0.78</td>
</tr>
<tr>
<td>SD</td>
<td>2.56</td>
<td>4.93</td>
<td>4.29</td>
<td>1.99</td>
<td>4.09</td>
</tr>
</tbody>
</table>

* p < 0.05

Figure 2.4 Mean percent change for each question category between treatment and control semesters (Fall 2016 – Fall 2014). Symbol (*) indicates significant change from zero at p < 0.05 level.

2.5 Discussion

**Major findings:** This study sought to isolate the effects of utilizing short, content-related videos to move foundational concepts in physical geology outside of the classroom, thus freeing up time during in-class meetings for deeper content coverage and active learning.
techniques. Results suggest a nuanced relationship between variables related to classroom flipping and one that, in many ways, highlight that the true “effects” of partial course flipping must be delineated by the goals of the instructor implementing the course changes. Is the goal to move presentation of material outside of the classroom? Or is the goal to increase student learning? The strong relationship between active learning strategies and student performance in higher education environments suggests that these pursuits must be intertwined for course flipping to truly benefit student outcomes in undergraduate STEM-related courses.

Though not statistically significant, students overall performed better in the Fall 2016 semester than in prior semesters of the course. We do see statistically significant changes in student performance and confidence when we break down the summative exam questions by category related to how content was presented both prior to class and during class. For Category 1, students maintained the same level of performance despite specific concepts not being presented by the instructor during class. Consequently, we infer that students are capable of teaching themselves some of the basic course content, provided that the multimedia they are consuming is well-aligned to empirical design suggestions (Mayer, 2003). Though performance was maintained across the semesters for Category 1 questions, there was a slight drop in student confidence. This suggests a need for more cognitive and metacognitive support for these topics throughout the course, perhaps through more formative assessment and feedback regarding how students are performing on concepts that were no longer covered in the course lectures.

Whereas performance on topics introduced in videos and removed from lecture remained constant, performance on questions related to concepts introduced in videos and
subsequently reinforced via in-class presentations and activities (Category 2) saw a significant increase between the comparison semesters. Perhaps videos have greater potential than readings to effectively engage students as they gather prior knowledge and generating their own representations of the concepts that they subsequently bring to class to serve as a foundation for instruction. The added value of supporting learning with videos is reinforced by the work of Mayer (2003), and with recent findings on the effectiveness of geoscience videos specifically in fostering learning of geology content. Wiggen & McConnell (in press) determined in a laboratory study that videos on the physical geology concepts of magma viscosity and fault identification were more effective in conveying information than equivalent reading passages from geoscience textbooks as measured by pre- and post-tests on the content. One such question included in Exam 1 of this study but related to the content covered in the video investigated in Wiggen & McConnell (in press) saw over a 23% improvement in performance (and 10% more confidence) when comparing performance on the question between Fall 2016 and Fall 2014. These results support video usage in this context.

There were mixed and unexpected results for exam questions related to topics that were introduced in videos and augmented in lecture. One category of questions, whose content was not introduced to students in videos (Category 4), exhibited large improvements in both student performance and confidence. Additionally, for these questions (as they are few), it is possible to isolate specific strategies that were employed during in-class meetings that likely helped contribute to the exhibited gains. Two of these questions were related to the concept of hot spots, with in-class augmentations on the topic in Fall 2016 coming in the
form of additional ConcepTest questions that were afforded via the flipping process. Later summative examination of the concept saw gains in performance on two related exam questions of 24.7% and 32.7% when compared to the Fall 2014 semester performance. This provides support to prior studies that claimed benefits seen in flipped courses may merely be those of active learning strategies (Jensen et al., 2015). Importantly, however, this effect was not homogenous across questions with augmented in-class activities. Category 3 questions, representing video topics that saw more depth or activities in-class for the Fall 2016 semester, were not significantly different than in Fall 2014 (although results increased on average). This suggests that even though pre-class videos can be more effective than text-based assignments in communicating requisite knowledge (as evidenced by results for questions in Categories 1 and 2), the diverse nature of the discipline, of the strategies that can be used in-class to communicate the discipline, and of students in general dictates that there may be a differential effect of these variables on the learning of geology content. The video used to communicate baseline knowledge, the activities used to build upon this foundation, or receptivity of student may all affect student learning. Others have reported that not all multimedia will have an equal benefit to learning as measured by an assessment (Chen et al., 2010). Further investigation into this line of inquiry and further isolation of effective strategies relating to individual topics of the discipline, however, is an important avenue for future work.

Finally, results from questions in the final category of analysis (Category 5) suggests that in spite of course design elements, there will always be exam questions and course concepts to which students will perform differentially. Although these questions saw a small
yet significant mean increase comparing Fall 2016 and Fall 2014 (3%), it is difficult to speculate to gains measured in these questions as no variables were altered in reference to these concepts within this study. However, it may be that students were able to devote more time and attention to studying and learning these concepts because of increased efficacy for the remainder of course content.

**Limitations:** Although this study did find support for partial classroom flipping at the college level, there are many limitations to consider in the interpretation of results. Though multi-semester-long studies conducted in situated learning environments can provide important insight into student learning in a real-world sense (as opposed to laboratory settings), there are many variables that cannot be controlled. Although this study attempted to control many of the situational factors related with the course setting (instructor, time of year, time of day, exams, etc.) implications and notions of causation should be approached with caution. Additionally, although each dataset was relatively robust in terms of its origin (student responses for 85 questions across four semesters), the categorization of questions cannibalized the statistical power of the tests. Though many results of the study’s statistical tests were still significant, it is important to consider these smaller samples in the interpretation of results.

As students were not given pre-measures or post-measures at the beginning and end of each semester, gains seen during the Fall 2016 semester could simply be due to the sample population of students having increased prior knowledge of the discipline or increased motivation for learning science and/or geology. Additionally, the target course, even in its Fall 2013 iteration, is one that has been designed to include a suite of research-based teaching
practices and student-centered pedagogy and as such there may be a ceiling effect for student performance in the course (mean performance for Fall 2016 was above 80%). Though students performing too well is certainly not a limitation in itself, it is important to consider this relatively high performance in the consideration of results. Future work will seek to control for prior knowledge and teaching practices to better isolate effects of target variables.

Finally, though many of the features of the course were equivalent across the four iterations of the course, students were not subject to identical homework assignments or formative assessment support, introducing the potential for confounding variables relating to student performance and confidence in the course. Platforms offering students optional formative quizzes was experimentally altered for a concurrent project investigating student metacognitive monitoring accuracy (Jones & McConnell, 2016), which also may have potentially influenced student performance on exam questions during the Fall 2016 semester.

**Recommendations/Future work:** In sum, this study provided support to the notion that assigning students the responsibility to learn introductory physical geology concepts outside of the classroom via consumption of short-content videos does not hinder their learning of the concepts as long as the target course utilizes well-designed videos that adhere to empirically-supported aspects of multimedia design and uses time gained from the flipping process to utilize active learning strategies during in-class meetings. In other words, effective flipping practices should add depth, complexity, and student activity to the target course, not merely shift static actions to different settings. It is suggested that future investigation into the effect of flipping on student performance should include further controls of student-level variables such as prior knowledge, motivation and multimedia consumption, and further
strive for the best balance of in-class active learning strategies and out-of-class interactive multimedia learning.
CHAPTER 3: EVALUATING THE EFFECT OF ITERATIVE COURSE CHANGES ON STUDENT METACOGNITIVE MONITORING ACCURACY AND ACHIEVEMENT

3.1 Background

Metacognition is one’s ability to recognize the workings and characteristics of their own knowledge and thought processes (Flavell, 1979) and is generally separated into the two distinct components of knowledge of cognition and regulation of cognition (Schraw, 1998). Within regulation of cognition, there exist three processes, planning, monitoring, and evaluation (Jacobs & Paris, 1987). The concept of ones “self-regulation” represents the cycling of these processes during a learning task. Functionally, the cycle is the sum of a student’s awareness and knowledge of their own thinking (metacognitive awareness) and their approach to monitoring and management of their thinking and their control over motivations and behaviors related to learning (Zimmerman, 2008). A student’s self-regulatory behaviors influence how the student interacts with and internalizes course content. Self-regulation provides a student with the ability to isolate important information, identify gaps in their knowledge, and inform their decisions on what to study, when to study, and how to study in order to fill identified deficiencies (Zimmerman, 1990). Importantly, these abilities have been shown to be malleable and responsive to explicit and sustained training and feedback (Schraw, 1998; Nietfeld & Schraw, 2002; Nietfeld et al., 2006; Gutierrez & Schraw, 2015; Callender et al., 2016).
A student’s metacognitive monitoring accuracy, their ability to monitor their knowledge and to accurately compare this to how they perform during an assessment task, is an important learning skill that correlates well with performance both from analyses of self-report measures provided by learners (Zimmerman, 2008) and from work comparing student perceptions and performance (Kruger & Dunning, 1999; Bol & Hacker, 2000). This monitoring accuracy, however, is not guaranteed. In fact, it is sometimes faulty and has been shown to be independent from general ability in a discipline (Pressley & Ghatala, 1990; Schraw, 1998). Within the discipline or science, students with better metacognitive skills outperform others in introductory undergraduate biology (McCrindle & Christensen, 1995), chemistry (Hawker et al., 2016), and psychology (Hacker et al., 2000; Nietfeld, 2005; Nietfeld, 2006). This pattern is also evident in meta-analyses involving high school and college students (Kuncel et al., 2005).

When a student asks themselves, “Am I prepared to take this exam?” they are engaging in an act of self-regulation and attempting to monitor the quality of their understanding of the principal components of course content and to use this information to take corrective steps if needed. Unfortunately, for many students, this is not an easy question to answer, and many assess tasks incorrectly and are poor judges of their strengths and weaknesses (Kruger & Dunning, 1999; Hacker et al, 2000). That is, they are poorly calibrated. Calibration accuracy is one measure of metacognitive monitoring accuracy, and represents the absolute value of the gap between a student’s confidence and their measured performance on an individual assessment item (Schraw, 2009; Alexander, 2013). This measure is considered to be a gauge of an individuals’ “local” accuracy, often referred to as
an “on-line” measure of monitoring during a task (as opposed to before or after the task; Nietfeld et al., 2005).

Practically, for most students in situated (i.e., classroom) education settings, a recognition of any potential gaps in monitoring will not become evident until after they have completed some form of summative assessment (e.g., midterm exam) and have received performance feedback. Given the situational factors associated with higher education environments, it is often too late to go back and address identified deficiencies after an exam as the focus of the course moves on to new concepts. High-performing students generally exhibit high confidence and low calibration values (i.e., a small gap between the prediction of their performance and their actual result), and tend to slightly underestimate their performance (Kruger & Dunning, 1999). In contrast, low-performing students generally exhibit the widest calibration gaps as they are more likely to be over-confident in their knowledge (Kruger & Dunning, 1999; Hacker et al., 2000; Hadwin & Webster, 2013). Often these low performing students do not recognize that they are poorly calibrated, and therefore do not take steps to self-regulate their performance and close their calibration gap (Dinsmore & Parkinson, 2013).

Negative effects are not isolated to low-performers, high performers who may be uncalibrated may use study time unproductively to review material that they have already learned (Maki et al. 2005, 2009). Additionally, whereas calibration accuracy represents the unsigned value of this gap between a student’s performance and their confidence, bias takes into account the directionality of this disparity, thus communicating whether it represents over-confidence (a positive value) or under-confidence (a negative value; Schraw, 2009).
Calibration and bias represent important metrics for measuring metacognitive monitoring accuracy in students. We used the analysis of calibration values of local monitoring to investigate the calibration gap as related to the level and accuracy of student learning in an introductory geology course.

In addition to item-level, on-line monitoring accuracy measured during each question of a task, many studies on calibration accuracy take measures of global monitoring (e.g., Hacker et al, 2000; Bol et al., 2012; Callender et al., 2016; Hawker, 2016). Global monitoring measurement asks students to predict (before they see any items of an assessment) or postdict (as the last question of an assessment) the score they anticipate receiving for the assessment (Dinsmore & Parkinson, 2013; Pieschl, 2009). Global measures allow students to encapsulate their feelings about the assessment and an approximation of their performance in a singular judgement of confidence that can then be compared to their eventual score on the assessment to generate global values of calibration accuracy. Global estimates are generally more accurate than local measures of accuracy (Schraw, 1994; Nietfeld et al., 2005). Similar to local measures, global accuracy can be reported as an absolute value of the disparity between confidence and performance or as a signed bias value which can then be interpreted as related to study goals (e.g., Bol et al., 2012).

While many of the skills associated with metacognition are often considered to be domain-general, it is presumed that higher levels of content knowledge assist the acquisition and use of metacognitive skills (Schraw, 1998). Given our interest in promoting the discipline of the geosciences, it is important to investigate the domain specificity of metacognitive awareness. Most commonly, however, when metacognition is studied in the
specific application of science education, it is studied via prompting of students during science instruction (Zohar & Barzilai, 2013) and few studies have looked intensively at monitoring accuracy in the discipline, with none investigating the concept as applied to learning the geosciences. Towards this end, this study seeks to begin to characterize student monitoring accuracy in this domain at the undergraduate level.

**Baseline data collection:** To begin to investigate characteristics of students’ monitoring within an undergraduate introductory geoscience context, preliminary data was collected to serve as the baseline for future investigations into potential discipline-specific monitoring trends within the geosciences. During the Fall 2014 semester, item-level confidence judgments were collected via a continuous scale under each question (Figure 2.1) of three paper-based midterm exams administered as part of an introductory physical geology course at North Carolina State University (NCSU). Simple regression analysis revealed a high correlation between students who averaged higher scores across all three exams and lower calibration values ($r^2=0.79$; Figure 3.1), supporting results of domain-general literature in educational psychology (Kruger & Dunning, 1999; Bol & Hacker, 2000; Nietfeld et al., 2005). Additionally, as no feedback or training was provided regarding student judgments throughout the semester, it was also found that students did not become more accurate on later exams by virtue of repetition, a result consistent with previous work (e.g., Bol et al., 2005). Similar findings regarding high performers having more metacognitive skills have been obtained in the geosciences by instructors who utilized knowledge surveys to evaluate students’ metacognitive knowledge (e.g., Nuhfer & Knipp, 2003; Wirth & Perkins, 2005).
Additionally, there is a defined issue regarding gender inequality in the pursuit of STEM learning with female students not only less likely to select a STEM field as a major but also less likely to persist in that major (NCES, 2009). Consequently, we sought to investigate any gender disparities in monitoring accuracy and performance as female participants progressed through each semester in hopes to further develop activities to support equity in future geoscience courses.

Figure 3.1 Mean performance vs. mean calibration for the baseline Fall 2014 semester

Research Questions: Building upon this baseline correlation between the performance and local monitoring accuracy, this study sought to further elucidate potential relationships between these variables and iteratively begin attempts to positively affect
students’ monitoring accuracy and potentially improve student learning outcomes. To begin to address this challenge, we developed the following research questions moving into the experimental semesters of the study. In general, this study strives to investigate the iterative effects of two instructional interventions: a) Implementing metacognitive monitoring checkpoints (prompts) into students’ required homework assignments (22 in total) during the Fall 2015 semester; and, b) Altering the platform of optional formative assessment via the adoption of a web-based quizzing tool that automatically communicates the performance results and monitoring accuracy feedback. Specifically, we sought to investigate the following questions:

1. What are the effects of the instructional interventions on students’ mean exam performance between each semester (e.g., Fall 2015 vs Fall 2016)?

2. What are the effects of the instructional interventions on students’ mean local and global monitoring accuracy between each semester (e.g., Fall 2015 vs Fall 2016)?

3. What are the effects of the instructional interventions on mean performance, confidence, and local calibration accuracy among exams within each semester (e.g., Exam 1 vs Exam 2 vs Exam 3 in Fall 2016)? Was there any variation on the basis of gender?

3.2 Methods

To investigate these research questions and to further characterize the discipline-specific monitoring behaviors of learners in the setting, a situated quasi-experimental mixed-methods study was designed and implemented across two treatment semesters, with each sequentially altering an aspect of the course materials to potentially elicit and isolate effects on student performance and monitoring accuracy variables. In attempts to mitigate
confounding variables, efforts were made to control for many situational factors affecting each cohort. Each target semester shared the same instructor, was offered during the same academic semester (Fall) and time of day (early afternoon), and measured student performance via equivalent exams. Course lecture materials were slightly altered experimentally for another work pertaining to methods of content delivery and effects of partial course “flipping” (Chapter 2), but topics pertaining to physical geology covered during in-class meetings were equivalent across study semesters. This report focuses upon the quantitative data collected. Qualitative data represented by student interview results will be reported separately.

Participants: Participant cohorts consisted of three convenience samples of undergraduate students from North Carolina State University enrolled in an introductory physical geology course and spread across one baseline (2014) and two experimental (2015 and 2016) Fall semesters. Students self-selected into the course via enrollment. Students varied in academic rank and in age, with the majority of students being either freshman or sophomore (>75% in each semester) and non-STEM majors (approximately 75% in each semester). As is historically the case for this course, each semester’s sample population contained a male majority (74.5% in 2014, 59.7% in 2015 and 60.4% in 2016). From each course population, study-related subsets were included for analysis based on consent and complete datasets (provided in Table 3.2). Full details regarding all students enrolled in the course is provided in Table 3.1.
Table 3.1 Gender, academic rank, and declared major for students in each course

<table>
<thead>
<tr>
<th></th>
<th>Fall 2014 (n = 94)</th>
<th></th>
<th>Fall 2015 (n = 77)</th>
<th></th>
<th>Fall 2016 (n = 91)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>count</td>
<td>%</td>
<td>count</td>
<td>%</td>
<td>count</td>
<td>%</td>
</tr>
<tr>
<td>Male</td>
<td>70</td>
<td>74.47</td>
<td>46</td>
<td>59.74</td>
<td>55</td>
<td>60.44</td>
</tr>
<tr>
<td>Female</td>
<td>24</td>
<td>25.53</td>
<td>31</td>
<td>40.26</td>
<td>36</td>
<td>39.56</td>
</tr>
<tr>
<td>Freshman</td>
<td>52</td>
<td>55.32</td>
<td>51</td>
<td>66.23</td>
<td>47</td>
<td>51.65</td>
</tr>
<tr>
<td>Sophomore</td>
<td>22</td>
<td>23.40</td>
<td>16</td>
<td>20.78</td>
<td>25</td>
<td>27.47</td>
</tr>
<tr>
<td>Junior</td>
<td>9</td>
<td>9.57</td>
<td>8</td>
<td>10.39</td>
<td>12</td>
<td>13.19</td>
</tr>
<tr>
<td>Senior</td>
<td>11</td>
<td>11.70</td>
<td>2</td>
<td>2.60</td>
<td>7</td>
<td>7.69</td>
</tr>
<tr>
<td>STEM</td>
<td>23</td>
<td>24.47</td>
<td>20</td>
<td>25.97</td>
<td>22</td>
<td>24.18</td>
</tr>
<tr>
<td>non-STEM</td>
<td>71</td>
<td>75.53</td>
<td>57</td>
<td>74.03</td>
<td>69</td>
<td>75.82</td>
</tr>
</tbody>
</table>

Student measures: Performance – Student performance was measured via three summative midterm exams that were distributed in time throughout the course, each assessing five to eight 75 minute lessons of geology content. Each exam consisted of either twenty-eight (Exams 1 and 2) or twenty-nine (Exam 3) dichotomously scored (correct or incorrect) multiple choice questions and two short answer questions that are not considered in this analysis. The multiple choice questions ranged in difficulty from lower-order recognition questions akin to the Knowledge level of Bloom’s taxonomy (Bloom et al., 1956) to relatively higher-order questions that asked students to apply concepts in novel contexts or assessed multiple concepts simultaneously (akin to Bloom’s Application level of complexity). Upon exam completion, student performance was calculated traditionally with performance for each question earning either a 0 for an incorrect response or a 1 for the one correct answer present in the list. All items for which a student selected the correct answer were summed and divided by the total number of items to generate a mean score for that exam. Kuder-Richardson Formula 20 values for each of the exams across the target semesters
were 0.6 or greater, indicating reliability and relative consistency of performance across students on each item of each exam.

**Local monitoring accuracy** – To measure students’ monitoring accuracy on-line during each step of the assessment process, a five-inch line was placed below each item of the paper-based exam with the origin and terminus of the line being labelled “Not at all confident in my answer (0%)” and “Very confident in my answer (100%),” respectively (Figure 3.2). Students were then instructed to draw an intersecting line that best represented the level of confidence they maintained that their selection was the correct answer to the question. Collecting this measure of student confidence allowed for the calculation of calibration accuracy values for each question. This was accomplished by measuring the length from the origin of the continuous line to the student’s intersecting confidence indicator (Figure 3.2) for each exam question for every student.

This measurement was then converted to a decimal percentage of the whole line length and compared to the individual’s performance on the question to generate the calibration value for each question of the assessment (Schraw & Roedel, 1994). Each calibration value represents the absolute value of the difference between a student’s confidence judgment and their performance on the question (a correct answer was represented by a 1 and an incorrect answer was represented by a 0; Nietfeld et al., 2006). For example, if a student was relatively confident of their answer and placed a mark 75% of the way along the line (0.75) and were determined to have scored a correct answer, their calibration score for that question would be 0.25 (i.e., |0.75 – 1.0|; Figure 3.2). Similarly, if a student had little confidence in their answer and placed a mark 15% of the way along the line
(0.15) and were determined to have submitted an incorrect answer, their calibration score would be 0.15 (|0.15 - 0|). These values were then summed and divided by the total number of items to generate exam-wide calibration scores that represent the absolute level of local monitoring accuracy demonstrated by the student on the exam (Schraw, 2009).

![Figure 3.2 Example question and calibration calculation for local accuracy judgments](image)

**Global monitoring accuracy** – Beginning in Fall 2015, student confidence related to global performance was collected via a request to make a prediction of the score they will earn on the exam after completing the assessment (a task “postdiction”; Pieschl, 2009; Dinsmore & Parkinson, 2013). Students were instructed to reply with a numerical value between 1-100. This value was then compared to their mean performance on the exam to generate a measure of their global monitoring accuracy for the task. This disparity was then recorded as both an absolute value (without indication of over- vs under-confidence) and as a signed value (thus maintaining the directionality of the discrepancy; Bol & Hacker, 2012).
These values were calculated for every student and across each course exam for both experimental semesters (Fall 2015 and Fall 2016) and averaged within each student to generate a mean value representing global accuracy across each semester.

**Course revisions:** In attempts to foster improvements in monitoring accuracy during the course of learning geology content in the target course, two primary alterations were iteratively integrated into the course, with each being added in subsequent years the course was offered:

*Metacognition checkpoints (2015):* Prior to the 2015 semester, metacognitive monitoring prompts were added to the end of required online course homework assignments. These asked students their level of confidence related to a learning objective associated with the homework assignment (e.g., On a scale between 1-100, how confident are you that you could successfully write a definition of the term "tectonic plate" in your own words?). There were a minimum of two prompts per homework assignment and eighteen assignments in total. Given the prominence of prompting in prior science education studies focusing on metacognition (Zohar & Barzilai, 2013), this addition was designed to be the first iterative step to investigate which actions could potentially elicit effects in this specific domain setting.

*CLASS (2016):* Traditionally, students in the course were provided with an opportunity to take mastery quizzes aligned to the learning goals for each unit of geology content material via the course management system (CMS) used for the class (Moodle). These quizzes would consist of ten multiple choice questions randomly selected from a quiz bank ranging from approximately 25 to 50 questions, with each attempt generating a different
set of questions. They were provided to give students an opportunity to assess their learning of the content in a no-stakes environment as the quiz attempts were not considered in the final course grade. For the Fall 2016 semester, these quizzes were replaced with equivalent quizzes (i.e., the same bank of content questions) in a web-based assessment tool developed by the authors entitled the Confidence-based Learning Accuracy Support System (or CLASS). In addition to providing students multiple choice questions to practice, CLASS automates the confidence data collection and calibration calculation processes as students complete the online quiz, allowing for immediate and robust feedback related to monitoring accuracy to be provided to students following individual assessments and globally across all quizzes that the student has completed for the course.

For students, CLASS is an account-based quiz-taking website that allows them to take quizzes created by their instructor and shared using the system and also provides access to any quizzes made public by the program creators or others. The quiz learning objectives are displayed multiple times throughout the quiz-taking process in order to frame the assessment task and allow students to make better-informed confidence judgments than on the basis of the quiz title alone. Once a student logs into the system and selects a quiz, they are transferred to the quiz-taking screen which displays the learning objectives for the quiz and asks them to make a global task prediction (e.g., “Please predict your score on this quiz”). The student uses a continuous slider bar to predict their score as a percentage, thus providing a confidence measure on the basis of learning objectives alone, prior to the student seeing any of the quiz questions (Figure 3.3). The student must make a selection before being allowed to proceed to the body of the quiz.
After making their prediction, the student is taken to the main body of the quiz where they are asked to select their responses for each question and make item-level confidence judgments using a slider bar in the same manner as described for collected paper-based data (Figure 3.4). After the student has provided responses and confidence judgments for each question included in the quiz, they are asked to make a global task postdiction during which they use the slider bar to communicate the score they think they earned (Figure 3.5). Once the student has responded to each question, made their postdiction and submitted their quiz, the attempt is auto-graded and the student is immediately provided feedback not only related to their performance (score percentage), but they are reminded of their global task pre- and postdiction judgments (Figure 3.6). Further, they are given feedback in relation to the accuracy of their local monitoring via auto-calculated calibration and bias scores (Figure 3.3 Example global task prediction within CLASS quiz attempts)
3.6). These latter values are color-coded in order to give students an idea of the quality of their accuracy on the attempt, with green representing a fairly accurate value of calibration or bias (<0.25), orange representing a fairly inaccurate value (0.25-0.5) and red representing a very inaccurate value (>0.5). The student also has the opportunity to review their answer choices for each question and the related calibration scores.

![Relative Dating and Unconformities](image)

*Figure 3.4 Sample question and item-level confidence measure within CLASS*
Figure 3.5 Global postdiction screen for a quiz within CLASS

Figure 3.6 Sample results screen automatically presented to students after taking a quiz in CLASS
Finally, all activity within the CLASS system was communicated to the student via a global results summary that provides averages of all of the metrics measured within CLASS (Figure 3.7) in addition to a graphical representation of their attempt history (Figure 3.8).

![Figure 3.7 Global results communicated to students in CLASS](image)

This scatter plot graphs the mean confidence (calculated from each item-level judgment) against their performance on each quiz attempt the student has taken in the system (similar to Hacker et al., 2000; Figure 3.8). The plot area over which these data points are graphed is color-coded similarly to the calibration and bias values, with green areas representing more-accurate and more-desirable areas, yellow representing intermediate areas, and red representing particularly inaccurate or low-performing results (Figure 3.8). These methods of providing direct and immediate feedback relating to the students’ performance
and monitoring accuracy were designed to give students an opportunity to be provided information regarding their local and global accuracy throughout the semester.

![Graphical representation of student quiz results in CLASS](image)

*Figure 3.8 Graphical representation of student quiz results in CLASS*

It is important to note that although students were introduced to the system during the first course meetings and provided information related to its use and the communication of results, there was no direct training or reflection-based intervention associated with the CLASS system for the Fall 2016 semester. Results on the effects of feedback in isolation on improving the accuracy of judgments have been mixed, with some studies seeing significance in feedback alone (Miller & Garaci, 2011; Callender et al., 2016) and others recommending either a pairing with distributed practice and/or training for effects to be seen
(Nietfeld et al., 2006; Huff & Nietfeld, 2009). Consequently, this study sought begin the process of inquiry into the effectiveness of CLASS and potentially isolate any effects of simply altering the formative assessment platform and communicating feedback to students regarding their accuracy of learning geology concepts.

3.3 Statistical methods

Quantitative data relating to student performance and local and global accuracy predictions across the study semesters were analyzed using IBM Statistical Package for the Social Sciences (SPSS). All inferential statistics were run at an alpha level of .05. Effect size considerations for $d$ follow recommendations from Cohen (1988), with sizes being defined as “small” ($d = .2 - .49$) “medium,” ($d = .5 - .79$) and “large” ($d > .8$).

**Student performance:** To answer research questions pertaining to student performance, independent sample t-tests were used to compare means of average student performance across the three course exams for each study semester. Performance data failed to reject null hypotheses for both the Shapiro-Wilk test for normality and Levene’s test for the assumption equal variances with no significant outliers.

**Local accuracy across semesters:** For measures of local accuracy (calibration), a procedure for the compiling and collation of student confidence data was followed rejecting participants with more than a fifteen-percent non-response rate due to the requirement for paired data in repeated measures designs. As a result, students who failed to complete a sufficient number of confidence judgments for questions on each exam were rejected from the larger sample population used to compare measures of global accuracy and performance due to their datasets being incomplete. For students with missing data between 0-15% of total
questions, missing calibration values for individual questions were replaced with each student’s mean calibration value for the exam. This lead to attrition for local accuracy analysis in both study semesters (30 individuals in Fall 2015 and 7 in Fall 2016). Related sample sizes for each variable are summarized in Table 3.2. All variables relating to local accuracy failed to reject the null hypotheses for the Shapiro-Wilk test for normality and Levene’s test for assumption of equal variances with no significant outliers present. To compare students’ level of absolute calibration accuracy between each study semester (e.g., Fall 2015 vs Fall 2016), independent sample t-tests were performed on students’ mean calibration values averaged across the three course exams.

**Table 3.2** Total sample size and related subsets for each study semester

<table>
<thead>
<tr>
<th></th>
<th>Performance (n)</th>
<th>Global accuracy (n)</th>
<th>Local accuracy (n)</th>
<th>Total enrollment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall 2014</td>
<td>57</td>
<td>-</td>
<td>45</td>
<td>94</td>
</tr>
<tr>
<td>Fall 2015</td>
<td>64</td>
<td>64</td>
<td>34</td>
<td>77</td>
</tr>
<tr>
<td>Fall 2016</td>
<td>55</td>
<td>55</td>
<td>49</td>
<td>91</td>
</tr>
</tbody>
</table>

**Global accuracy across semesters:** Absolute accuracy data at the global level was not normally-distributed and violated the assumption of kurtosis (2015 excess kurtosis = 5.74) and as a result the data for each semester was subjected to a square-root transformation to increase normality. This transformation reduced the range, mean, and standard deviation of the variable. While the resulting distribution was more normal than the raw distribution (reducing excess kurtosis to 0.26), the transformed data still rejected the null hypothesis of the Shapiro-Wilk normality test and one outlier was identified via the outlier labelling rule and g-factor described in Hoaglin and Inglewicz (1987). First level winsorization (1.6% winsorized) was used for this outlier (Dixon & Massey Jr., 1969) because the parameters for
trimming were not met according to Hawkins (1980). Winsorization had minimal impact on data. Transformed absolute accuracy mean and standard deviation for Fall 2015 decreased slightly to from 2.81 (1.03) to 2.79 (0.99). Post-winsorization, data failed to reject the null hypothesis of the Shapiro-Wilk normality test. Thus, winsorized scores are used for global analysis. Global signed measures of accuracy data (the signed difference between a student’s postiction judgment and their performance on each exam) were normally distributed, failing to reject the null hypothesis for both the Shapiro-Wilk test for normality and Levene’s test for the assumption of equal variances. To determine if students’ global signed accuracy represented over-confidence or under-confidence in each semester, one-sample t-tests were performed on each set of signed accuracy values for each semester.

**Local accuracy variables within semesters:** To analyze the within-subjects effect of time on the participants within each study semester (e.g., Exam 1 vs Exam 2 vs Exam 3, 2016), two-way mixed-design ANOVAs were performed for student results on each exam as a proxy for time and gender as a between-subjects factor in order to identify potential trends in how students were reacting to changes implemented during each iteration of the course and if this effect was equal across gender. Post-hoc tests used the Bonferroni adjustment for multiple comparisons. Due to the requirement for paired data across the repeated measures ANOVA design, the sample for this analysis was smaller than the total sample populations used in across-semester comparisons of performance and global accuracy (detailed in Table 3.2).
3.4 Results

In general, students demonstrated at least equivalency between the baseline semester and the two target semesters in variables related to performance and monitoring accuracy, with some positive changes evident in the Fall 2016 semester. Results are reported in relation to each of the study’s primary research questions:

1. What are the effects of the instructional interventions on students’ mean exam performance between each semester (e.g., Fall 2015 vs Fall 2016)?

Students in the Fall 2015 semester performed similarly to the students in the Fall 2014 baseline semester (M=77.2 SD=12.23 in 2014 vs M=76.4 SD=12.30 in 2015) with no statistically significant difference in mean performance present (t(119) = -0.348; p = 0.728). Students in the Fall 2016 semester, however, statistically significantly outscored their predecessors in Fall 2015 (mean difference = 4.49; t(117) = 2.204; p = 0.030) with a moderate effect size (d = 0.408). The same comparison between the Fall 2016 and Fall 2014 semester was non-significant (t(110) = 1.792; p = .076).

Though students in the Fall 2016 semester performed better on average than for Fall 2015 for all three exams, only the difference for Exam 2 was individually statistically significantly higher (t(107)=2.832, p =0.006) with a moderate effect size (d = 0.55). Additionally, Fall 2016 Exam 3 scores were significantly higher than Exam 3 scores in Fall 2014 (t(110) = 2.40, p < 0.05, d = 0.41). No other comparison between the semesters was significant. Means and standard deviations relating to performance are presented in Table 3.3.
Table 3.3 Performance variable mean and standard deviation across semesters

<table>
<thead>
<tr>
<th></th>
<th>Fall 2014 (n= 57)</th>
<th>Fall 2015 (n= 64)</th>
<th>Fall 2016 (n = 55)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Exam 1</td>
<td>77.21</td>
<td>15.11</td>
<td>77.97</td>
</tr>
<tr>
<td>Exam 2</td>
<td>79.81</td>
<td>12.26</td>
<td>74.76</td>
</tr>
<tr>
<td>Exam 3</td>
<td>75.47</td>
<td>14.13</td>
<td>78.40</td>
</tr>
<tr>
<td>Student mean</td>
<td>77.20</td>
<td>12.23</td>
<td>76.43</td>
</tr>
</tbody>
</table>

2. What are the effects of the instructional interventions on students’ mean local and global monitoring accuracy between each semester (e.g., Fall 2015 vs Fall 2016)?

**Local accuracy:** When compared to the baseline semester (Fall 2014), students in the Fall 2015 semester displayed nearly identical values of mean calibration to their predecessors (Figure 3.9; M= 0.285 SD=0.09; mean difference = 0.0009; t(77)=-0.041, p = 0.967).

Additionally, there was no statistically significant difference between students’ semester-long mean calibration accuracy between the Fall 2016 and Fall 2015 semesters (t(81)=-1.212, p = 0.229), although means did decrease in Fall 2016 (Figure 3.9; M=0.256 SD=0.10, indicating a trend towards increasing accuracy).
When comparing local accuracy results across each treatment semester by exam, however, there were significant differences. Students in the Fall 2016 semester were similarly accurate to their Fall 2015 predecessors during Exam One (M=0.243 in 2016 vs M=0.266 in 2015), but were significantly more accurate for Exam Two ($t(80) = -2.459, p = 0.016, d = 0.55$; Figure 3.10). This increase in accuracy dissipated, however, for Exam Three as the accuracy values were again equivalent (Figure 3.10).
Global accuracy: There was no significant difference in absolute global accuracy between the Fall 2015 and Fall 2016 semesters, though the mean value was lower in Fall 2016. Additionally, despite square-root transformation of the data, it still rejected the null hypothesis of Levene’s test for the assumption of equal variances so adjusted degrees of freedom for the independent samples t-test were used in the interpretation of results ($t(109.108) = -1.044; p = 0.284$). Signed accuracy between students’ postdictions and their performance in 2015 showed that they were on average overconfident in their global performance across exams (Figure 3.11; M=3.20; SD=9.37), as one-sample t-tests of their mean signed accuracy were significantly different from zero ($t(63)=2.73; p = 0.0082$). This measure of accuracy improved for the Fall 2016 semester with students averaging less than one-percent global over-confidence as a group (Figure 3.11; M=0.96; SD=5.97), a difference not significant from zero via a one-sample t-test ($t(54)=1.19; p = 0.24$).
Considering each exam individually, in Fall 2015 Exam 1 (M=4.09) and Exam 2 (M=4.45) signed postdictions significantly differed from zero ($t(61)=2.067, p = 0.043$; $t(60)=3.089, p = 0.003$) whereas students were globally more accurate for Exam Three (Table 3.4). All three exams’ signed postdictions in Fall 2016 were not significantly different from zero, suggesting more accurate global judgments (Table 3.4).

**Table 3.4** Global accuracy variables mean and standard dev. across study semesters

<table>
<thead>
<tr>
<th></th>
<th>Fall 2015 (n= 64)</th>
<th></th>
<th>Fall 2016 (n= 55)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute (raw)</td>
<td>Signed</td>
<td>Absolute (raw)</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Exam 1</td>
<td>8.84</td>
<td>9.79</td>
<td>3.90</td>
</tr>
<tr>
<td>Exam 2</td>
<td>9.57</td>
<td>7.17</td>
<td>4.33</td>
</tr>
<tr>
<td>Exam 3</td>
<td>8.03</td>
<td>10.36</td>
<td>0.49</td>
</tr>
<tr>
<td>Student Mean</td>
<td>8.93</td>
<td>6.80</td>
<td>3.20</td>
</tr>
</tbody>
</table>
3. **What are the effects of the instructional interventions on mean performance, confidence, and local calibration accuracy among exams within each semester (e.g., Exam 1 vs Exam 2 vs Exam 3 in Fall 2016)?** Was there any variation on the basis of gender?

**Correlations:** Table 3.5 provides the Pearson correlations calculated from the paired datasets for the investigated variables relating to changes in students’ local accuracy and performance in each target semester. As predicted by domain-general work regarding monitoring accuracy (Kruger & Dunning, 1999; Hacker et al., 2000; Nietfeld et al., 2005; Nietfeld et al., 2006), undergraduate students learning within an introductory geoscience course exhibit strong correlations between their mean confidence, performance, and local accuracy as measured by calibration values. Furthermore, the strength of these correlations extends across paired exam-related variables, with students behaving relatively consistently across the semester in relation to their performance, confidence regarding the material, and the accuracy of their judgments. This is particularly the case in the Fall 2016 semester with $r$ values ranging from -0.72 to -0.81, $p < 0.01$ indicating a very strong negative relationship between performance and calibration for each course exam, with students performing better (higher score) being more accurate in their local judgments (lower calibration values). The relationship between student confidence and performance for each exam of the Fall 2015 semester is neither strong nor significant, whereas these relationships are fairly strong in Fall 2016. This suggests that students in the Fall 2016 semester were more correlated between their relative confidence on each question and their eventual performance regardless of their eventual accuracy (higher confidence for exam questions and a higher score on the exam).
Table 3.5 Pearson correlations for local accuracy variables

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fall 2015 (n = 34)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Exam 1 Performance</td>
<td>-</td>
<td>-.512**</td>
<td>0.223</td>
<td>.608**</td>
<td>-.381*</td>
<td>0.08</td>
<td>.487**</td>
<td>-.378*</td>
<td>0.272</td>
</tr>
<tr>
<td>2. Exam 1 Mean Calibration</td>
<td>-</td>
<td>-.843**</td>
<td>-.418*</td>
<td>.621**</td>
<td>-.640**</td>
<td>-.509**</td>
<td>.658**</td>
<td>-.629**</td>
<td></td>
</tr>
<tr>
<td>3. Exam 1 Mean Confidence</td>
<td>-</td>
<td>0.226</td>
<td>-.577**</td>
<td>.796**</td>
<td>0.337</td>
<td>-.584**</td>
<td>.659**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Exam 2 Performance</td>
<td>-</td>
<td>-.643**</td>
<td>0.216</td>
<td>.602**</td>
<td>-.399*</td>
<td>0.145</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Exam 2 Mean Calibration</td>
<td>-</td>
<td>-.748**</td>
<td>-.524**</td>
<td>.667**</td>
<td>-.527**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Exam 2 Mean Confidence</td>
<td>-</td>
<td>0.313</td>
<td>-.565**</td>
<td>.704**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Exam 3 Performance</td>
<td>-</td>
<td>-.777**</td>
<td>.548**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Exam 3 Mean Calibration</td>
<td>-</td>
<td>-.824**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Exam 3 Mean Confidence</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fall 2016 (n = 49)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Exam 1 Performance</td>
<td>-</td>
<td>-.811**</td>
<td>.648**</td>
<td>.452**</td>
<td>-.501**</td>
<td>.361*</td>
<td>.521**</td>
<td>-.432**</td>
<td>.388**</td>
</tr>
<tr>
<td>2. Exam 1 Mean Calibration</td>
<td>-</td>
<td>-.866**</td>
<td>-.531**</td>
<td>.686**</td>
<td>-.602**</td>
<td>-.456**</td>
<td>.582**</td>
<td>-.614**</td>
<td></td>
</tr>
<tr>
<td>3. Exam 1 Mean Confidence</td>
<td>-</td>
<td>.357*</td>
<td>-.597**</td>
<td>.660**</td>
<td>.347*</td>
<td>-.609**</td>
<td>.686**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Exam 2 Performance</td>
<td>-</td>
<td>-.724**</td>
<td>.538**</td>
<td>.593**</td>
<td>-.556**</td>
<td>.419**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Exam 2 Mean Calibration</td>
<td>-</td>
<td>-.780**</td>
<td>-.577**</td>
<td>.756**</td>
<td>-.659**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Exam 2 Mean Confidence</td>
<td>-</td>
<td>.397**</td>
<td>-.650**</td>
<td>.769**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Exam 3 Performance</td>
<td>-</td>
<td>-.735**</td>
<td>.508**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Exam 3 Mean Calibration</td>
<td>-</td>
<td>-.825**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Exam 3 Mean Confidence</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** p < 0.01 (2-tailed) * p < 0.05 (2-tailed)
Fall 2015: Performance - There was no significant main effect of time on student performance from exam to exam, $F(2, 64) = 1.491, p = 0.233$. That is, students did not become progressively more calibrated from Exam 1 to Exam 3. There was additionally a non-significant main effect of gender ($F(1, 32) = 1.065, p = 0.310$), although males did outperform females on average during all three exams. Beginning the consideration of gender between semesters, though students overall performed best during Exam 1, this exam exhibited the largest disparity between males and females, with male students performing on average six percent better than females.

Confidence – There was a significant main effect of time on student confidence from exam to exam, $F(2, 64) = 4.887, p = 0.011, \eta_p^2 = 0.132$. Bonferroni corrected post-hoc tests...
showed that student confidence between Exams 1 and 3 did not differ ($p = 1$), but students were significantly less confident in their performance for the second exam when compared to the first and third (both $p < 0.05$). Males were generally more confident than females although the main effect of gender was not significant, $F(1, 32) = 0.957, p = 0.335$.

Local Accuracy – Comparing mean calibration accuracy to mean performance, there was a lower correlation in Fall 2015 than Fall 2014 (Figure 3.12 vs Figure 3.1). Analyzing student results within the semester, there was a significant main effect of time on student calibration accuracy throughout the semester, $F(2, 64) = 4.713, p = 0.012, \eta^2_p = 0.128$, with post-hoc tests determining that students were significantly most calibrated in their judgments during Exam One (Figure 3.10). Post-hoc tests determined that this disparity was significant when compared to Exam Two ($p = 0.022$) but not Exam Three ($p = 0.138$). There was a non-significant main effect of gender on calibration accuracy ($F(1, 32) = 0.238, p = 0.629$), and though males were generally more accurate than females on Exam 1, that contrast effectively disappears on later exams and general trends of calibration accuracy essentially mirror each other for each exam (Figure 3.13).
Figure 3.13 Mean calibration by gender for Fall 2015 semester

Figure 3.14 Mean performance vs. mean calibration for all exams for Fall 2016
**Fall 2016:** *Performance* – Similar to Fall 2015, there was no significant main effect of time on student performance from exam to exam (Table 3.2), $F(2, 94) = 0.185, p = 0.832$, though the overall performance was better for each exam than in the previous semester. There was additionally a non-significant main effect of gender ($F(1, 47) = 0.538, p = 0.467$). Again, the students as a group performed best during exam one (as they did with Fall 2015), which again exhibited the largest disparity between males and females, however the magnitude of the disparity was lower compared to Fall 2015 with male students performing on average just three percent better than females.

**Confidence** – In contrast to Fall 2015, there was not a significant main effect of time on student confidence from exam to exam (Table 3.6, Figure 3.10), $F(2, 94) = 1.074, p = 0.346$. Students, however, were most confident for Exam One (their best performance). Males were again generally more confident than females although the main effect of gender was not significant, $F(1, 47) = 0.674, p = 0.416$.

**Local Accuracy** – Mean calibration accuracy was more-highly correlated to mean performance in Fall 2016 than Fall 2015 (Figure 3.13 vs Figure 3.12). Analyzing student results within the semester, unlike Fall 2015, there was a non-significant main effect of time on student calibration accuracy for each exam of the Fall 2016 semester, $F(2, 94) = 1.983, p = 0.143$, though students were generally better-calibrated across the three exams. The main effect of gender on calibration accuracy was non-significant ($F(1, 46) = 0.034, p = 0.854$), but there was a significant Exam * Gender interaction, $F(2, 94) = 3.787, p = 0.026, \eta^2_p = 0.076$, indicating that the calibration values for the three exams significantly differed between males and females. Follow-up tests revealed that males were statistically significantly more
accurate during Exam One than females (M=0.223 for males vs M=0.283 for females; \( t(46) = 1.834, p = 0.03 \)). This discrepancy between genders disappeared between Exam One and Exam Two, as female students increased their accuracy. Males became increasingly less calibrated throughout the semester (Figure 3.15).

![Figure 3.15 Mean local accuracy (calibration) by gender for each exam of Fall 2016](image)

### Table 3.6 Local accuracy variables mean and standard deviation from 2014-2016

<table>
<thead>
<tr>
<th></th>
<th>Fall 2014 (n=45)</th>
<th>Fall 2015 (n=34)</th>
<th>Fall 2016 (n=48)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Exam 1 Performance</td>
<td>78.81</td>
<td>14.09</td>
<td>80.88</td>
</tr>
<tr>
<td>Exam 1 Calibration</td>
<td>0.29</td>
<td>0.12</td>
<td>0.27</td>
</tr>
<tr>
<td>Exam 1 Confidence</td>
<td>78.04</td>
<td>11.40</td>
<td>76.40</td>
</tr>
<tr>
<td>Exam 2 Performance</td>
<td>80.24</td>
<td>12.17</td>
<td>77.63</td>
</tr>
<tr>
<td>Exam 2 Calibration</td>
<td>0.28</td>
<td>0.11</td>
<td>0.31</td>
</tr>
<tr>
<td>Exam 2 Confidence</td>
<td>76.80</td>
<td>11.56</td>
<td>72.55</td>
</tr>
<tr>
<td>Exam 3 Performance</td>
<td>75.56</td>
<td>14.77</td>
<td>80.02</td>
</tr>
</tbody>
</table>
Table 3.6 Continued

<table>
<thead>
<tr>
<th></th>
<th>Fall 2014</th>
<th>Fall 2015</th>
<th>Fall 2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam 3 Calibration</td>
<td>0.29</td>
<td>0.28</td>
<td>0.27</td>
</tr>
<tr>
<td>Exam 3 Confidence</td>
<td>75.81</td>
<td>76.87</td>
<td>76.81</td>
</tr>
<tr>
<td>Mean Performance</td>
<td>78.20</td>
<td>79.51</td>
<td>81.52</td>
</tr>
<tr>
<td>Mean Calibration</td>
<td>0.29</td>
<td>0.28</td>
<td>0.26</td>
</tr>
<tr>
<td>Mean Confidence</td>
<td>76.88</td>
<td>75.27</td>
<td>77.32</td>
</tr>
</tbody>
</table>

3.5 Discussion

**Major findings:** This analysis sought to characterize the relationship between the performance, confidence, and metacognitive monitoring accuracy of undergraduate students within a large-enrollment introductory geoscience course across three semesters with iterative addition of prompts and feedback regarding their metacognitive judgments. In addition to gaining insight about this relationship, we hoped to begin the process of identifying both domain-general and domain-specific practices that could be employed to provide students with information regarding their monitoring behaviors, and to potentially increase these skills through interaction with the target course. Towards this aim, we implemented two iterative changes hypothesized to make students more aware of their metacognitive monitoring. We will consider implications in relation to each of the study’s primary research questions:

1. *What are the effects of the instructional interventions on students’ mean exam performance between each semester (e.g., Fall 2015 vs Fall 2016)?*

   There was no significant change in student performance as a result of the introduction of the monitoring prompts into the required homework assignments. Students in the Fall 2015 semester performed largely in the same manner than students in the Fall 2014 semester.
aside from a lower performance in Exam 2. Students who had access to CLASS formative quizzes during the Fall 2016 semester, however, performed significantly better than students in the Fall 2015 semester (and better than Fall 2014 but not to a statistically significant level). This suggests that either these two samples of students with similar demographic characteristics were simply better students coming into the course, or that the focus on the CLASS quizzes and related-feedback affected study habits and content mastery. More investigation on CLASS and its effect on study habits and achievement is needed to support or refute suggestions of performance improvements from this study.

2. What are the effects of the instructional interventions on students’ mean local and global monitoring accuracy between each semester (e.g., Fall 2015 vs Fall 2016)?

There was no significant effect of the addition of judgments of learning prompts (“metacognition checks”) into students’ required homework assignments during the Fall 2015 semester as no measured variables showed a statistically significant change when compared to the baseline Fall 2014 semester and many of the means and standard deviations were near identical. This is consistent with other work regarding monitoring accuracy, as it has been shown that having students be reflective in their learning does not lead to increases in accuracy in later assessments (Bol et al., 2005).

There were mixed results with the introduction of quizzing through the web-based assessment designed to communicate monitoring accuracy feedback to students (CLASS) during the Fall 2016 semester. Student performance on all course exams was significantly higher in Fall 2016, with students outperforming their peers from 2015 which experienced a similar class structure and format. Though there were some interesting trends and a general
(non-significant) increase in mean accuracy between the Fall 2015 and Fall 2016 semesters, there was no defined trend towards increased local monitoring accuracy across semesters.

This disparity between the increase in performance but a lack of significant gains in local monitoring may be attributed to the focus on the practice quizzes brought about by the introduction of the CLASS quizzing system. Low- or no-stakes practice quizzing has been shown to increase student performance in many educational settings and studies (Dunlosky et al., 2013). Perhaps the introduction of the system increased students’ focus and attention on the previously extant quizzes and therefore positively affected performance. Without the explicit and distributed training suggested by domain-general work to increase students local monitoring accuracy (e.g., Nietfeld et al., 2006; Huff & Nietfeld, 2009), however, the information and feedback provided within the CLASS system regarding monitoring accuracy may have been largely ignored by students. This would provide support to the recommendation by the aforementioned authors that feedback alone may not be able to increase students’ local accuracy on-line during summative assessment tasks.

Expanding consideration to students’ global accuracy of metacognitive judgments, results from this study suggest that feedback regarding judgments may be able to increase students’ signed accuracy on summative assessment task postdictions. This result supports other work seeing increases in student accuracy via similar postdiction-based data collection with the provision of feedback (Callender et al., 2016). Again, however, given the fact that the feedback on students’ accuracy was provided via the quizzing platform and not directly addressed in course meetings nor were students forced to formally interact with this information, we are unable to speculate to the level of attention paid to the feedback.
Formalizing interactions with this feedback to identify its effect is an important avenue for future work.

3. *What are the effects of the instructional interventions on mean performance, confidence, and local calibration accuracy among exams within each semester (e.g., Exam 1 vs Exam 2 vs Exam 3 in Fall 2016)? Was there any variation on the basis of gender?*

Important to the pursuit of potential discipline-based variations in student performance and metacognitive monitoring, students in the experimental semesters of this study were best-calibrated in their local calibration accuracy during the first exam of the semester. This runs contrary to recent work in chemistry suggesting that students may enter college-level science courses with a misalignment in expectations displayed by lower accuracy during the first exam, which they were shown to correct during subsequent exams (Hawker et al., 2016). Though Hawker et al. (2016) focused upon global postdictions across two sequential semesters of a chemistry course, this trend was not observed in the global accuracy results in this study, with absolute global accuracy being either lowest during Exam 1 (Fall 2016) or second lowest (Fall 2015). When parsing gender from the total sample, however, the aforementioned pattern was indeed supported in females as they were statistically significantly less calibrated than males for Exam 1 in Fall 2016 as they were in Fall 2015 yet to a non-significant level), only to correct this discrepancy for Exam 2 and then mirror a decrease in accuracy for Exam 3 shown in males. This adjustment in women between Exams 1 and 2 in Fall 2016 vs. the relatively equal fluctuations in accuracy between genders in Fall 2015 is an important avenue for future investigation.
Expanding this consideration of gender though the rest of the study variables, while many are not statistically significant, there are general trends that warrant further study, given the importance of gender equity and its influence in females’ less frequent persistence in STEM fields at the college level (National Science Board, 2007). Towards this end, males generally outperformed females and were more confident during each of the course exams for both semesters, although it is important to note that potential interactions between gender, academic rank, and declared major were not taken into account. Females were significantly less accurate than males on Exam 1 in Fall 2016, yet corrected this inaccuracy for Exam 2. Fall 2015 accuracy was inconsistent with 2016, with Exam 2 being the least accurate for both genders, while males in Fall 2016 were increasingly less accurate as the semester progressed (though all participants were generally more accurate than their predecessors from 2015). Increasing performance and confidence in females during the process of taking a geoscience course (and potentially isolating the root of this discrepancy via qualitative research) is another potential line of inquiry moving forward.

**Limitations:** Though situated, semester-long studies provide authentic, real-world data pertaining to the target variables, they are obviously prone to the complexities of the environment in which the data is collected. As a result, changes in target variable cannot be exclusively attributed to experimental changes in the target course. For this study in particular, as no pre-measures of student content knowledge or metacognitive awareness were taken, it is impossible to say that the students in Fall 2016 (who performed statistically better and displayed significantly more-accurate signed postdictions) were not simply more skilled when they arrived in the course. This may begin to be mitigated by the similarities in
the situational factors associated with the course and the similarity of participant
demographics, but it is a definite limitation to be addressed via pre-measures in future studies
regarding the effects of the CLASS quizzing system and any interventions designed around
the feedback it provides. Additionally, future work regarding potential gender differences
should be triangulated with pre-post measures and/or further categorization of participants in
each gender in regards to confounding factors such as academic rank and declared major
(STEM vs non-STEM), an endeavor not feasible in this study due to low sample sizes.

Another limitation was the significant attrition experienced for the comparisons of
local accuracy between the semesters as a result of the requirement for paired data in
repeated measures ANOVAs. Students in the Fall 2015 semester were particularly prone to
ignoring judgments on the exams, with 30 participants being lost as a result of missing local
judgments. Given the situational factors of proctoring a course exam for a group of 90+
students, it was outside of the instructors’ ability to check all questions prior to the
submission of exams. This problem was less pervasive in Fall 2016 (7 participants lost),
perhaps due to the focus on judgments brought about by the CLASS quizzing system. This is
important to mitigate for future studies as many of the participants who did not complete the
dataset were often lower-performing students, which likely skewed considerations of
performance in the mixed-design ANOVA. This trend, however, was equivalent for each
semester.

Finally, though many of the features of the course (semester, exams, instructor,
meeting time) were equivalent across the three iterations of the course (baseline and
experimental), students were not subject to identical lectures or in-class activities,
introducing the potential for confounding variables relating to student performance in the course. Also, short, content-related videos were added to the course across the semesters that may have affect student learning of the course content. This “partial flipping” of the course was a part of a concurrent study on the effects of the practice on student performance (Jones & McConnell, in prep) and also potentially influenced the demonstrated increase in performance seen in the Fall 2016 semester. It is important to note, however, that the primary change made in the course, however, was between the Fall 2014 and Fall 2015 semester where seven short videos were added to the required pre-class homework assignments. The homework assignments related to these experimental changes were consistent between Fall 2015 and Fall 2016.

**Recommendations/future work:** As the primary goals of the study were to potentially increase performance and monitoring accuracy at both the global and local levels in the context of an introductory geoscience course in a university environment, the study saw mixed results. Though many of these outcomes were hypothesized via prior work in metacognitive monitoring, such as students not increasing their accuracy simply via the virtue of making judgments (Bol et al., 2005), the novelty of the CLASS quizzing system warranted an iteration-based approach to the investigation of its effects as a tool. Though the focus on mastery quizzing and the provision of monitoring feedback to students via the system appeared to increase student performance and signed global measures of accuracy, students’ local accuracy was not significantly affected. This may be expected, however, as there were no protocols in place to ensure that the students utilized the information the system provided. As a result, future work implementing explicit and sustained metacognitive
training or weekly reflections centered around the monitoring feedback provided by the CLASS system is needed to truly investigate the ability of the system to potentially improve geoscience learners monitoring abilities. Similar approaches have elicited improvements in psychology courses (Nietfeld et al., 2006). Additionally, to further isolate and identify impetus for the increased performance measured during the Fall 2016 semester of the study, a study containing pre- and post-measures of content knowledge and metacognitive awareness is required to isolate effects.
CHAPTER 4: CONCLUSION

This work investigated two factors affecting student learning in the geosciences at the college level: features of effective course design and pedagogy and characteristics of student learning processes. In analyzing the effect of a partial flipping of an introductory geoscience course (Chapter 2), it was found that students can effectively teach themselves introductory concepts outside of the classroom with videos. As a result, student learning and confidence (as measured by summative exams) can be positively affected if time gained by adopting the flipped model is used to increase active learning strategies during subsequent class meetings. This benefit was not universal, however, and results towards this variation present opportunities for further investigation into effective practice and particularly problematic geoscience concepts for introductory students.

The second study sought to characterize and improve student performance, confidence, and metacognitive monitoring accuracy related to their learning in a geoscience course. This was attempted through the iterative addition of metacognitive prompts into required homework assignments in one semester and the adoption of a novel, researcher-designed formative assessment tool providing accuracy feedback to students. Results indicated a significant increase in performance and certain measures of accuracy and a general trend towards increased accuracy overall. These results suggest promise for future study and that implementations of interventions designed to further the effects observed in the study may further provide support for student learning in introductory geoscience courses.
REFERENCES


Chen, X. & Weko, T. (2009). “Students who study science, technology, engineering and


classroom may simply be the fruits of active learning. *CBE Life Sciences Education*,
v.14, p.1–12.

computer-assisted learning program - a novel approach to teaching gynecological

48, #7, doi: 10.1130/abs/2016AM-279609


recognizing one’s own incompetence lead to inflated self-assessments. *Journal of
Personality and Social Psychology*, v.77, #6, p.1121–1134.

averages, class ranks, and test scores: A meta-analysis and review of the literature.


