ABSTRACT

JONES, JASON PATRICK. Recalibrating Student Learning in Introductory Geoscience Courses Through the Use of a Web-based Assessment Tool. (Under the direction of Dr. David McConnell.)

In the past couple of decades, the geoscience education community has made great strides towards investigating how to provide effective student learning experiences in the setting. Whether these investigations be student-oriented (e.g., new effective activities for student) or instructor-oriented (e.g., new evidence-based pedagogical techniques), this type of initiative is essential if we are to work towards maximizing student learning. While such teaching strategies and course design elements are useful for the instructor, however, they may not make the learning process itself explicit to the student. To help remedy this issue, we designed and developed the Confidence-based Learning Accuracy Support System (CLASS) to provide students explicit feedback related to their mastery of geology content and the accuracy of their perceptions of their abilities. For instructors, CLASS can provide novel information regarding student learning (e.g., overconfidence, underconfidence) as it is happening, with enough time to potentially intervene prior to course exams.

CLASS leverages robust evidence from education psychology regarding student metacognition and self-regulated learning (SRL). This work provides a scaffold for geoscience educators to access the theoretical frameworks of SRL and metacognition and describes how these concepts can be supported in practice via the design elements and implementation of CLASS. To investigate its effects, we applied CLASS to two different introductory geology course settings (research institution and community college) and investigated the relationship between students' judgments of their performance and their actual performance during summative exams. Each study collected student confidence data

for every question of students' midterm exams and compared this confidence to performance via multiple empirically-derived measures of the disparity between students' perceptions of their performance and their actual performance.

In addition to exam-based data, we developed and provided CLASS quizzes to each environment (with varying requirements) to provide students with feedback regarding their learning and accuracy during the target courses. Results indicated that students utilizing CLASS at the research institution viewed CLASS as a benefit to their learning. In addition, they performed better than their predecessors for the first two exams and were generally more accurate in their approximations. The two-year college setting results were more variable, however, suggesting the need for further investigation to determine the most effective strategies for impacting learners in the setting. Overall, results provide support for CLASS's potential to serve as a tool for increasing student metacognitive awareness, self-regulation and performance in undergraduate geoscience courses

Recalibrating Student Learning in Introductory Geoscience Courses Through the Use of a Web-based Assessment Tool

by
Jason Patrick Jones

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Marine, Earth, and Atmospheric Sciences

Raleigh, North Carolina
2020

APPROVED BY:

Dr. David A. McConnell Committee Chair	Dr. Karl Wegmann
Dr. John L. Nietfeld	Dr. Eric Wiebe

DEDICATION

To my wife, my sons, our family and our future.

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 geosciences, 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 (David) and the rest of the Geoscience Learning Process Research (GLPR; or phonetically "glipper") 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.

Almost five years, a Master's, and a PhD program later, Desmond is five and his second son Bear (who was born just before the first semester of this project) is two. He and his family have loved their time in Raleigh and are looking forward to continuing lifelong friendships with David, Sandy, GLPR alumni, and all those in Raleigh they have met along the way.

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 curious and intelligent son Desmond, who put up with long days and nights repeatedly asking, “When are you going to be done working, gah!” while writing this work. I want to thank my sweet and stubborn little Bear boy, whom I am so excited to see grow and learn how to say all the things he wants to eat all the time. Finally, for 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 have graciously been patient whilst living 1400 miles away with both their child and grandchildren.

Next, I would like to acknowledge my advisor Dr. David McConnell, for being an amazing mentor and selfless friend; 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 to work together moving into the future. 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 when my young family got here knowing no one, but who have since become life-long friends.

Doug Czajka, LeeAnna Chapman, Mike Pelch, and Jennifer Wiggen, over the past five years you have made this rewarding experience at NCSU that much better even though you each have moved on as you have graduated/finished I deeply appreciate your friendship and having the opportunity to learn from your individual personalities and skills. Katherine Ryker, you have been a great “research sister” and I am lucky to know and learn from you

whenever we get the chance to work together. Laura Lukes, it has been great to get to know you better and to indulge our shared passion for SRL/Metacognition in the geosciences via our current/future collaborations. Moving from GLPR before me to GLPR who will remain, Stephanie Sabatini, it has been great getting to know you these last few months and I am excited to see where your vision and ambition for geoscience education research takes you as you move through the rest of your PhD and beyond. Amanda Crenshaw, while I have known you a bit longer, it has been great having you "on the team" and getting to know you better. Good luck with everything!

Finally, I would like to acknowledge my committee for their mentorship and guidance throughout this project. Dr. John Nietfeld for his support and mentorship in furthering my scholarship of education psychology, Dr. Karl Wegmann for his support and mentorship in geology, and Dr. Eric Wiebe for his efforts towards bringing his STEM-education and technology expertise to the project.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES x

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: METACOGNITION AND SELF-REGULATED LEARNING (SRL): A REVIEW AND FUTURE DIRECTIONS FOR GEOSCIENCE 3

2.1 Introduction 3

2.2 Analysis 7

2.3 Investigating SRL in DBER contexts 23

2.4 Discussion and recommendations 32

2.5 Conclusion 37

CHAPTER 3: THE DESIGN AND IMPLEMENTATION OF A WEB-BASED TOOL TO SUPPORT SRL AND STUDENT LEARNING ACCURACY: "CLASS" 39

3.1 Introduction 39

3.2 Metacognition and self-regulated learning 39

3.3 CLASS 44

3.4 Project description 60

3.5 Methods 62

3.6 Student views of CLASS 70

3.7 Conclusions 73

CHAPTER 4: CLASS, COURSE SETTING AND COURSE STRUCTURE: INVESTIGATING BEST PRACTICES IN UTILIZING A WEB-BASED ASSESSMENT TOOL IN AN INTRODUCTORY GEOLOGY COURSE 75

4.1 Background 75

4.2 Methods 87

4.3 Results 99

4.4 Discussion 115

CHAPTER 5: CLASS AND INSTITUTION TYPE: USING A WEB-BASED ASSESSMENT TOOL TO ANALYZE THE RELATIONSHIP BETWEEN STUDENT PERCEPTIONS OF ABILITY AND EXAM PERFORMANCE ACROSS DIFFERENT TYPES OF INSTITUTIONS 123

5.1 Background 123

5.2 Methods 129

5.3 Results 136

5.4 Discussion 152

CHAPTER 6: CONCLUSION 158

REFERENCES 160

LIST OF TABLES

CHAPTER 3

Table 3.1 CLASS use, course requirements, and utilized question bank for each target semester by institution. 66

CHAPTER 4

Table 4.1 Course structure, CLASS, and gamification features by semester. 96

Table 4.2 Aggregate CLASS quiz question attempts across semesters by exam 101

Table 4.3 Performance variable mean and standard deviation across semesters..... 105

Table 4.4 Local accuracy variables mean and standard deviation from 2017-2019..... 109

Table 4.5: Wilcoxon signed ranks of global bias by performance quartile 112

Table 4.6: Global postdiction incentive bonus percentage by semester 114

CHAPTER 5

Table 5.1 Student gender, sample size and lead instructor for each target semester..... 130

Table 5.2 Exam features and example reliability for each exam/instructor 134

Table 5.3 GCI pre- and post-test mean (and standard deviations) and gains for each institution. 138

Table 5.4 Performance variable mean and standard deviation across semesters..... 140

Table 5.5 Confidence variable mean and standard deviation across semesters..... 141

Table 5.6 Unstandardized coefficients (and standard errors) of multilevel models of calibration accuracy from RCC setting. 147

Table 5.7 Unstandardized coefficients (and standard errors) of multilevel models of calibration accuracy from the research university setting. 148

Table 5.8 Aggregate CLASS quiz question attempts across semesters by exam. 150

LIST OF FIGURES

CHAPTER 2

Figure 2.1 Theorized role of self-regulation in student outcomes (Adapted from Zusho, Pintrich & Coppola, 2003)..... 8

Figure 2.2 General composite model of Self-regulated Learning (SRL) 11

CHAPTER 3

Figure 3.1 Example question and calibration calculation for local accuracy judgments..... 43

Figure 3.2 CLASS (classforlearning.com) landing page and login screen..... 45

Figure 3.3 Mean performance vs. mean calibration for the baseline Fall 2014 semester..... 47

Figure 3.4 Example global task prediction within CLASS quiz attempts 49

Figure 3.5 Sample question and item-level confidence measure within CLASS 50

Figure 3.6 Global postdiction screen for a quiz within CLASS 51

Figure 3.7 Sample results screen automatically presented to students after taking a quiz in CLASS..... 52

Figure 3.8 Global results communicated to students in CLASS.....53

Figure 3.9 Graphical representation of student quiz results in CLASS 54

Figure 3.10 Results by learning objective table presented to students within CLASS..... 55

Figure 3.11 Dynamic quiz generation form for student use CLASS 56

Figure 3.12 Results by learning objective summary statistics presented to instructors within CLASS..... 58

Figure 3.13 Mean performance and confidence for individual learning objective questions presented to instructors within CLASS. 58

CHAPTER 4

Figure 4.1 Mean performance vs. mean calibration for the baseline Fall 2014 semester..... 83

Figure 4.2 Mean performance vs. mean calibration for the Fall 2016 semester..... 86

Figure 4.3 Example question and calibration calculation for local accuracy judgments 91

Figure 4.4 Course content grouping and exam structure for Fall 2017 semester 93

Figure 4.5 Course content grouping and exam structure for Fall 2018/2019 semesters..... 94

Figure 4.6 Distribution of CLASS quiz attempts during the Fall 2017 semester 102

Figure 4.7 Distribution of CLASS quiz attempts during the Fall 2018 semester 103

Figure 4.8 Distribution of CLASS quiz attempts during the Fall 2019 semester 104

Figure 4.9 Mean performance between each exam across each target semester 107

Figure 4.10 Mean item-level confidence between each exam across each target semester collected from course exams 108

Figure 4.11 Mean calibration between each exam across each target semester 109

Figure 4.12 Dot plot of Exam 1 postdiction values by semester 111

Figure 4.13 Global postdiction values (%) by exam and earned grade for Fall 2019..... 113

CHAPTER 5

Figure 5.1 Mean performance vs. mean calibration for the baseline Fall 2014 semester.... 127

Figure 5.2 Example question and calibration calculation for local accuracy judgments..... 135

Figure 5.3 Mean performance vs. mean calibration for the research institution Fall 2019 semester 139

Figure 5.4 Mean exam performance and item-level confidence by RCC course exam 142

Figure 5.5 Mean performance vs mean calibration for all RCC subsets and for research university Fall 2019 semester 143

Figure 5.6 Distribution of RCC CLASS quiz attempts during the Fall 2019 semester 151

Figure 5.7 CLASS attempts, mean performance, and confidence by RCC students during the Fall 2019 semester 152

CHAPTER 1: INTRODUCTION

The adoption of empirically validated instructional practices can contribute to improvements in student learning (e.g., Pollock & Finkelstein, 2008; Derting & Ebert-May, 2010; Freeman et al., 2011, 2014), increased retention rates (Russell et al., 2007), and a reduction in the achievement gap among different student populations (Haak et al., 2011; Eddy & Hogan, 2014). However, translating these practices to the classroom represents a challenging hurdle in higher education (Garet et al., 2001; Dancy & Henderson, 2010; Singer et al., 2012). Despite a trend to a greater diversity of teaching approaches, approximately half of undergraduate teaching faculty continue to rely heavily on lecture in their courses (Eagan et al., 2014; Hurtado et al., 2012).

Many students do not know how learning happens, nor what they have to do to make it happen (e.g., Nilson, 2013). Work in education psychology investigating student learning processes has suggested that improving students' knowledge of how they learn can compensate for low initial ability in a discipline (Schraw, 1998). Self-regulated learning (SRL) represents the sum of a student's awareness and knowledge of their own thinking (metacognition), their approach to monitoring and management of their thinking and their control over motivations and behaviors related to learning (Zimmerman, 2008). Utilizing technology to develop tools designed to foster metacognitive abilities and inform students' SRL behaviors represents a new and promising frontier for future teaching and learning endeavors.

Towards this goal, we developed and implemented the Confidence-based Learning Accuracy Support System (CLASS; www.classforlearning.com), a web-based tool that allows instructors of any discipline to measure students' metacognitive awareness via

confidence judgements on content-based quiz questions. This work represents both the culmination of the three-year project designed to investigate the utility of the tool and the beginning of continued development, use and research regarding its utility. It is hoped that information gained from these and future studies will be extended and generalized to provide recommendations for practice in better supporting learners within introductory geoscience courses. Contained in this volume are four works related to the CLASS project, its origins, and research projects design to investigate its use. Chapter 2 seeks to synthesize the literature on metacognition and SRL specifically for the geoscience audience. Chapter 3 describes the development of CLASS, its utility for students and instructors and provides some insight into students' views of using the tool. Chapter 4 describes a research project investigating the effects of altering the course structure and requirements surrounding CLASS within a research institution. Finally, Chapter 5 describes the first efforts into applying CLASS in a two-year college (2YC) environment.

CHAPTER 2: METACOGNITION AND SELF-REGULATED LEARNING (SRL): A REVIEW AND FUTURE DIRECTIONS FOR GEOSCIENCE

Submitted to the *Journal of Geoscience Education*

2.1 Introduction

There are many factors that can influence student learning in geoscience courses. There are differences between students such as age, gender, and previous experience. There are also many course context factors such as content, types of academic tasks, class format (e.g., face-to-face vs. online) and the reward structure (e.g., points for effort, assessment types). Finally, the nature of instruction itself can contribute to learning outcomes. Despite this, there are students in similar environments with similar characteristics and prior experiences that demonstrate significantly different results. Also, there are students who should be well placed to succeed who do poorly, and others who overcome challenges to earn a high grade. As a result, one may ask, "why do some students succeed, and others do not?" Research in educational psychology has investigated this phenomenon and provided the theoretical framework of self-regulated learning (SRL) to explain some of the disparities in student outcomes. SRL refers to the decisions and behaviors a student employs to accomplish a learning task (Zimmerman, 1990) and effective SRL decisions and behaviors have been correlated to increased learning (Panadero, 2017).

The National Research Council (NRC, 2012) have called for discipline-based education researchers to recognize the interdependence of cognitive (information processing) and affective (e.g., emotions, attitudes) processes related to SRL as a critical way to improve performance and persistence in STEM. McConnell and van der Hoeven Kraft (2011) identified the critical role affect and SRL processes play in students' use (or non-use) of

effective cognitive strategies when learning in the geosciences. SRL has recently been identified as a critical research theme and grand challenge for the future of geoscience education research and the potential to broaden participation in geoscience (St. John, 2018).

Why is a self-regulated learning model so important to the future of geoscience education? Consider a typical introductory geoscience class in which some students are high performing and others are not. SRL offers researchers and instructors a conceptual framework for understanding student learning and subsequent student success (or failure) in a manner that bridges affective, cognitive, and metacognitive domains. Instructors can then use an SRL framework to identify specific student approaches to class assignments across these domains and explicitly guide students towards adopting more effective learning strategies and tactics.

But what is the nature of student SRL in a typical geoscience classroom setting? What tools and strategies can be used to characterize a student's unique approach to SRL? How can instructors influence these student self-regulatory processes to positively impact student learning outcomes? While SRL has been a topic of interest in the educational psychology community since the 1970's (see Zimmerman and Schunk, 2011 for an overview), self-regulated models of learning have remained largely unexamined in college-level geoscience classrooms. To facilitate an expanded discussion of these issues within the geoscience education community and more broadly in DBER, the goals of this paper are to (1) provide a theoretical framework for SRL with an overview of current SRL models that can be used to situate future investigations; (2) explain how SRL processes have been documented and supported across STEM disciplines to facilitate student learning; (3) provide a roadmap that identifies previous work in this area, gaps in our understanding, and

guidelines and suggestions for future directions of SRL work in geoscience, specifically. This review is needed so that SRL models and tools can be used by researchers and instructors to improve student learning outcomes in the geosciences and to provide students with strategies that can be applied to help deepen learning in future courses. Additionally, we seek to ground consideration of SRL firmly within a geoscience context throughout this work via practical examples (hypothetical student vignettes) to help facilitate understanding of the concepts described.

Review methods and timeframe: One of the primary goals of this paper is to provide a conceptual framework for SRL in post-secondary science settings. SRL frameworks are relatively contemporary, beginning in the 1980's and producing a series of critical studies in the next two decades (e.g., Zimmerman, 2000). Work from these early SRL frameworks are considered the foundational models of SRL and form the basis of much of the work since, consequently, our review focuses on the exploration of the literature from these early works to the present.

While rooted in psychology, SRL is more recently being explored in discipline-specific contexts, including in geoscience. To capture a wider DBER lens of applied models of SRL, which potentially have more implications for future geoscience education community work on this topic, we reviewed articles from key journals in the Physics Education Research, Chemistry Education Research, and Biology Education Research communities from 2000 to present. The Geoscience Education Research (GER) community is relatively young and may be considered emerging compared to some other DBER fields (NRC, 2012; McConnell, 2019). Our review of research on SRL in geoscience started with the Earth and Mind book collections (Manduca & Mogk, 2006; Kastens & Manduca, 2012),

the outputs of the community conferences on the role of the affective domain (SERC, 2007) and metacognition (SERC, 2008), and a survey of geoscience work published within the *Journal of Geoscience Education* since 2006. Additionally, because we are exploring SRL in a post-secondary science context, we only included studies that involved participants in a post-secondary and/or discipline-based setting.

Groundwork for investigating SRL in geoscience contexts: While formal investigation into SRL behaviors of students in geoscience courses is in its early stages, the GER community has focused on related questions. Though each did not consider SRL directly, the GSA Special papers *Earth and Mind* (Manduca & Mogk, 2006) and its successor *Earth and Mind II* (Kastens & Manduca, 2012) brought attention to the relationship between cognitive and metacognitive components of learning in the geosciences and laid the groundwork for thinking about factors that affect student experiences in geoscience courses outside of the content itself (e.g., spatial skills).

Around the same time as Earth and Mind I, the geoscience education community began to place emphasis on the affective domain. The affective domain is broadly considered as the "interests, attitudes, appreciations, values, and emotional sets or biases" of learners that influence learning (Krathwohl et al., 1964). Through seminal workshops (2007, 2009) facilitated by the *On the Cutting Edge* project and scholarly works elucidating the concept of the affective domain and its role in geoscience learning (McConnell & van Der Hoeven Kraft, 2011; van Der Hoeven Kraft et al., 2011), the exploration of the affective domain in the geosciences provided the foundation for the consideration of student variables and their role in the geoscience learning process.

This early inquiry into the affective domain in the geosciences led to a group of researchers to collaborate in creating the Geoscience Affective Research NETwork (GARNET) to investigate student affect in introductory geology courses at a variety of institutions. Within the larger push towards characterizing student affect via constructs such as emotion, motivation, and connections to Earth, GARNET research also included components that were associated with SRL (particularly metacognition) and suggested that these components may be easier to facilitate in the introductory geology classroom via the integration of cognitive and metacognitive strategies into course assignments (McConnell & van Der Hoeven Kraft, 2011; van Der Hoeven Kraft et al., 2011; Gilbert et al., 2012; Lukes and McConnell, 2014). Finally, van Der Hoeven Kraft (2017) provided a theoretical framework for characterizing and potentially cultivating interest in the geosciences and highlighted the construct of interest as an influence on SRL with interested students more likely to exhibit effective SRL behavior and vice versa. In our analysis below, we will seek to distill the theoretical underpinnings of SRL and synthesize prior investigations in college science settings.

2.2 Analysis:

SRL models seek to explain why some students are successful and others are not. SRL was originally theorized as applied social cognitive theory (Bandura, 1986) in educational contexts (Zimmerman, 1990). In these SRL models of learning, self-regulation and motivation are thought to mediate the relationships between students, the classroom context, and student learning outcomes (Pintrich, 2000; Pintrich & Zusho, 2007; Winne and Hadwin, 2008; see Figure 2.1). The effects of differences between students' characteristics and the classroom environments they experience therefore may influence outcomes but may

not have a direct causal relationship. Instead, differences in an individual student's prior experiences and the nature of classroom contexts influence the student's motivation as well as how they plan and implement learning strategies (self-regulation), which in turn influence outcomes. So, any observed differences in student outcomes by age, gender, instructional methods, etc. are the result of the motivational and self-regulatory strategies and processes a student employs (Figure 2.1). Therefore, the more we can shape the learning environment to help students become better self-regulators, the more likely it is that students will become more skillful learners.

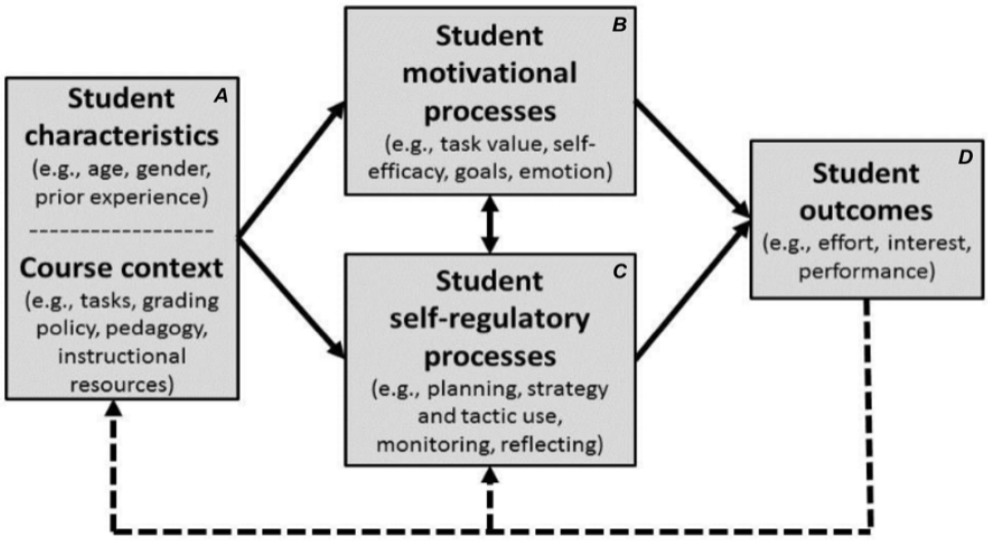

Figure 2.1 Theorized role of self-regulation in student outcomes (Adapted from Zusho, Pintrich & Coppola, 2003).

What are the characteristics of self-regulated learning? Despite some structural differences between commonly used models of SRL (e.g., Pintrich, 2000; Winne & Hadwin, 1998, 2008; Zimmerman, 2000; Panadero, 2017), there is consensus on the underlying assumptions that define the nature of self-regulated learning (see Pintrich, 2000). First, SRL is an active, constructive process. Students are agents, taking ownership of the learning process and making meaning of themselves and their environments while attempting learning

tasks. Second, students set goals related to their learning, so that learning itself becomes a goal-driven process. Third, students attempt to manage their thinking (i.e., cognition), motivation, emotions, and behavior through a series of monitoring and control processes, as dictated by their goals and the learning environment (Pintrich, 2000).

Successful self-regulated learners think about how they process information and how their learning directs their engagement: that is they are metacognitively aware of their learning processes. This metacognitive awareness is the fundamental driving force of successful SRL behaviors (Winne & Hadwin, 2008). The concept of metacognition refers to the interface between a learners' awareness and their underlying brain function (or *cognition*; Flavell, 1979). Metacognitive awareness allows students to reflect on how their learning outcomes compare to internal (e.g., "Do I understand that?") and external (e.g., grade on an exam) standards. For example, when reading a page in a textbook, a student who is metacognitively aware recognizes when they do not understand the text they are reading and decide to restart reading that page with more intention. Later, they may also recognize that they do not understand specific concepts from the reading by comparing their perceived performance to external, instructor-provided standards of performance such as a practice test or a list of learning objectives.

Discrepancies between these standards and their learning outcomes may prompt students to reflect, evaluating the effectiveness of their learning strategies, and perhaps to identify alternative strategies (Winne & Hadwin, 1998). Ideally, the results of these self-evaluations lead to regulation and/or control of their learning behaviors. Regulation is the change and the adoption of more successful learning strategies and tactics, while control refers to attempts to focus effort and attention to more successfully implement learning

strategies and tactics. These changes in behavior may vary in scale from small (e.g., rereading the page in the prior example) to more substantial (e.g., choosing to alter study setting or timing of study sessions). If there are no discrepancies between perceived success and an external standard of success, students can confirm the usefulness of their strategies and tactics and plan to continue to use them. For a more-thorough explanation of the role of standards in student self-assessment, see Winne and Hadwin (1998).

The principal assumptions outlined above can be structurally organized into a general, composite model of SRL (Figure 2.2). SRL is depicted as a recursive loop that links the commonly accepted macro-level student processes as phases: planning, action, reflection, and regulation (e.g., Greene & Azevedo, 2009). In the planning phase, students define the learning tasks, establish goals, and choose strategies and tactics to achieve those goals. These (and their variations) are examples of micro-level SRL processes. For example, students may decide that they need to know the concept of the rock cycle so they can meet their goal of earning an "A" on the exam. They choose to accomplish this goal by planning to utilize the learning strategy of re-reading their notes. In the second phase of SRL, students take action by employing their strategies and corresponding tactics (e.g., re-reading notes, choosing a quiet location so they can focus while reading), during which they monitor their perceptions of relative success during the learning task itself. In the third phase, or after the task is completed, students reflect on the perceived success of their actions.

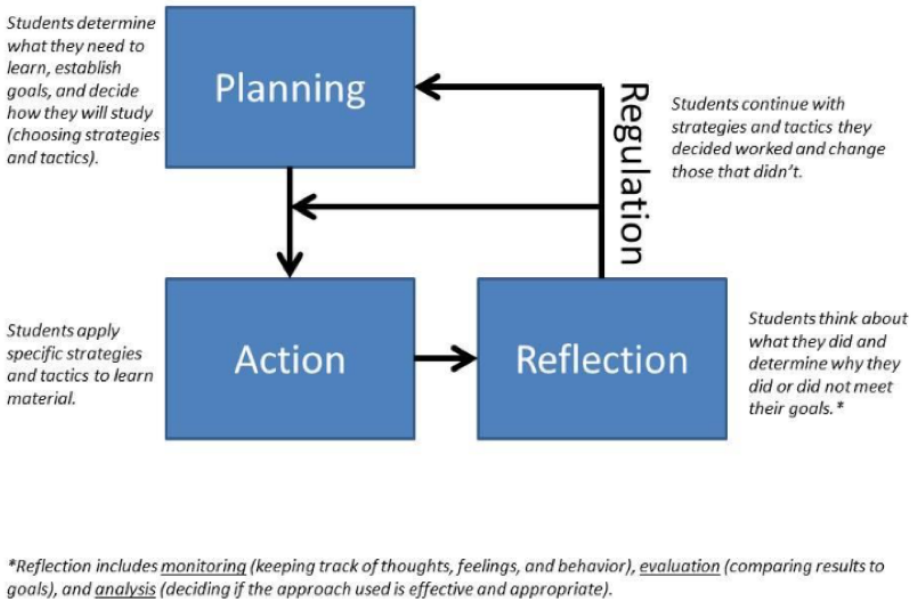

Figure 2.2 General composite model of Self-regulated Learning (SRL).

Reflection includes three sub-components: monitoring, evaluation, and analysis. Monitoring involves comparing what is being done to a standard or criterion. A student may monitor thoughts, feelings, and/or behavior related to learning, but often may not realize when they do so. For example, a student who is considering a figure in a textbook may finish examining it and determine that they just read and didn't process the information. This realization, termed a feeling of knowing or a judgment of learning, is a product of the monitoring process (Dunlosky & Tauber, 2014). Evaluation occurs when a student compares monitoring results to learning goals. In the example above, the student determined that they did not retain the information from their action. Analysis then occurs when the student examines the monitoring and evaluation results to either confirm the efficacy of learning strategy or determine if modifications and/or alternative strategies are needed.

The student considering the figure may have concluded that they did not meet their goal of learning the information because they were going through the motions of

implementing a reading strategy and were reading the words in the figure but did not attach meaning to the words. The realization that their learning goals have not been met may trigger the regulation phase in which the student seeks to improve future outcomes by changing their cognitive and affective tactics and/or adding control tactics to support their current learning strategies. For our example, the student may choose to change cognitive strategies and use a new tactic, such as sketching the figure and adding elaborative captions in their own words. Or, they may deem the learning strategy to be sufficient (e.g., reading), but they may add a control tactic, such as reading the words in the figure aloud to help or change a feature of their learning environment (e.g., turning background music on/off) to help them focus and process information more effectively.

It is important to note that the composite model of SRL presented in Figure 2.2 is not meant to be inclusive of all the micro-level SRL processes (each of the major SRL models developed by education psychology researchers have their own imagining of potential subprocesses); nor is the model meant to imply that SRL is a strictly sequential pattern. Rather, the model is meant to emphasize that self-regulation is an iterative, recursive process and that "closing the loop" through regulation is key to achieving effective student learning.

How is SRL measured? To best shepherd students towards fostering effective SRL strategies, researchers and educators must be able to accurately and reliably measure SRL behaviors. This is to not only isolate students' study skills, but to provide them useful feedback towards identifying the relative effectiveness of their behaviors. First and foremost, regardless of measurement technique and characteristics, measurement of SRL variables must be explicitly tied to a specific theoretical framework to identify and frame results (Rovers et al., 2019). Once a framework has been defined, there exist two primary

methodologies towards measuring SRL behaviors, each with their own features and demonstrated potential for inference: offline and online measures (Schellings, 2011).

Measurements are considered "online" when they are collected during a learning task itself, thus recording the actual sequence of SRL strategies and events as they occur (Schellings, 2011; Veenman, 2005). "Offline" measurements ask the subject to consider a prompt and to aggregate facets of their SRL abilities. As a result, these measurements are not tied to any specific learning task (i.e., they take place before or after any actual learning; Schellings, 2011; Veenman, 2005). It is important to note that this concept is not to be confused with an association with the internet. An "online" measurement of SRL can be taken during a learning task that has nothing to do with a computer or the internet (e.g., audio recording a learner working through a chemistry problem) and an "offline" measurement of SRL can be collected via the internet (e.g., an emailed self-report survey). Each type of measurement has been shown to have validity in measurement of subjects' SRL skills, but several disparities and important distinctions exist between the two types of measures.

Self-report surveys - The first to be significantly generated and utilized, offline measures of SRL processes are generally collected via self-report surveys of SRL variables (Rovers et al., 2019). Examples of such surveys are the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich & De Groot, 1990), the Metacognitive Awareness Inventory (MAI; Schraw & Dennison, 1994) and the Learning and Study Strategies Inventory (LASSI; Weinstein & Palmer, 2002). While each instrument has their own idiosyncrasies, each asks the subject to globally aggregate their level of agreement (usually on a Likert scale) to statements regarding aspects of SRL by asking participants how they usually approach

learning tasks (Schellings 2011; Schellings et al. 2013; McCardle & Hadwin 2015). These measures are easy to collect and allow for quick analysis via bubble sheets, etc. however, several researchers have expressed concerns with these measurements (Veenman 2005; Winne et al. 2002; Winne & Perry 2000). These concerns can be combined into two categories: 1) an incongruence between how SRL is theorized and how self-report surveys measure SRL (task specific SRL behaviors vs global SRL traits of learners); and 2) issues with subjects' ability to accurately approximate their own SRL abilities (Rovers et al., 2019).

Self-report surveys treat SRL as a stable trait that belongs to the individual and ask subjects to approximate their SRL practices to generate a global, aggregated report of their abilities via their response to a prompt (Rovers et al., 2019). However, many researchers acknowledge that SRL behaviors are often not stable. Students will utilize different strategies for different contexts or learning tasks and that they may not even use the same strategies in multiple iterations of similar tasks (McCardle & Hadwin 2015; Moos & Azevedo 2008; Winne & Hadwin 1998; Bråten & Samuelstuen 2007). The same strategies (and level of success) that a student brings to a geoscience learning task may not be the same exact strategies that they may choose to use to learn in an economics course. As a consequence, these self-report instruments generally lack the ability to answer research questions that require knowledge of students' actual SRL processes within and across learning tasks (Rovers et al., 2019). Additionally, as all theoretical models of SRL treat the construct as a procedurally-complex and cyclical interworking of planning, monitoring and regulatory behaviors (Panadero, 2017), this "clumping" of skills and past behaviors can present several opportunities for measurement error derived in the process of self-approximation of SRL behaviors by the subject.

This approximation process within the subject has been shown to impart several sources of error that emanate from the subject themselves (as opposed to theoretical issues with the instrument itself described above). This approximation has been shown to increase cognitive load on the subject causing subjects to potentially make errors as the process of taking the survey progresses (Winne, 2013). As each prompt asks subjects to consider their past behavior, they must retrieve their response from long term memory, the process of which has been shown to lead to aggregation issues as the subjects may believe their memories to be accurate when they are in fact exaggerated or even false (Perry & Winne 2006; Tourangeau et al., 2000). In addition, subjects may also feel pressured to be dishonest in their approximations due to the feeling that there is a "desired" result and that by exaggerating their report of their own skills that it would better reflect on either themselves or the researcher (Bråten & Samuelstuen, 2007). Finally, subjects may simply be unaware of their own strategy use, leading to lower levels of reported SRL in cases where its actual occurrence may be higher (Perry & Winne 2006). Similarly, subjects may incorrectly identify one strategy as another, thus leading to errors in the reported distribution of specific strategies (Veenman, 2011). These sources of error must be considered when making interpretations based on results of utilizing a self-report instrument.

One approach to overcome the complications of offline, self-report instruments is to use think aloud protocols (Winne, 2013). Think-aloud protocols are online measures of SRL where a subject is asked to work through a learning task while describing their process aloud while the researcher records the event and codes the events (e.g., Ericsson, 2006). Directly comparing the signals of a self-report instrument (Metacognitive Awareness Inventory; MAI; Schraw & Dennison, 1994) to a think-aloud protocol in evaluating the efficacy of a peer

tutoring intervention, De Backer et al. (2012) found that the MAI had no significant differences in metacognitive knowledge and regulation between the pre- and post-test. The think-aloud protocol, however, revealed significant increases in metacognitive skills in the same subjects as a result of the same intervention (De Backer et al., 2012). This highlights a need for studies to measure SRL variables both online and offline and to use both data sources to triangulate potential changes in skills. While think-aloud protocols are online measurements and have the ability capture the sub-processes of SRL and sequences of strategy use, one can imagine that the subjects are still being experimentally manipulated by the procedure and thus are not authentically engaging in learning behaviors as they would naturally (i.e., alone at home). Additionally, think-aloud protocols increase the cognitive load of the participants, thus negatively impacting the process of completing the required task (Winne, 2013). These complications posed by self-report and think-aloud measures can be potentially mitigated, however, via triangulation of SRL variables afforded by collecting and investigating trace data pertaining to subjects' online SRL activities.

The rise and subsequent proliferation of computing technology and the internet has provided new tools to researchers attempting to investigate the online SRL actions of learners in the form of the trace data recorded during the interaction with these technologies. For example, learning management systems (e.g., Moodle, Blackboard) record every user entry (e.g., click, file access, etc.) during a session. The pattern of these clicks can be investigated to elucidate the user's behaviors related to SRL strategies (Winne, 2013; Stief & Dollar, 2009). This approach records sequence of study events without biases of self-report surveys and provides insight into the user's actual activities as opposed to their (potentially faulty) recall and agglomeration of events.

What is metacognition and how is it related to SRL? Integral to effective SRL behaviors is the concept of metacognition. As mentioned previously, metacognition refers to the interface between a learners' awareness and their cognition (Flavell, 1979). Often over-distilled to the concept of "thinking about thinking," global metacognition is generally separated into two distinct theoretical sub-components: *knowledge of cognition* and *regulation of cognition* (Schraw, 1998). A learner may be metacognitively aware in regard to facets of one, the other, or both and may have unique strengths and weaknesses in each realm. *Knowledge of cognition* refers to information a learner knows about their own cognition or about cognition in general and *regulation of cognition* refers to the thinking that helps learners control their learning (Schraw, 1998). Similar to SRL, these sub-components of metacognition have been subdivided into categories or processes. While listing all of them is not within the context of this work, it is important to note three sub-processes of knowledge of cognition for future consideration of SRL. Metacognitive awareness attributed to knowledge of cognition can be either declarative (i.e., "I have a good memory"), procedural (i.e., "I know which study strategies I use") and/or c) conditional (i.e., "I have different study strategies for different situations."; Schraw, 1998). Finally, it is important to clarify the distinction between metacognition and SRL as they are often used interchangeably with one another (Dinsmore et al., 2008).

For our considerations in this paper, we adopt a theoretical relationship between metacognition and SRL similar to the Winne and Hadwin model of SRL (Winne & Hadwin, 2008) in that metacognition is the primary driver of regulatory processes. In other words, a student with metacognitive skills uses those tools to be effective at SRL. These metacognitive skills (internal thought processes of the student) are used to generate the

regulatory behaviors that the student chooses to utilize during a study task. The greater the metacognitive awareness, the more efficient and successful the cycling of SRL planning, monitoring, and evaluation becomes. In this sense, the metacognition can be considered as an internal mental process that is utilized to generate outward (SRL) behavior. Therefore, effective metacognitive knowledge and skills are necessary prerequisites for effective SRL behaviors.

To ground this distinction in a practical geoscience example, consider two college students preparing for a mineralogy lab exam. Student A demonstrates metacognitive awareness while Student B does not. A week before the exam, Student A realizes the exam is near and that her knowledge of carbonates and evaporites is relatively weak, while her knowledge of silicates is much stronger. She uses this metacognitive information to inform a SRL behavior and generate a study plan (i.e., planning phase of SRL) that incorporates a greater emphasis on her self-identified problem areas (i.e., carbonates and evaporites) and less of a focus on concepts she is more confident about. During her study, she self-tests her knowledge of carbonates and evaporite minerals by generating flash cards for each of the carbonate and evaporite minerals featured in her lab (i.e., monitoring phase of SRL) and re-evaluates her knowledge level on these groups as her performance improves. In this scenario, Student A's metacognitive awareness and conditional knowledge of what strategies and tactics to use directly influenced the features of her study practice during the planning, monitoring, and regulation phases of the SRL cycle she utilized to prepare for the exam.

Now consider the alternative scenario of Student B. Student B, without the requisite metacognitive awareness, realizes the exam is a week away and at first decides to do nothing, basing this judgment on the cue that she has attended each lab and therefore must at least

have sufficient knowledge to perform adequately on the test. When she begins to study, she decides to re-read all of her notes (an ineffective study strategy; Dunlosky et al., 2013) as many times as she can prior to the exam in hopes that the behavior will provide the knowledge necessary to succeed. After the exam, results show that Student A succeeded on the exam and Student B did not. While Student B generated a study plan, it was not based on accurate metacognitive thoughts (i.e., Student A's accurate judgments of knowledge regarding certain mineral groups) and instead was based on naive theories of knowledge construction and deficient and/or inaccurate declarative and procedural metacognitive knowledge. As metacognitive thought was not driving study behaviors, other phases of the SRL cycle were not employed (i.e., no monitoring of progress towards study goals nor evaluation of results).

Why is understanding student self-regulation important? As described above, some students may have greater metacognitive awareness and knowledge of SRL than others (Schraw, Crippen, & Hartley, 2006). Regardless of whether they do so knowingly, all students engage in some aspect of self-regulation (Winne, 1995). When completing a task such as studying for an exam, most students go through multiple SRL cycles, until they decide that they have met their goals (Winne & Hadwin, 1998). During a semester in a single course, students will engage in multiple "units" of task-level SRL cycles and will receive instructor-provided feedback. In its most basic form, feedback is represented by a grade or score on a task, while more comprehensive responses may take the form of instructor comments, a grading rubric, or sample answers. Students may incorporate these external evaluations from these tasks into their self-reflections to improve their self-regulation strategies through adaptation (Winne & Hadwin, 1998). Unfortunately, students may bring

faulty beliefs about learning strategies and tactics to the course (Karpicke, Butler, & Roediger, 2009) that can short-cut their attempts at effective self-regulation. Failure to correctly define the task initially, poor learning strategy choices, inaccurate reflections, and/or gaps in SRL cycles, can lead to poor student outcomes (Pintrich & Zusho, 2007; Figure 2.2).

Instructors can introduce activities into classes that scaffold the self-regulation skills students need to complete these SRL cycles and perform successfully. Knowledge of SRL and a delineated SRL model (Figure 2.2) can help researchers and instructors identify key measurable macro-level phases (planning, action, reflection, regulation) in student learning. Further, by gathering information related to specific micro-level components of SRL, instructors may be able to identify potential weak points for students in the SRL cycle and then select appropriate intervention strategies to help remedy identified deficiencies. Additionally, by illustrating the sequence of effective student strategies in an SRL cycle format, students and instructors can examine a more holistic picture of individual student learning processes. The cycle format allows students and instructors to get a better idea of why or how learning is or is not occurring. Instructors could potentially use the SRL model (Figure 2.2) as an academic counselling tool when working with struggling students. Researchers can use the SRL model to frame their research, design interventions, or inform measurement of SRL in geoscience settings. By providing a concrete visual for the abstract conceptual model of the SRL process, students can have an opportunity to increase their metacognitive awareness and knowledge of their learning processes. By raising student awareness, instructors can help students better understand and monitor their own learning

processes. Geoscience education research framed within SRL can seek to achieve these benefits.

SRL in collaborative learning environments: While SRL refers to the cycling of cognitive and metacognitive processes during an individual's interaction with a learning task, recently researchers have begun to investigate how SRL behaviors are utilized and constructed within group settings (Jarvela & Hadwin, 2013). Within undergraduate STEM education settings there has been an increased focus on collaborative work in the classroom via active learning strategies that have been shown to positively influence student variables (see Freeman et al., 2014 for a review). As a result, seeking to understand how group members interact during a group task has particular importance. Considering that learning does not just occur individually in these settings, researchers have viewed group interactions via a social-constructivist perspective on learning (Jarvela & Hadwin, 2013; Winne, Hadwin, & Perry, 2013) during which students collaborate and work together to construct SRL behaviors as a group entity (e.g., planning, monitoring of progress, evaluation of results) rather than solely as individuals.

Based on the Winne and Hadwin (1998) model of SRL, Jarvela and Hadwin (2013) expanded consideration of SRL into how group members interact during a group task. While effective SRL in a solo environment relies solely upon an individual's cognitive and metacognitive skills, group situations present opportunities for group members to contribute different behaviors related to the SRL cycle. Regulation in this context is theorized as a continuum from the completely solo or individual to the completely collaboratively constructed behaviors (Jarvela & Hadwin, 2013). Along this continuum are three distinct types of regulation: 1) individual SRL where all regulatory behaviors occur within the

individual; 2) co-regulation of learning (Co-RL) where peers may temporarily support one another's SRL behaviors via feedback or confirmation of SRL strategy use; and 3) socially-shared regulation of learning (SSRL) that occurs when regulatory activities are evenly-distributed and constructed amongst members within the group (Jarvela & Hadwin, 2013; Panadero & Järvelä, 2015 for a review).

Though still in its relative infancy, increased instances of SSRL within groups have been correlated to increased group criterion performance (i.e., higher grades; e.g., Janssen et al., 2012; Volet, Summers, & Thurman, 2009) and lower perceptions of task difficulty which can facilitate more effective group behavior (e.g., Hurme et al., 2009), in addition to positive affective responses (e.g., increased enjoyment; Panadero & Jarvela, 2015). However, both the original purveyors of the construct and subsequent researchers reviewing recent work have suggested that additional work needs to be done to further constrain the construct and its potential benefits for students (Panadero & Jarvela, 2015).

One may now pose the question, "what can I do as an educator to foster metacognitive abilities to better inform the SRL behaviors of my students?" One approach employed by instructors as they begin to provide support for student learning behaviors is to pose reflection questions to students to prompt them to evaluate their progress on particular concepts, learning goals, questions, etc. during class (e.g., minute papers; McConnell et al., 2017). While an important first step, work in educational psychology has demonstrated that metacognitive and SRL behaviors can be fostered via explicit and direct in-class training and feedback (e.g., Callender, Franco-Watkins & Roberts, 2016; Nietfeld, Cao & Osborne, 2006) and that this training should reflect the phases of SRL being promoted (Dignath & Büttner, 2008). Self-reflection is important, but is only one phase of the SRL cycle and if instructors

are looking to promote SRL behaviors in their students, the activities and prompting being employed should go further than simple metacognitive prompts (e.g., "what is the most important thing you learned today?") and move towards more-complete SRL behaviors (e.g., generating a study plan, providing monitoring tips, etc.). Further suggestions for how instructors can work to foster effective SRL behaviors are outlined in the "Discussion and Recommendations" section below.

2.3 Investigating SRL in DBER contexts

While SRL has been investigated broadly within the educational psychology community, investigation into SRL in discipline-based education research (DBER) settings is still emergent (NRC, 2012). While many of the cognitive and metacognitive skills that inform SRL have been demonstrated to be domain general (Schraw, 1998), other sources suggest that there is some discipline-specific variability in the application of SRL behaviors in different contexts (e.g., Schraw, Krippen, & Hartley, 2006). Towards this end, there have been several recent examples of studies investigating SRL in specific STEM-related settings. As one might imagine, these studies do not approach this type of investigation in the same way. To synthesize the findings of these efforts for this review, we will discuss them organized by their essential design elements. First we will consider studies that simply attempted to characterize college science students' SRL behaviors ("Measuring SRL without intervention") with or without correlating with other variables like performance. Then we will consider studies that went a step further to attempt to foster SRL behaviors via an experimental intervention ("Measuring SRL with intervention"). Finally, we will consider studies that not only measured SRL and attempted to foster SRL behaviors via an intervention but compared measured SRL behaviors to students' grades/performance.

Measuring SRL without intervention: Often, the first step towards understanding a phenomenon is to simply attempt to characterize it in a specific setting. Towards this end, we will now consider studies that specifically investigated students' SRL in undergraduate science education settings, using real students enrolled in real courses (as opposed to compensated volunteers in lab settings, etc.). For example, the Chemistry Education Research (CER) community has approached the question of how college students utilize SRL behaviors during chemistry learning. In the early 2000s, the originators of the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich & De Groot, 1994) partnered with faculty in chemistry at the University of Michigan to conduct an early investigation into discipline-specific SRL in two large enrollment (200+) chemistry courses (Zusho, Pintrich, & Coppola, 2003). Using the MSLQ to measure motivation and learning strategies, they found that while student motivation surrounding the course decreased as the semester progressed, their self-reported level of self-regulatory strategies increased over time (Zusho, Pintrich, & Coppola, 2003).

More recently, Lynch & Trujillo (2011) employed the MSLQ to investigate the conceptions of students toward the end of an organic chemistry course. They reported pronounced gender differences, with males reporting higher levels of SRL-related variables such as task value (Lynch & Trujillo, 2011). Overall and across genders, the authors noted correlations between the self-reported levels of SRL variables and performance, with higher reported levels of MSLQ subscales such as intrinsic motivation and task value being correlated with higher academic performance in the course (Lynch & Trujillo, 2011). Again, while important, these studies only attempted to correlate self-reported MSLQ scale variables across students and did not attempt to elicit change in the participants being studied. In

Biology Education Research (BER), Sletten used the MSLQ to investigate biology students' self-reported SRL strategy use in one large enrollment (75+) flipped biology course (Sletten, 2017). Results suggested that students' perceptions of the flipped classroom positively predicted their self-reported strategy use, but that SRL (as measured by the MSLQ) did not predict academic performance (Sletten, 2017).

In geoscience, as part of the GARNET project, Lukes (2012) used a grounded theory approach informed by student interview data, artifact analysis, and benchmarking, to characterize student choices about learning and influences on their learning, as well as participant MSLQ data, Geoscience Concept Inventory data, and academic data to explore the relationship between SRL and performance. The emergent model from the student data (n=63; 26 from two community colleges and 37 from two research-intensive universities) converged with existing SRL models of learning but provided additional characterizations of the differences in SRL processes of high and low performing students. This work suggested that there is a disconnect between course design, pedagogy, and student strategy use (relationship A to C in Figure 2.1) and that there were differences between community college and research intensive student populations in terms of timing of learning strategy acquisition and monitoring of learning (metacognitive skills relevant to SRL).

Lukes and McConnell (2013; 2014) also analyzed student interview data (n=42) through a motivation and emotion theoretical framework, characterizing a strong relationship between motivation and emotion in SRL decision-making (relationship B to C in Figure 2.1). Lukes and McConnell (2014) utilized student interviews to characterize the sources of students' motivation to study for geology exams. Interview results revealed that while all students harbored a performance goal orientation (i.e., all were motivated by their course

grade), higher performing students had higher levels of mastery goal orientations (i.e., they were motivated to master the content in addition to getting a good grade). Additionally, these mastery goal orientations were theoretically linked to more effective use of SRL strategies (Lukes & McConnell, 2014).

In contrast, there have been a few SRL studies that have been more exploratory in nature. For example, as a large portion of introductory physics courses at the college level leverage problem sets to guide student learning (e.g., Adams & Wieman, 2015), some have sought to utilize them to measure SRL variables within college environments. Mota et al. (2019) added reflection exercises to the end of students' homework assignments in an undergraduate physics course. These assignments consisted of a suite of problems related to the physics concepts being addressed in the course and were followed by five prompts (e.g., "list assistance you sought while working on this problem set"; Mota, 2019). Additional activities elicited by the added prompts included students color coding their responses for areas where they struggled and asking them to summarize the information they learned from the assignment (Mota, 2019). After coding students' responses to the added reflection exercises, the researchers found that students' metacognitive comments increased as the semester progressed, specifically comments related to knowledge of cognition (i.e., comments coded as declarative, procedural, and conditional knowledge; Mota, 2019).

As a whole, these studies demonstrate a continuing thread in extant research into college-level science students' SRL behaviors: characterization and measurement without intervention through a confirmatory approach to documenting SRL in course settings. Studies that rely solely on self-report instruments such as the MSLQ are limited by the narrow SRL content of the survey instruments and their underlying assumptions about SRL. The MSLQ

contains a limited number of statements about very specific SRL subcomponents and there has been debate about the validity of the instrument to measure the constructs it purports as a result (e.g., Rovers et al., 2019; Hilpert, Stempien, and van der Hoeven Kraft, 2013). While valuable, these correlational studies only serve as baselines in our understanding of SRL in these settings.

Measuring SRL behaviors with intervention: While correlational studies help us to characterize SRL in college science environments, if the ultimate goal is increased student learning in these environments and learning is connected to effective SRL practices, the logical next step is to attempt to foster effective SRL in students who are not exhibiting these behaviors. Many of the attempts to foster effective college student SRL behaviors through instructor/course activity intervention have been conducted in relation to the common student experience of the course exam. Many researchers working in undergraduate STEM settings have asked students to reflect on their exam preparation behaviors post-exam to gain insight into their SRL strategy use and other behaviors. Often called "exam wrappers," these are assignments that have students considering (for example) recent exam results, the study strategies the students used to prepare for the recent exam, and how they plan to alter their study habits for subsequent exams (e.g., regulation; Lovett, 2013). Within DBER settings, studies implementing versions of these instruments have primarily taken place in biology settings (e.g., Stanton et al., 2015; Smith, Metzger, & Soneral, 2019; Sabel, Dauer, & Forbes, 2017; Metzger et al., 2018). There have been a number of conference presentations related to the use of exam wrappers in geology courses to support student study behaviors (e.g., Wirth & Perkins, 2011; Perkins & Wirth, 2011; Marton, 2015; Nunez, Lukes, & Rushing, 2015). Results generally converge on two common themes: 1) variability in students' ability to

identify effective SRL strategies to use; and 2) variability in the level of which they follow through with the effective strategies they may have identified. For example, Sebesta & Smith (2017), found that students largely failed to follow through with planned study strategy changes reported in the exam wrappers for two sequential exams (Sebesta & Smith, 2017). Contrastingly, Smith, Metzger, & Soneral (2019) found the opposite was the case, with a majority of students self-reporting planned and executed changes in study strategies that led increased exam performance from one exam to the next (Smith, Metzger, & Soneral, 2019). Specifically, for geoscience, Nunez, Lukes, and Rushing (2015) reported that students were mainly overconfident in their learning and most commonly reported using the learning strategy of rereading notes (a less effective learning strategy; Dunlosky et al, 2013).

Stanton and co-authors (2015) employed "metacognitive assignments" following each of the two course exams in a large-enrollment (250+ students) biology course. The first of these was given after the first course exam for students to monitor and evaluate their study behaviors to prepare for the exam and to make a study plan for the next exam (Stanton et al., 2015). The second, given after the second exam, asked students to reflect on the extent of which they followed through with the plans they outlined in the first assignment (Stanton et al., 2015). Results revealed that approximately half of the students displayed SRL behaviors (e.g., planning future learning strategy use, monitoring effectiveness of learning strategy use) during the assignments. While almost all of the students suggested that they would change their study strategies for future exams, many did not identify learning strategies shown to be effective (see Dunlosky et al., 2013 for a review of effective learning strategies), nor did many report following through with their altered study plans in subsequent assignments (Stanton et al, 2015). While each post-exam instrument and/or procedure in these "exam

wrapper studies" had their differences, they all generally converged upon practices of comparing expected and actual exam outcomes along with explicit communication of study strategies and planned alterations to a study plan for future exams with the general goal of increasing performance.

In addition to the practice of including exam wrappers to support increased awareness of their SRL behaviors after a course exam, some have approached fostering SRL via explicit training programs. This training generally consists of providing students with specific study strategies and feedback on the efficacy of these strategies throughout the training program. For example, an experimental study by Dörrenbächer and Perels (2016) sorted students from diverse fields of study (including natural sciences) into groups that; (a) received an 8-week SRL training program focusing on strategies and their effective use; (b) completed a learning diary cataloging their study strategies; (c) both, or; (d) neither. While the control group unsurprisingly saw no change in SRL outcomes, the learning diary group also demonstrated no significant improvement. While the training group saw improvements in students' self-reported SRL, the group that completed both the training and the learning diaries reported the highest gains in SRL variables (e.g., self-monitoring; Dörrenbächer & Perels, 2016). This suggests that while more reflection-based strategies may not be enough to impact students' SRL behaviors, explicit training in effective learning strategy use, particularly in combination with reflection-based strategies, more effectively supports student SRL skill growth. Although these training approaches have been widely employed in non-discipline-specific work in education psychology (e.g., Nietfeld, Cao, & Osborne, 2006; Bellhaüser et al., 2016; Costa Ferreira et al., 2016), similar targeted training interventions in college STEM settings have yet to see such a diversity in application.

Other studies attempted to utilize a more-detailed measurement strategy towards investigating student SRL behaviors. Throughout a semester of organic chemistry, Lopez et al. (2013) gave students a number of tools (study diaries, problem sets, and concept maps) to help them navigate the semester and then collected/analyzed those tools to measure students' authentic SRL behaviors throughout the course (Lopez et al., 2013). By collecting completed tools, they were able to analyze them to determine students' study strategies and behaviors. They found that, for example, students were underusing higher quality strategies (e.g., metacognitive reflection and peer learning) and over-utilizing lower quality ones (e.g., reviewing the text; Lopez et al., 2013). While expanding the types of data collected beyond a self-report Likert-style survey, to include direct evidence of behavior through artifact analysis, they were able to gather a more-nuanced, and potentially more valid, understanding of chemistry students' SRL behaviors.

Finally, the construct of SSRL has been investigated within an online introductory geology course via the use of Google Hangouts. Spencer, Thomson and Jones (2018) embedded regulatory prompts into group activities facilitated by synchronous online chatting. Results suggested that prompting alone exerted enough of an influence on the interactions between group members to generate more examples of co-regulation (co-RL) but was not enough of an influence to generate more complex episodes of socially-shared regulation (SSRL; Spencer, Thomson, & Jones, 2018).

SRL interventions and correlations with performance: In addition to documenting the change in SRL knowledge or behaviors as the result of intervention, other researchers have implemented their interventions with the additional goal, and thus prediction, of increasing student performance in science courses. Pape-Lindstrom, Eddy & Freeman (2018)

reported the results of a multi-semester study at a two-year community college (2YC) implementing reading quizzes into an undergraduate biology course. The multiple-choice quizzes on reading assignments were presented as tools to increase course structure and support students' SRL by providing feedback during self-study that students would theoretically use to regulate their learning, therefore better preparing them for future active learning lecture sessions (Pape-Lindstrom, Eddy & Freeman, 2018). Results revealed that the semesters with the reading quizzes (controlling for other student variables such as non-biology GPA) saw a significant increase in exam performance (~5% increase; Pape-Lindstrom, Eddy & Freeman 2018). Though there were increases in performance, this study did not measure SRL variables directly, however, given the increase in performance and importance of 2YC environments in geoscience, this approach could be used by the GER community, yet could be improved as a tool to identify SRL behaviors if explicitly connected to SRL components (for a 2YC investigation within this work, see Chapter 5).

Other studies, however, did directly measure SRL strategy use in undergraduate biology settings. Sebesta and Speth (2017) developed their own survey instrument based on an existing interview protocol designed to evaluate student's SRL strategy use (Zimmerman and Martinez-Pons, 1986; Zimmerman, 1989) and evaluated the strategies used by students in their flipped biology course twice during the semester (after the first and second exams). They evaluated both the cumulative reported use of each specific strategy (e.g., seeking instructor assistance, reviewing notes), compared these to the students' grades, and calculated how these patterns changed between each survey. Results revealed significant differences based on exam performance, with both high performing students and those who improved from exam to exam reporting using more strategies than lower-performing students (Sebesta

& Speth, 2017). Additionally, lower-performing students reported a lower adherence to following through with the strategies they planned on implementing (Sebesta & Speth, 2017). These studies suggest that direct instructional interventions that are designed to foster student SRL strategy use have the potential to improve learning outcomes in undergraduate biology settings.

2.4 Discussion and recommendations

Goal #1: Providing a Theoretical Framework for SRL to Situate Future Investigations: SRL models provide a way to explain observed differences in student performance, and therefore learning (by proxy). If we seek to improve student learning and performance in geoscience, we, as a community, need to explore these SRL conceptual frameworks in geoscience contexts. While Lukes (2012) and Lukes and McConnell (2014), have reported some mapping of such SRL models in a geoscience discipline context, in general, there is little data about whether these models work in college-level science discipline contexts. Mapping out the nexus between conceptual SRL frameworks and actual learning experiences is a key part to advancing our understanding of SRL broadly, as well as student learning and success in geoscience specifically. It is important to explore these SRL frameworks in practice with real students situated in formal and informal learning experiences (e.g., courses, lab/field/research projects, museum exhibits, science center programs, clubs, citizen science projects) rather than experimental settings because understanding the realities of student experience is what provides the link to choices in pedagogical design and instructor practice. In other words, translating SRL research into geoscience teaching and learning practice. As discussed above, targeted interventions have been shown to increase both the frequency and efficacy of SRL-related strategy use in

students in other college science contexts (e.g., Dörrenbächer & Perels, 2016; Sebesta & Speth, 2017), but these need to be tested and refined for geoscience-specific contexts. Tools and pedagogical approaches can therefore be developed and refined to better support students for success in geoscience. Beyond simply improved performance, such success has implications for recruitment and retention.

Aside from promoting effective SRL behaviors in the geoscience classroom, it is important for researchers interested in this space to consider the specific features of the geosciences and how research into SRL can help augment geoscience instruction. Specific attributes of geoscience education (e.g., field work, spatial representations, abstraction) pose unique challenges to effective learning. Targeted research studies investigating: a) how students approach common geoscientific tasks via specific SRL strategies and; b) what kind of interventions can help foster effective SRL strategy-use to enhance learning are presently sparse (Lukes, 2014) and warrant future work.

One particularly salient opportunity for geoscience education researchers is in the developing realm of SSRL. Given the importance placed upon effective group work in both educational and professional geoscience settings, geoscience education researchers are particularly well-positioned to further the understanding of socially-constructed regulatory behaviors not only within our discipline, but in educational psychology more broadly. What kind of SSRL strategies are employed during effective field work? ...collaborative mapping activities during field camp? ...in active learning environments? Though some preliminary work has been conducted (Spencer, Thomson, & Jones, 2018), these are questions that can be investigated to help remedy this gap in the literature.

Goal #2: Methods of Studying and Supporting Effective SRL Practices in Learners: We need to understand what SRL strategies learners are currently using in geoscience learning experiences so that we can ultimately focus on developing targeted support for effective SRL strategy use. We see from the educational psychology community and other STEM DBER communities that a variety of self-report tools (survey instruments, reflection prompts/diaries, interviews, think aloud protocols) exist for use in documenting and measuring the SRL practices of learners. Less-explored are tools for artifact analysis, SRL-specific pedagogical observation, and participant behavior observation technologies (trace data collection and analysis from learning management systems, learning analytics from adaptive tutoring tools, eye tracking, digital badges and biometric monitors). With the prevalence of online learning experiences and emerging technologies related to virtual reality, augmented reality, and mixed reality, these tools are increasingly important to incorporate into our understanding of SRL.

In terms of broader methodological approaches, SRL research is largely emergent in the GER community and currently based on more qualitative methods. Referencing the Strength of Evidence Pyramid outlined in St. John and McNeal (2017) from an educational psychology lens, the research conducted on SRL spans all levels and is robust in volume. Narrowing focus to college level geoscience applications, this abundance becomes a paucity. The majority of extant SRL research conducted in geoscience contexts are either cases of practitioner wisdom or preliminary data presentations (e.g., conference presentations; Wirth & Perkins, 2011; Perkins & Wirth, 2011; Marton, 2015; Nunez, Lukes, & Rushing, 2015) or in some isolated cases published case studies (e.g., Spencer, Thomson, & Jones, 2018). The only example of a cohort study was that of Lukes & McConnell (2014), though the analysis

therein was primarily associated with sources of motivation and goal orientations (i.e., SRL was only considered as a potential influence on results). Future work should seek to move the community-wide understanding of SRL in geoscience up the strength of evidence pyramid via increased case and cohort studies investigating SRL variables.

Goal #3: Summary of Future Directions for SRL Research in Geoscience: The future directions for SRL research in geoscience contexts can be summarized in three broad areas: lines of inquiry, corresponding methods, and strategies for translating research into practice. In terms of lines of inquiry, several fundamental framework questions emerge from this review.

First, *do SRL models explain differences in learning and performance in geoscience learning contexts?* There are several priority research questions for the GER community. Namely, how are students regulating/not regulating their learning in geoscience contexts both individually and in group settings? What micro-level processes do they use? Do processes vary with specific types of learning activities or attributes of geoscience (e.g., field work, spatial abilities, abstraction)? What influences their underlying beliefs about how learning works and their corresponding SRL choices in geoscience contexts?

Second, *what methods can we employ to best document SRL in geoscience learning contexts?* There has already been initial work in GER around the use of self-report SRL data in the form of exam wrappers (e.g., Marton, 2015), strategy-use surveys (Nunez, Lukes, & Rushing, 2015), and student interviews that include think aloud protocols (Lukes, 2014; Lukes and McConnell, 2014). However, the relationship between self-reported SRL practice and actual practice remain unknown in these studies, which is consistent with the criticism of SRL studies in general that self-report SRL data is less trustworthy data in capturing

students' SRL sub-processes and applications (Rovers et al., 2019). There is a clear need in SRL studies and therefore the GER community engaging in such studies to attempt to triangulate self-report data with trace evidence (e.g., artifact analysis, learning management/learning analytics metadata, direct observation). Future studies in GER should therefore consider such limitations and explore the use of additional methods to validate findings through triangulation. The GER community is encouraged to test existing methods and tools for documenting SRL from other fields such as educational psychology, as well as develop and validate new ones. Think-aloud protocols or written reflection artifacts in which participants describe their thought process while engaging in a task or when reviewing video documentation of doing a task allow researchers to identify mismatches in participant belief and SRL action (Ericsson, 2006). Online learning activities provide trace data (e.g., resource access, learning tool use frequency, etc.) that could also be used to triangulate self-report data from participants (e.g., Jones & McConnell, 2019). Other sources of data could include artifact analysis, learning analytics metadata, and direct observation.

Finally, we may ask *how can these findings be translated into practice? What interventions can be employed to increase effective SRL skills and strategy/tactic use? And, can SRL support interventions be used to increase learning/performance?* Some preliminary work has been reported through conference presentations and workshop discussions, but there is a clear gap in the published peer-reviewed record for these questions. Providing peer reviewed evidence of intervention value is critical to the successful translation of SRL research into practice to improve student learning in geoscience. Once (un)successful intervention studies are documented, then practitioners can use the reported findings to inform their learning support task selections and teaching practices, to ultimately benefit

students. Given the importance of metacognitive skills and knowledge in effective SRL processes, priority should be given to testing interventions that aim to increase learner SRL-related skills such as metacognitive awareness and knowledge; control and learning strategy/tactic use; learning judgment accuracy; and task analysis. Priority should also be given to interventions that seek to support student SRL in the learning tasks or situations common to geoscience contexts such as field and lab settings; visualization and/or spatial thinking activities; abstraction activities; and collaborative group learning situations. How can instructors best design and scaffold learning support tasks in these contexts to support SRL, SSRL, and Co-RL?

2.4 Conclusion

Though prior work has laid a solid theoretical foundation for the investigation of students' SRL behaviors in college geoscience courses, only the first tentative steps towards identifying best practices in supporting student SRL strategy use have been undertaken. Future work should seek to identify several important, step-wise, phenomena that influence student SRL behavior. For example, we should seek to determine the suite of metacognitive skills undergraduate students bring to an introductory geoscience course due to the importance of metacognition in driving effective SRL behaviors. Additional work could then solicit student rationale for the selection of various study strategies. Finally, we could investigate what kind of instructor-led interventions could potentially "level the playing field" by promoting students SRL abilities for not only general success in the course but for increased opportunity for equity in instruction. Though these pursuits could be approached via traditional (e.g., survey-based) data collection activities, with the rise and proliferation of new educational technology we encourage geoscience education researchers to "think outside

the survey" and seek additional data sources that capture a more-authentic record of behavior (e.g., online trace data, etc.). Given the potential benefits for both geoscience students and geoscience education research, we encourage geoscience educators and geoscience education researchers alike to support SRL interventions to potentially improve student learning.

CHAPTER 3: THE DESIGN AND IMPLEMENTATION OF A WEB-BASED TOOL TO SUPPORT SRL AND STUDENT LEARNING ACCURACY: "CLASS"

Prepared for submission to *Internet and Higher Education*

3.1 Introduction:

In this chapter we will first introduce readers to research on an important cognitive process that, when activated, has the potential to help students adopt more effective learning strategies that can lead to enhanced performance. Next, we will examine how the traditional methods used to conduct this research has made it difficult to provide prompt student feedback on their learning accuracy. We will discuss how this led us to create a web-based tool to help students immediately reflect on their learning judgments so that they could readily identify potential misconceptions that might hinder their learning. We will introduce the tool, abbreviated as CLASS, describe its characteristics, and discuss the project we designed to investigate its utility and design elements. Finally, we will consider students' perceptions of the tool's utility collected via end-of-semester interviews in an introductory physical geology course.

3.2 Metacognition and Self-Regulated Learning:

Research reveals that metacognition, a learner's ability to recognize the workings and characteristics of their own knowledge and thought processes (Flavell, 1979), is an important influence on the level of success a student realizes from a learning task. Metacognition is generally separated into the two distinct components of *knowledge of cognition* and *regulation of cognition* (Schraw, 1998). Regulation of cognition is further subdivided into three processes: planning, monitoring, and evaluation (Jacobs & Paris, 1987). The concept of "self-regulation" could be viewed as the cycling between these three

processes during a learning task. Functionally, this cycle is the sum of a student's awareness and knowledge of their own thinking (metacognitive awareness), 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, and, consequently, determines how much a student will gain from a particular learning task. 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 address identified deficiencies (Zimmerman, 1990; Serra & Metcalf, 2009). Within an educational context, the process is referred to as self-regulated learning (SRL; Zimmerman, 1990).

Instructional interventions that focus on metacognitive support have the potential to foster student learning gains (e.g., Bannert & Reimann, 2012; Künsting, Kempf, & Wirth, 2013). A robust body of research has confirmed the beneficial effects of metacognitive support on learning in computer-based learning environments (Devolder, van Braak, & Tondeur, 2012; Zheng, 2016). Several studies have suggested that learners who are better able to identify and employ effective SRL strategies demonstrate higher academic achievement (e.g., Mega, Ronconi, & De Beni, 2014; Zimmerman & Martinez-Pons, 1988), exhibit higher levels of intrinsic motivation (Pintrich & Zusho, 2002) and demonstrate greater persistence (Pintrich & De Groot, 1990). Importantly, abilities related to SRL have been shown to be malleable and responsive to explicit and sustained training and feedback (Callender et al., 2016; Gutierrez & Schraw, 2015; Nietfeld & Schraw, 2002; Nietfeld et al., 2006; Schraw, 1998). During a geoscience course, a learner must process course content that

contains a considerable amount of specific and unfamiliar terminology (Kortz, Grenga, & Smay, 2018), thus making the ability to employ effective SRL practices an important factor for success.

Regulation of cognition relies upon accurate metacognitive knowledge for the SRL cycle to function properly (Serra & Metcalf, 2009). This information is provided via active metacognitive monitoring of a learning task as it is happening, which then influences important decisions elsewhere in the SRL cycle (Zimmerman, 1990). To make these decisions, students must make accurate monitoring judgments. For example, a student who is overconfident during these judgments of learning may choose to terminate their studying prematurely, leading to a poor performance on subsequent assessments (Dunlosky & Rawson, 2012). In contrast, underconfident students may invest unnecessary time and effort reviewing content that they already understand sufficiently well. Similarly, a learner's confidence during judgments was found to predict information-seeking in decision-making contexts (Desender, Boldt, & Yeung, 2018). Unreliable judgments lead to poor study decisions (see Bjork et al., 2013, for a review).

Student Confidence Judgments: *Measurement* - Traditionally, researchers have investigated student confidence judgments in courses by collecting confidence measures for individual exam questions using a "confidence line". The standard procedure for collecting and collating item-level confidence data is labor and time intensive, involving the researcher measuring the length between the origin of a continuous line and the student's intersecting confidence indicator (Figure 3.1) for each exam question for every student (e.g., Nietfeld, Cao & Osborne, 2006). The requirement to physically measure each of the lines associated with a judgment constrained these studies to cycles of data collection and subsequent data

processing before results could be obtained. Consequently, the measurement and subsequent analysis of the student line-measurements occurs long after the learning event and researchers would often not be able to isolate effects until after participants have completed the course.

Parameters: When a student asks themselves, "Am I prepared to take this exam?" they are engaging in 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 (Bol & Hacker, 2012). That is, they are poorly calibrated. Calibration is one measure of monitoring accuracy and represents the absolute value of the gap between a student's confidence and their measured performance on an assessment item (Schraw, 2009; Alexander, 2013). This gap is physically represented by the relative positions of a student's confidence indicator and the end of the confidence line (Figure 3.1). This measurement is 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. 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 3.1). 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|). In a relatively large class, an instructor might have to repeat this measurement process thousands of times just to collect the appropriate data for a single exam.

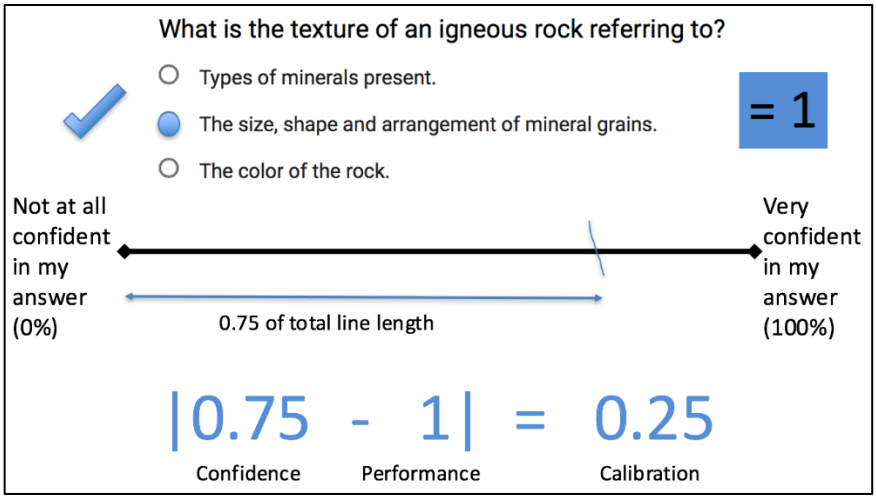

Figure 3.1 Example question and calibration calculation for local accuracy judgments.

For most students, this gap will not become evident until sometime after they have completed a summative assessment. Given the situational factors associated with higher education environments, this is too late to go back and address identified deficiencies 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), while low-performing students generally exhibit the widest calibration gap as they are more likely to be over-confident in their knowledge (Kruger & Dunning, 1999; Hadwin & Webster, 2013) and do not recognize that they are poorly calibrated, and therefore don't take steps to improve their self-regulation and close their calibration gap (Dinsmore & Parkinson, 2013). Additionally, whereas calibration represents the unsigned value of this gap between a student's performance and their confidence, a metric called 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 accuracy is a malleable skill but one that is only likely to improve when an intervention containing a variety of feedback, training or incentives is used to support learning and monitoring activities (Gutierrez & Schraw, 2015). Development of self-regulation and metacognition skills may be neglected if instructors assume that students already possess these skills or that such skills are too complex to teach in a typical course. The good news is that students can develop metacognitive skills over time when strategies are built into instruction (Bielaczyc et al., 1995; McCrindle & Christensen, 2005). Instructional interventions can increase metacognitive monitoring accuracy and improve learning (Huff & Nietfeld, 2009; Thiede et al., 2009) and improved metacognitive skills can potentially compensate for low ability and insufficient knowledge in a discipline (Schraw, 1998).

3.3 CLASS:

We developed the Confidence-based Learning Accuracy Support System (CLASS; classforlearning.com; Figure 3.2) to meet the goal of supporting students' metacognition and SRL behaviors. CLASS is a web-based tool that allows instructors of any discipline to measure students' metacognitive awareness via confidence judgements on content-based assessment questions. CLASS is grounded in the findings of educational psychology research and a self-regulated learning theoretical framework and has potential benefits for both students and instructors. For the student, CLASS has the appearance of an online quizzing program in which students also estimate their confidence in the accuracy of each of their answers. CLASS then determines their score on the quiz, communicates results in relation to

their confidence, and calculates calibration and bias scores - two empirically-derived measures of the gap between student confidence and performance (Schraw, 2009).

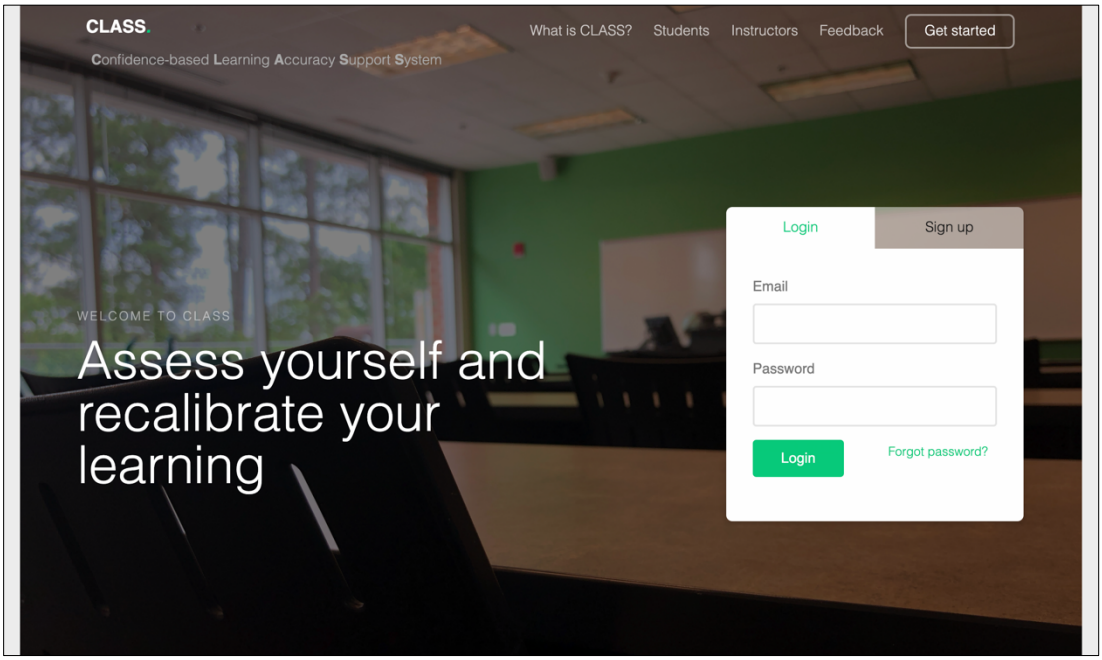

Figure 3.2 CLASS (classforlearning.com) landing page and login screen.

This communication makes what was once nebulous and approximate defined and quantitative for the student, giving them valuable and instantaneous feedback on not only their level of knowledge, but the accuracy of their perceptions. For the instructor, CLASS provides numerical and graphical summaries of students' performance and confidence relating to specific content areas of the course in which it is applied. CLASS can provide instructors with a wealth of information regarding course teaching and learning processes (e.g., what content students are overconfident or under-confident about) that cannot be readily ascertained from traditional assessment methods. Instructors can use this information to redesign lessons that target problematic content or to provide supplementary instruction or specific learning interventions.

To empirically investigate the effects of using the CLASS tool in a college course environment, we developed a project that iteratively paired conscientious course design and direct metacognitive instruction to investigate CLASS's potential to serve as a tool for increasing student metacognitive monitoring accuracy (i.e., closing the gap between confidence and performance), self-regulation and performance in undergraduate geoscience courses. The project applied the tool at both a two-year college and research-intensive institution in both face-to-face and online settings. Additionally, the implementation of CLASS will be analyzed as a novel metric for assessing relative effectiveness of instructional strategies in eliciting student learning gains and monitoring accuracy in the target courses.

We collected preliminary data in the geosciences that mirrored the correlation between student performance and calibration discussed above. During the Fall 2014 semester, item-level confidence judgments were collected via a continuous scale under each question of three paper-based midterm exams administered as part of an introductory physical geology course at a large research institution in the southeastern United States. Bivariate regression analysis found that students who averaged higher scores across all three exams were better-calibrated than low performing students and that 79% of the variability in calibration was explained by performance (R^2=.79; Figure 3.3). Additionally, as no feedback was given regarding these judgments throughout the semester, mixed-design ANOVA analysis of subsequent exams revealed that students did not become more accurate on later exams by virtue of repetition. Similar results 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).

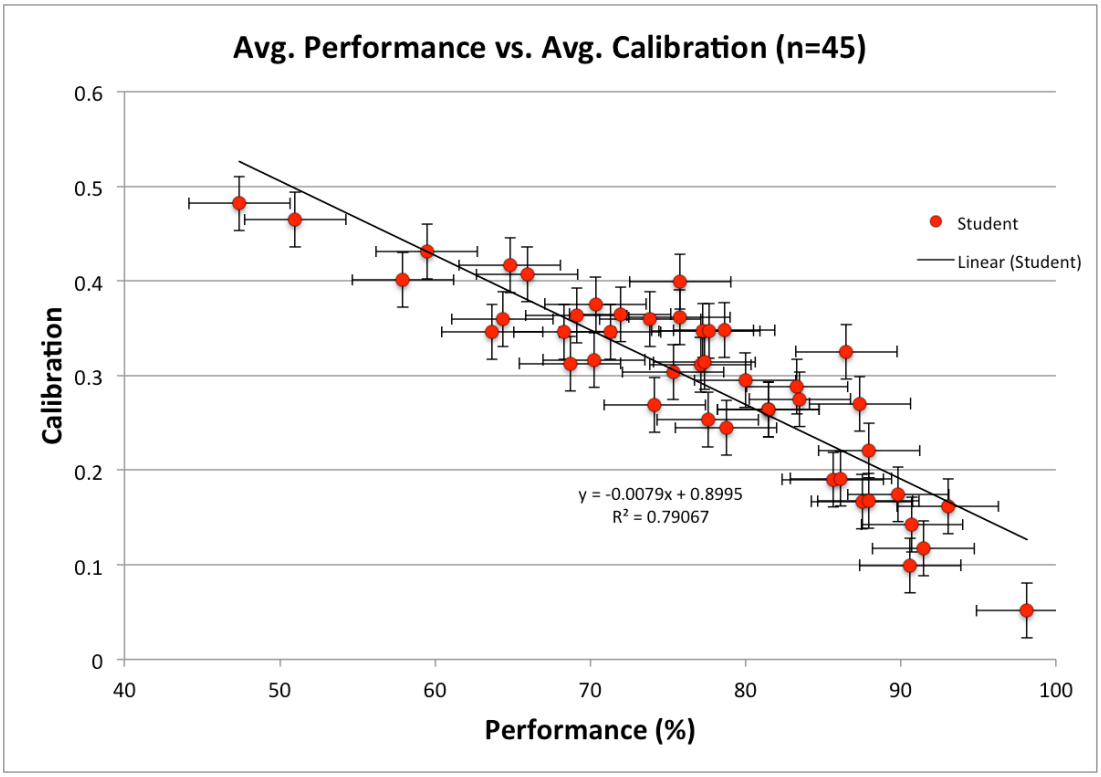

Figure 3.3 Mean performance vs. mean calibration for the baseline Fall 2014 semester.

In this work we describe a project where we paired iterative course changes with varying student use of the CLASS tool to provide more frequent student feedback of the accuracy of their calibration. Specifically, using the feedback the tool provides related to course learning objectives to providing students with a better sense of the success criteria for their lessons, which can be an important driver of subsequent study behavior (Alexander, 2013; Hattie, 2013).

Characteristics of "CLASS": As discussed previously, the normal procedure for collecting and collating item-level confidence data is labor and time intensive, involving the researcher measuring the length between the origin of a continuous line and the student's intersecting confidence indicator (Figure 3.1) for each exam question for every student. Calibration research has been hindered by this drawn-out process of data collection and subsequent analysis, delaying the opportunity to provide students with actionable feedback.

While collecting and analyzing student confidence data during another research project, we began to generate plans to create a web-based solution to this problem. After six months of research and development working in tandem with a web developer, we created the Confidence-based Learning Accuracy Support System (CLASS). CLASS automates the confidence data collection and calibration calculation processes as students complete an online quiz, allowing for immediate feedback to be provided to students following individual assessments and globally across all quizzes that the student completed for a course. Depending on how a quiz is designed, it may contain a set number of questions given in random or sequential order, or it may be a randomly-presented subset of questions from an instructor-generated question bank, thus making it unique for each attempt. Based in the concept of backwards design (Wiggins & McTighe, 2005), each question entered into CLASS must be connected to a specific learning objective designated by the instructor who generated the quiz. This allows for topic-wide consideration of results across one or multiple quizzes depending on the level of interconnectivity designed into each instructor's quiz suite. The utility of the CLASS web tool can be viewed from two perspectives: the student perspective and the instructor perspective.

Student use of CLASS: 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 global confidence judgments (i.e., score predictions) 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 3.4). The student must make a selection before being allowed to proceed to the body of the quiz.

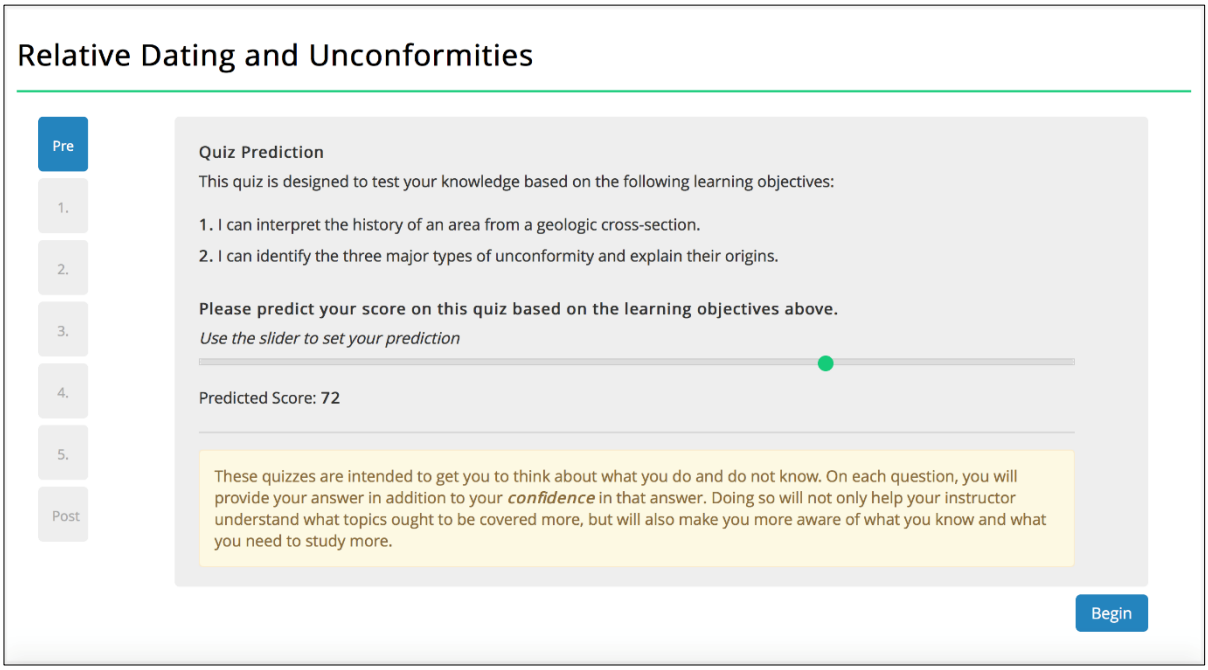

Figure 3.4 Example global task prediction within CLASS quiz attempts.

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 preliminary data above (Figure 3.5). Replicating the intuitive nature of the paper-based method of data collection, the continuous slider bar between zero and complete confidence is not preloaded with a value (as not to bias their response) and the location the student selects to place the

confidence indicator is not labelled with a numerical value (though for the researchers, this value is measured as a decimal percentage to the thousandths place). These features were designed to render the process of making an item-level judgment as intuitive as possible, in addition to eliminating the potential for the missing data that commonly plagues the process of collecting confidence-based data (i.e. the student ignoring the judgement line on paper exams).

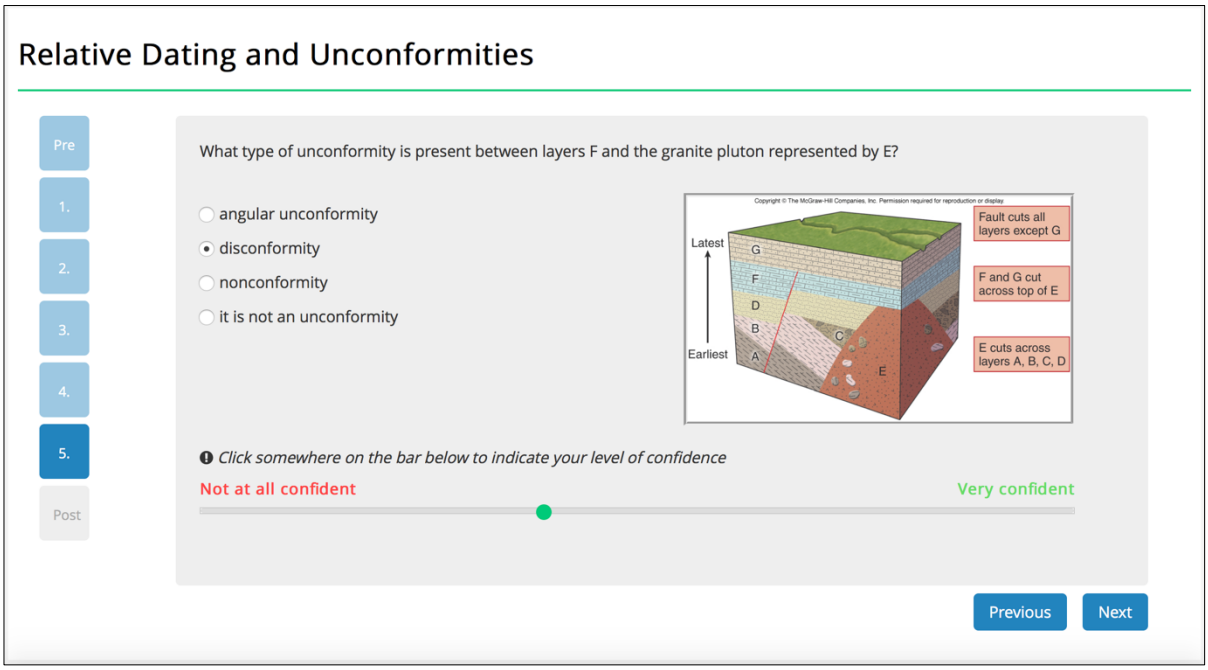

Figure 3.5 Sample question and item-level confidence measure within CLASS.

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 use the slider bar to communicate the score they think they earned (Figure 3.6). Generally understood to be more accurate (Bol & Hacker, 2012; Hacker et al., 2000; Pressley & Ghatala, 1990), this judgment, when compared to the global prediction and the average confidence derived from each item-level judgement, allows for important inferences to the student's experience during the quiz process. Did the quiz itself elicit an increase in

confidence or a decrease in score prediction? Were these values in-line with the average value derived from how confident they were in each question? Each of these questions, in addition to the student's calibration, references different facets of monitoring accuracy and are important lines of inquiry within the project.

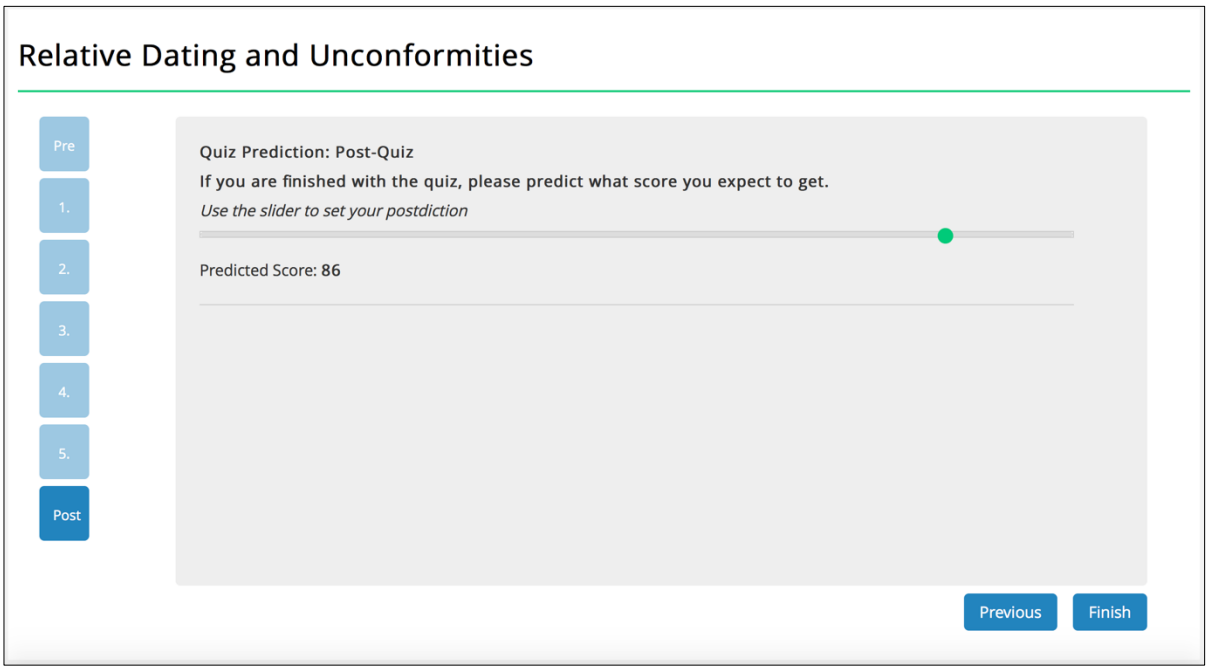

Figure 3.6 Global postdiction screen for a quiz within CLASS.

After the student has 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 3.7). Further, they are given feedback in relation to the accuracy of their metacognitive monitoring via auto-calculated calibration and bias scores (Figure 3.7). 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 sees their answer choices, the correct answer and the related calibration scores from each question.

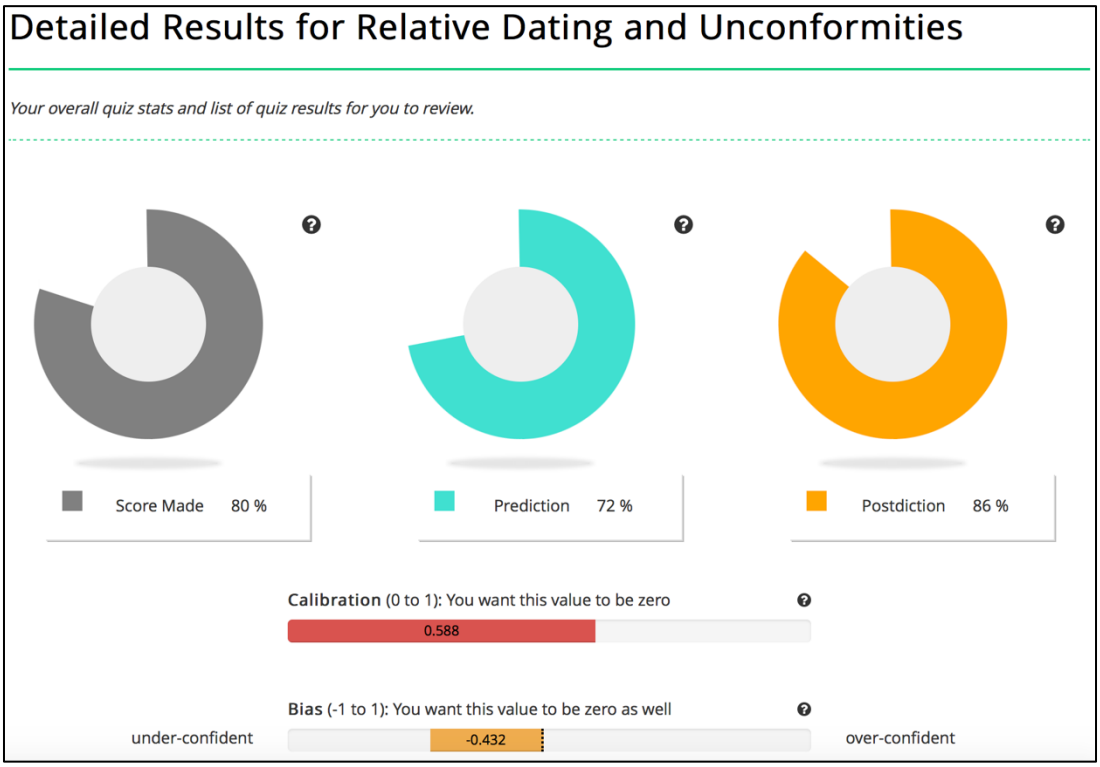

Figure 3.7 Sample results screen automatically presented to students after taking a quiz in CLASS.

All activity within the CLASS system is communicated to the student via a global results summary that provides mean values of all of the metrics measured within CLASS (Figure 3.8) in addition to a graphical representation of their attempt history (Figure 3.9).

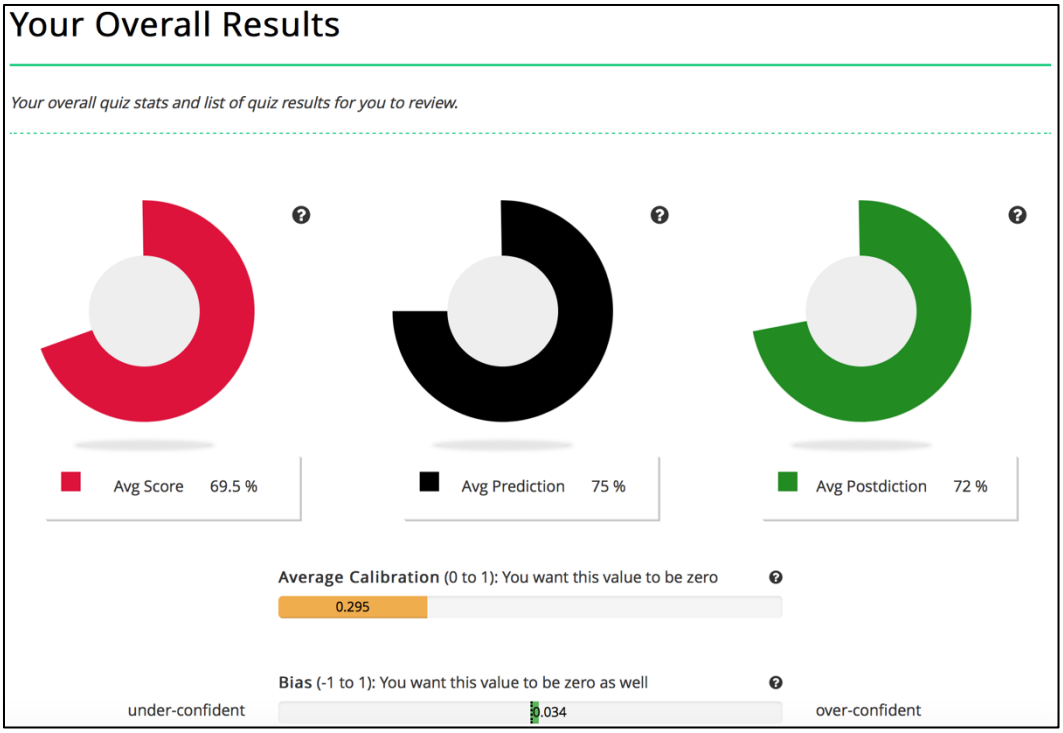

Figure 3.8 Global results communicated to students in CLASS.

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 (Figure 3.9). 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 3.9). These methods of providing direct and immediate feedback relating to the students' performance and monitoring accuracy will be used within the interventions associated with this project to leverage student awareness of their learning process.

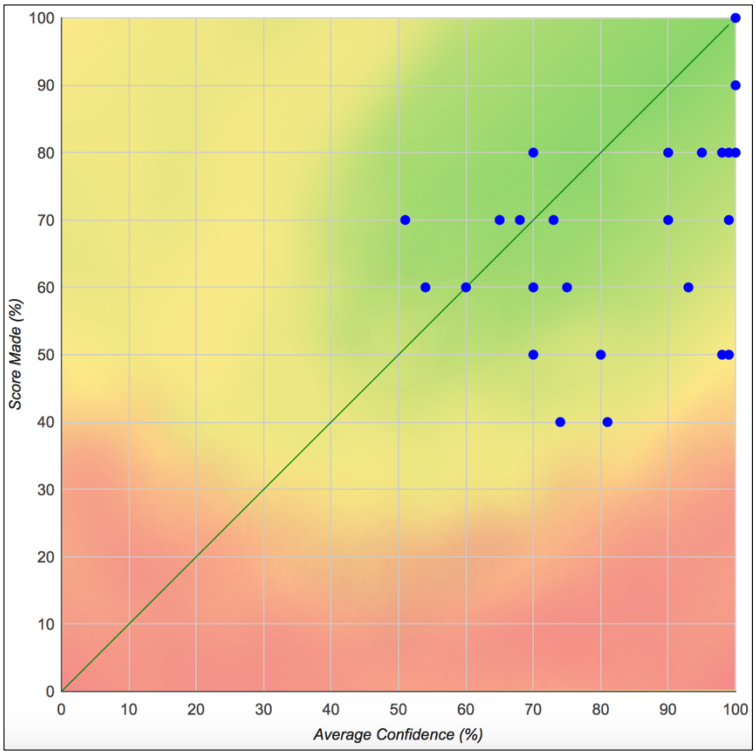

Figure 3.9 Graphical representation of student quiz results in CLASS.

In addition to being able to view results and associated accuracy metrics at the global and quiz level, students can view their results by learning objective. Importantly, this allows students to be able to specifically determine where their knowledge is inadequate or which topics are contributing to gaps in their perception accuracy. These are presented via a table summarizing the performance, confidence and bias (e.g., amount of under- or overconfidence) from all attempted questions delineated by their connected learning objective (Figure 3.10). In addition to the student's mean score, confidence and bias for the objective, the quality of the these values is indicated via a stop light style indicator tied to the bias value (green = < ±24.9% bias, yellow = ±25% - 49.9% and red = > ±50% bias). As with all tables in CLASS, the learning objective results table is able to be sorted by the ascending or descending values of any particular column. This allows students to isolate which objectives elicit the lowest values of performance, confidence, or positive or negative bias

(i.e., under or overconfidence). This information can be used as an external criterion to inform SRL and study decisions.

Attempted Learning Objectives

Learning Objective	Linked Quiz(s)	Mean Score (%)	Mean Confidence (%)	Mean Bias (%)	Full Results
I can identify the age of the Earth and cite evidence used to determine this age	- Divisions of Geologic Time	33.3	60.9	+27.6	View
I can name the four Eons geologic time and identify when they began and ended.	- Divisions of Geologic Time	100	56.2	-43.8	View
I can explain how life on Earth changed during the last 542 million years as older species became extinct and new ones evolved.	- Divisions of Geologic Time - Test ME!	27.3	71.8	+44.6	View
I can identify the four major units used to subdivide geologic time and order them by relative size.	- Divisions of Geologic Time	50	67.5	+17.5	View
I can name the three Eras of the Phanerozoic eon and identify when they began and ended.	- Divisions of Geologic Time - Test ME!	33.3	65.2	+31.9	View
I can identify the three major types of unconformity and explain their origins.	- Relative Dating and Unconformities	60	75.1	+15.1	View
I can describe the physical features and geological processes at a divergent plate	- Plate Boundaries	60	66.7	+6.7	View

Figure 3.10 Results by learning objective table presented to students within CLASS.

Associated with these objective-level results, beginning in the Fall 2019 semester, a new feature was made available on CLASS: dynamic quizzing. Whereas prior usage of CLASS was completely instructor-generated (i.e., instructors selected objectives and added questions related to quiz banks), dynamic quizzing allows students to select which objectives on which they want to be assessed from a list of all learning objectives for which they have attempted (Figure 3.11). This allows students to take control of their assessment and generate quizzes related to topics they either feel unconfident about or (thanks to the feedback CLASS provides) the topics they have demonstrated the lowest performance, lowest confidence, or highest levels of bias. Once they attempt their generated quiz, CLASS randomly recalls questions related to their selected learning objectives to generate unique quizzes for each attempt.

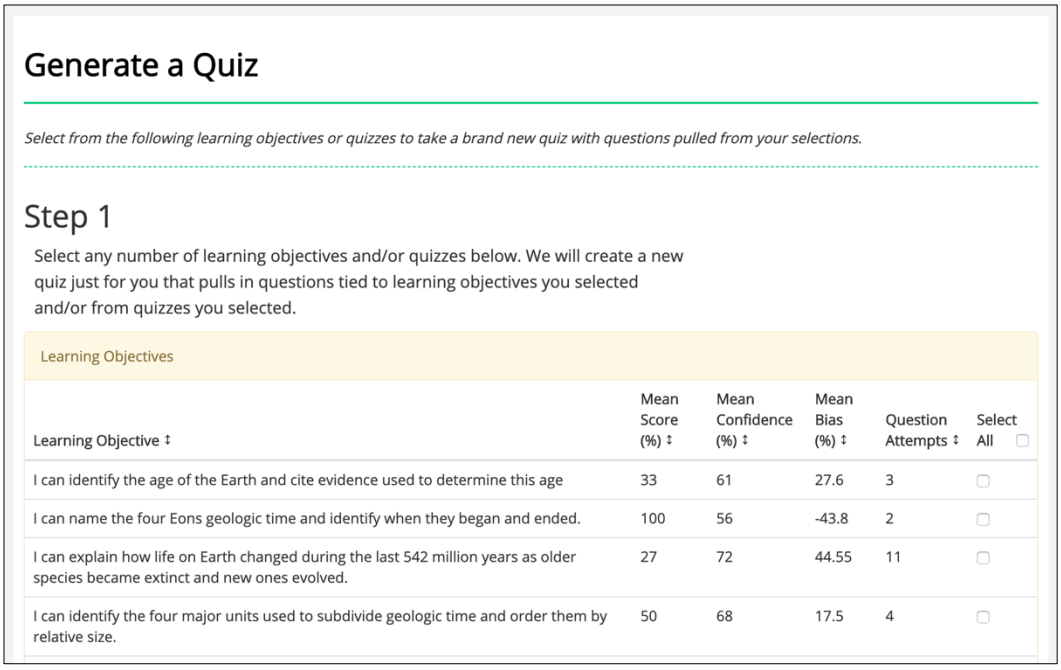

Figure 3.11 Dynamic quiz generation form for student use CLASS.

Instructor use of CLASS: For instructors, CLASS provides an opportunity to learn significantly more information than is possible from traditional assessment methods. Whereas a traditional assessment, once graded, will only provide performance-based data, CLASS provides instructors with student confidence data regarding the content they are being tasked with learning. This is communicated to instructors in three distinct modes. First, the instructor can review aggregate results for each of their quizzes. Similar to how global results are communicated to students, instructors are provided with mean performance, and mean global task prediction and postdiction values in addition to mean calibration and bias scores across all attempts of the quiz. This allows instructors many novel interpretations, such as whether or not taking the quiz produced gains in confidence from prediction to postdiction or if students were generally uncalibrated on the quiz content. The mean bias score also signals the directionality of any lack of accuracy (over- vs. under-confident). Second, instructors are also provided with confidence vs performance X-Y scatter plot for all

quiz attempts (same as Figure 3.9 but for all students who have taken the quiz). At a glance, instructors can determine how their students are collectively responding to the quiz material. In addition to considering data by quiz, CLASS communicates results to instructors (in the same manner as described above; Figure 3.8) in reference to global student results, allowing them to view each student's results across all quizzes they have created and/or shared with the student. Third, instructors can consider results across all of their quizzes in reference to each learning objective. This results screen displays both mean performance and mean confidence across all questions that reference that objective (Figure 3.12). In addition to these global statistics, instructors are provided a bar chart comparing the mean confidence and mean performance for each question associated with that objective (Figure 3.13). This allows instructors to isolate not only *topics* (i.e., learning objectives) that are leading to students being inaccurate in their monitoring accuracy, but any individual questions that may be causing particular problems. This information then presents an opportunity for the instructor to intervene via the augmentation or alteration of instruction or a check on students' study habits.

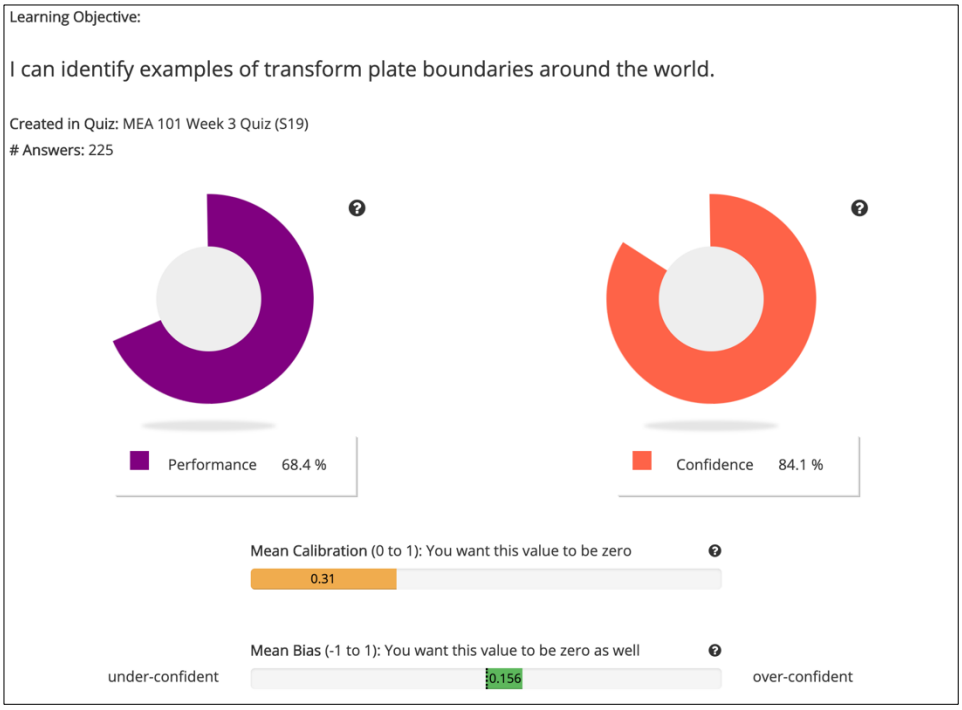

Figure 3.12 Results by learning objective summary statistics presented to instructors within CLASS.

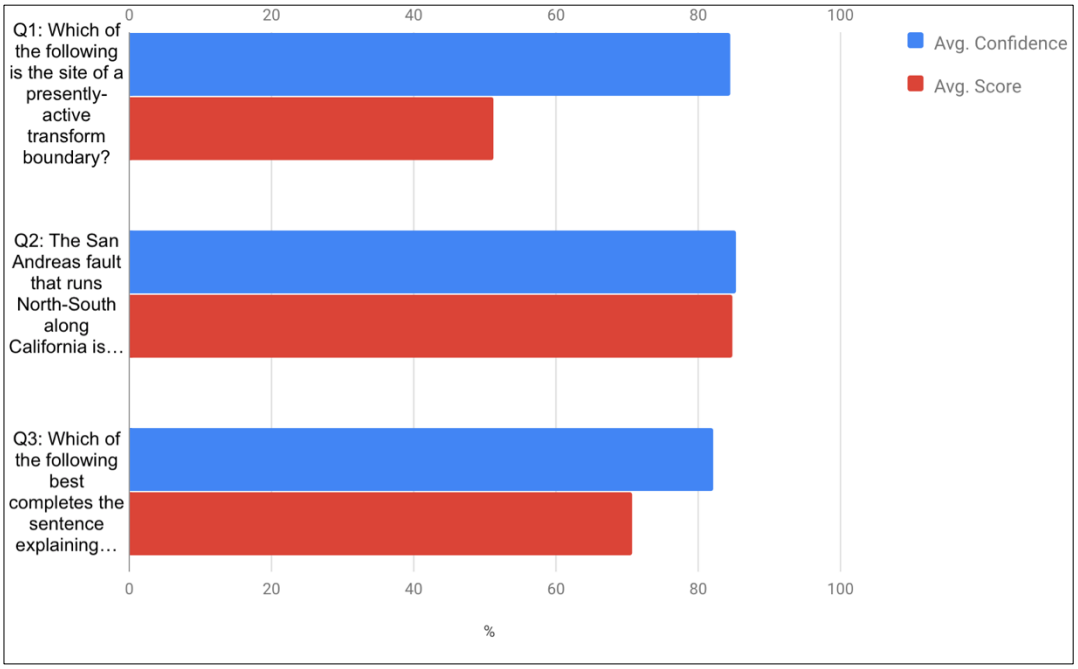

Figure 3.13 Mean performance and confidence for individual learning objective questions presented to instructors within CLASS.

CLASS and metacognitive monitoring support: From a theoretical perspective, CLASS provides students with instant feedback pertaining not only to performance, but to the level of accuracy of their judgments of learning (JOLs) and feelings of knowing. This feedback, if considered, could be used by the student to help close gaps in their understanding of course concepts and improve their overall performance as similar training programs have shown metacognitive monitoring ability to benefit from such approaches in the past (Nietfeld, Cao, & Schraw, 2006). Whereas feedback related to a specific question is generally useful and overall trends related to general accuracy may be of interest to the student (e.g., "I am generally overconfident in my knowledge of geology."), CLASS has the ability to tell students directly which topics are leading to their metacognitive dissonance via aggregated results by learning objective (e.g., "I can explain the rock cycle."). This allows students to seek to close gaps in their understanding and help them make more accurate judgments in light of targeted information.

CLASS and SRL: With conscientious student use, CLASS has the ability to provide students information that can help direct their study habits. A student's results on a particular quiz can inform them of which broad areas need the most attention during study and others that require less focus. At a finer level, learning objective results provide students specific topics or skills that can be addressed during preparation for course exams. For high performers who may feel like they need to study everything or low performers who may contain low motivation to study or not know where to begin, CLASS can help students determine which topics to study thus making the exam preparation process much more efficient for both populations. From an SRL lens, providing students a hyper-detailed external control of their monitoring and regulation processes can help students know how

and when to focus study and provide information on when to cease study, a decision often informed by questionable cues (Dunlosky & Tauber, 2014).

The introduction of dynamic quizzes in Fall 2019 added an additional layer to this control. Whereas the strategies a student utilizes while studying for an exam are often less effective (e.g., rereading of all notes or textbook; Dunlosky et al., 2013), the features of CLASS itself (e.g., practice testing) and the ability to directly practice test topics where the student has demonstrated the weakest understanding or highest metacognitive dissonance has the ability to direct student study habits in a more productive direction. We seek to leverage these features to investigate their effects on students' introductory geology learning within multiple higher education settings.

3.4 Project description

To test the effects of utilizing the CLASS tool in introductory geoscience courses, we integrated CLASS quizzing and a variable mix of CLASS-related activities and incentives into multiple courses at two different institutions over the course of three years. The goal was to test the effectiveness of CLASS in improving learning outcomes and general study skills. This process took the form of multiple intervention semesters during which data was be collected from populations of introductory geoscience learners at a research university and a regional two-year college (Table 3.1).

Research 1 Institution (R1)*:* We collected data from student-participants enrolled in introductory geoscience courses offered in a large research institution in the southeastern United States. The institution has an enrollment of more than 34,000 students and offers bachelor's degrees in geology, marine science and meteorology as well as master's and doctoral degrees. The undergraduate student body is 45% women and 55% men with 16% of

students identifying as a member of an underrepresented minority (URM) group. Students at this institution were sampled in a large (n= ~90 students) face-to-face introductory physical geology course taught by the same instructor every fall semester. This population consists of largely freshman and sophomore students (e.g., 76% freshmen/sophomores in Fall 2016). Additionally, the majority of these students are traditionally non-STEM majors (e.g., 72% non-STEM in Fall 2016). As a result, this intervention has the potential to provide the participating students with valuable metacognitive skills that they can carry with them through their college career.

Regional Community College (RCC): Students attending two-year colleges represent a large and important subset of the college student population. Students attending two-year colleges represent 42% of all undergraduates and 25% of full-time undergraduate students (Ma & Baum, 2016). These students are more likely to be members of a URM group (44% of black and 56% of Hispanic vs. 29% from these groups combined in the public four-year schools), first-generation students (36% of dependent students in the public two-year sector vs. 24% in the public four-year sector), and a large portion of them do not complete their desired program defined at the outset of their college careers (66% of students who started at two-year institutions have no degree after six years; Ma & Baum, 2016). Given the selection effect of degree requirements and introductory courses, it is likely that this population will find their way into introductory science courses at the 2-year college level and that this exposure may be their only exposure to college-level scientific reasoning.

Additionally, these institutions often have less stringent entrance requirements than four-year colleges and universities, which can yield students who are less academically prepared (Ma & Baum, 2016). This project aims to research means to improve the success of

these students in introductory physical geology courses offered in a face-to-face format at a regional two-year college about two hours away (driving distance) from the R1 institution. This institution has an enrollment of more than 12,000 students and offers a variety of associate degrees, diplomas and certificates. The student body is 60% women with 43% of students identifying as a member of URM group. Students were sampled from several small (n = ~20 students) face-to-face introductory physical geology classes taught by the same pair of instructors co-teaching the course over multiple semesters (Fall 2017 - Fall 2019; Table 3.1).

3.5 Methods

Course requirements and procedures: In each target course included in the project, there was an initial period where we collected baseline confidence data without deployment of the CLASS tool via initial summative assessments (exams) that measured student confidence and performance. This allowed for the calculation of students' calibration tendencies prior to using CLASS. Next, we tracked exam data from several semesters to determine if there were gains in student performance and/or calibration as a result of CLASS-related course interventions intended to support student self-regulation processes.

Subsequently, students completed content quizzes using the tool and had the opportunity to review their performance and calibration data. The requirements surrounding this use varied across the study semesters and will be outlined in detail below. In a global sense, we provided students with a variable suite of activities and resources intended to promote learning accuracy and reduce the gap between performance and calibration scores. These interventions were iterative in nature, with results from each intervention providing support and suggestions to be implemented in future applications. These specific activities

varied depending on situational factors related to the course or instructor. All activities included multiple opportunities for students to review lesson learning objectives and to complete related quiz questions using the CLASS tool so that they could have an opportunity to improve their metacognitive skills and become more intentional learners.

Regardless of institutional setting, each course required the following characteristics prior to using CLASS:

- Defined and communicated learning objectives for each lesson (class meeting; ~100 per semester) that directly stated what students should know or be able to do after completing the lesson (e.g., "I can define the water table."). As CLASS is built upon the principals of backwards design, each target course must have explicit learning objectives to align with assessment activities.
- Formative (CLASS) and summative (exam) assessments must be matched with lesson learning objectives.
- Students were to be provided lines under each exam question to complete confidence measures to generate baseline accuracy metrics (e.g., calibration, bias; similar to those represented in Figure 3.1).

Once these criteria have been met and baseline confidence/accuracy data collected during a pilot semester, each target course first provided students the opportunity to complete optional (no grade) CLASS quizzes in association with each of the modules of course content. This was to gain insight into how students chose to utilize the resource without the influence of the instructor or the course's design (i.e., being required to do so). These semesters served as a baseline for investigating how students chose to utilize CLASS solely to assess their knowledge of course content or prepare for upcoming summative assessments (exams).

For the R1 institution, students were assessed via four course exams (three midterms and one final). From these, we will focus upon exam performance and related judgments from the three midterm exams. The Regional Community College, being tethered to a lab component as part of its course structure, had more exams throughout the semester (six course exams and one lab practicum). Similarly, we will focus upon data collected from first four exams. In this work, we seek to determine how students in each setting chose to use CLASS prior to each exam.

Finally, after at least one semester of allowing students to use CLASS optionally, the placement of CLASS within the course structure was altered at each institution. This was to either force distribution of practice testing by requiring students to take at least five attempts per assigned quiz (an intermediate change at the R1) or to focus student effort and intention during quiz attempts by allowing unlimited quiz attempts for students to earn the highest grade (both institutions). Quiz grades in this final iteration were counted as part of the overall course grade in both settings. See Table 3.1 for a complete record of alterations to CLASS usage and requirements throughout the project.

Quiz bank generation and augmentations: To enable CLASS use in the target courses, instructors participated in creating databanks of multiple choice, true/false, fill-in-the-blank questions that were used to generate quizzes in the CLASS tool directly related to the course's learning objectives. To ensure that the quiz questions and learning objectives had relevance for physical geology instructors beyond those involved in this project, we surveyed content covered in several popular physical geology textbooks and also included questions on appropriate topics listed in the Earth Literacy documents (Wysession et al., 2012). In the R1 setting, there were multiple rounds of question development, increasing the

number of associated quiz questions per learning objective after each iteration while keeping the relative number and breath of the learning objectives similar. Beginning first with a bank of 358 questions that were used in prior semesters and administered via the course's learning management system (LMS; Moodle), the first round of question development in Spring 2018 increased the number of questions by 57.5% to 564 questions. The final iteration of question development in Fall 2019 further and significantly increased the number of questions to 1301 questions (a 130.7% increase) to achieve a criterion of at least ten questions available per course learning objective.

These augmentations of the question bank from which the quizzes were generated allowed students to be able to experience more unique attempts before seeing repeat questions via the randomization of questions to generate each ten-question attempt. By having at least ten questions per learning objective, the utility of the objective level feedback was increased. First, the summary statistics related to students' performance, confidence, and bias related to the objective were based on more attempts at the objective. Additionally, the mean values generated to form objective-level results were calculated from a wider diversity of questions, thus providing a more-complete picture of a student's understanding. For example, prior to the question augmentations, some objective-level metrics generated by CLASS only represented students' results from a few questions. This may or may not represent the true level of a student's understanding of that topic depending on the features of the questions or the actual elements of the topic that are being assessed. This can lead to a potentially false conclusion of a lack of understanding on an objective for both the student and instructor alike. If non-desirable results (e.g., low performance, low confidence, high levels of positive or negative bias) are generated from multiple attempts of more than ten

questions, it is much more likely that the phenomena signaled by the metrics represent true gaps in understanding.

Table 3.1 – CLASS use, course requirements, and utilized question bank for each target semester by institution.

		Fall 2016	Fall 2017	Spring 2018*	Fall 2018	Spring 2019*	Fall 2019
R1	**CLASS Use?**	Yes	Yes	Yes	Yes	Yes	Yes
	Question bank size?	358 questions	358 questions	564 questions	564 questions	564 questions	1301 questions
	Course Requirement?	Optional	Optional	5 non-graded attempts	5 non-graded attempts	Highest grade	Highest grade
RCC	**CLASS Use?**	N/A	No; Baseline exam data	No; Baseline exam data	Yes	Yes	Yes
	Question bank?	N/A	N/A	N/A	668 questions	668 questions	668 questions
	Course Requirement?	N/A	N/A	N/A	Optional	Optional	Highest grade

*= RCC semesters only; R1 Spring semesters are not considered in this report

We seek to analyze CLASS quiz results to identify the greatest calibration gaps among different iterations of the courses and specifically adjust course learning activities and/or formative assessments and assess if these interventions reduce calibration gaps. Finally, to triangulate cumulative effect of the activities included in each intervention, we administered pre/post surveys on both prior content knowledge (e.g., Geoscience Concept Inventory; Libarkin and Anderson, 2005) to all participants. This measure was collected to be used as a control variable during comparative statistical analyses and to determine net gains pertaining to geoscience knowledge pre- to post-participation in the study.

Interviews: During the final weeks of both the pilot R1 semester (Fall 2016) and final R1 semester of the study (Fall 2019) semi-structured interviews were conducted with students. Students identified interest in being interviewed via their consent forms. All students who were interested in being interviewed were sent an email with the opportunity to self-select a timeslot during the week of final exams. These interviews lasted approximately 10-20 minutes depending on student responses. Each followed a standard interview protocol that was structured into five categories:

1. Questions related to frequency and timing of use (4 total questions); this category included questions that probed student's decision-making regarding when they chose to take the quizzes and their perceptions as to what influenced the timing of their practice testing behavior. For example, this category included such questions as "*In relation to course exams, when did you use CLASS to assess your knowledge of the topic/module content?"*

2. Questions regarding the student's awareness of their learning process (9 questions); this category sought to target student perceptions regarding how they perceive their learning process, particularly related to how they considered the metacognitive feedback provided by CLASS (e.g., calibration, bias scores). Additionally, questions asked students about how they viewed making confidence judgments. Finally, as we were interested in how students' interactions with CLASS affected their approach to learning in other settings, we asked students questions such as, *"Can you think of examples from other settings (since your experiences with CLASS) where you have considered the accuracy of your confidence in an answer while you were completing a task?"*

3. Questions regarding preferences (2 questions); this category of questions asked students to reflect on what they liked and didn't like about the online quizzing process using

CLASS. These questions asked students whether or not they preferred quizzes that included confidence judgments and what they perceived to be the advantages and disadvantages of using CLASS. Specifically, *"As a student, what would you say were the advantages and disadvantages of using CLASS?"*

4. Questions regarding potential improvements to CLASS or course design (2 questions); the first question of this category asked students to suggest potential improvements in the CLASS system that they would be interested in seeing incorporated in future development of the tool ("*Can you suggest potential ways we can improve CLASS as a system?").* The second asked students to provide any improvements in how we choose to employ CLASS in the target course ("*Can you suggest potential ways we can improve the use of CLASS in the target course?"*).

5. Other thoughts not covered in prior questions and summary (2 questions); The final category of questions were designed to gather any latent student perceptions not captured as a part of the prior protocol (*"Please describe any other thoughts related to your experiences with CLASS not previously covered in this interview."*) and to have students summarize their experience in a general sense (*"Please summarize your experiences with CLASS this semester."*). The latter was added during the Fall 2019 interview process to insulate against students simply stating "no" as an answer to the first question and to force students to generate a summary of their perceptions of CLASS to bring to the fore the most salient aspects of their experiences.

As CLASS was designed to support student learning, metacognitive monitoring accuracy, and SRL behaviors, we will focus on student responses to Categories 1-3 in the section that follows. Questions in these categories were most directly related to determining

whether these design elements corresponded to students perceiving CLASS as providing the proposed utility of supporting student learning in the setting.

Following a grounded theory approach to the analysis of the interviews (Oktay, 2012), the semi-structured nature of the questions provided students opportunities to free answer open-ended questions and for student-derived themes to emerge. Follow-up questioning, however, was strictly maintained to asking students to further elaborate if they responded with simple "yes" or "no" answers, this was in attempts to control the experience across multiple participants. The interviews were audio recorded and transcribed before recordings were destroyed. After transcription, each participant was assigned a generic participant number and their study-wide identifier was removed.

Interview participants: In total, 25 student-participants were interviewed (11 in Fall 2016 and 14 in Fall 2019). Demographically, they were 44% female and 56% male (6F/5M in Fall 2016; 5F/9M in Fall 2019). Students in Fall 2016 were on average relatively high performers (Fall 2016 interviewee exam M = 89.1 SD= 5.6 Min = 81.0 Max = 98.8), while there was a greater diversity of performance represented in the Fall 2019 interviews (Fall 2019 interviewee exam M = 81.8 SD = 13.0 Min = 52.3 Max = 93.3). The majority of interviewees were non-STEM majors (57.7%) and were predominantly underclassmen (57.8%), which represented the larger demographics of the course as a whole.

Below we describe some global trends and patterns of perceptions of CLASS gleaned from these student interviews to provide, from a student's perspective, the efficacy of CLASS's design goals and potential to aid student learning. Detailed quantitative analysis pertaining to metacognitive monitoring accuracy and performance data collected from exams and on CLASS will be reported in later chapters. For example, results related to the iterative

course changes within the R1 institution and analyses comparing the different types of institutions (R1 vs RCC) are reported in future sections of this work (e.g., Chapter 4 and 5 of this volume, respectively). Finally, a more-thorough qualitative comparison of student responses from the pilot semester to final semester of the study will be reported in a later work. In essence, in this chapter we seek to demonstrate that students considered the CLASS tool to be useful and that they communicated evidence of SRL and metacognitive strategies as intended via the design of the tool.

3.6 Student views of CLASS

After the consideration of the interview responses from interview protocol categories 1-3 *en masse*, four common themes that emerged from this qualitative dataset:

Mechanics of CLASS as a practice testing tool: Students often mentioned that they thought CLASS worked well because of many of the features of how the tool was utilized in the target course. The first of these features was appreciation for the large number of questions tied to course learning objectives (a consequence of prior efforts to augment the question bank). For example, Student 6 attributed an advantage of the tool being the question bank, stating "**there are a lot of questions** and I really like that. Yeah, it was… **it was definitely hard to kind of see the same question twice, which definitely helped.**" Student 7 agreed by saying, "The fact that **there's so many questions that were like, able to choose from** and I could have seen that would have helped me like, study for tests better" Additionally, the randomization of questions per each quiz attempt along with the ability to take unlimited attempts for the highest grade was another oft-cited benefit of how CLASS was utilized in the course, with Student 14 touching upon both by stating "I think, being that **we could go back in and use it as many times as we wanted** too… that's unlike any other

class I've ever taken because usually when you do the homework is just one and done and that's it. But, that, that helps and especially that **it generates like specific questions [and] it's different every time**. That's, I think it was constructed very well to help us succeed."

Student perceptions regarding the depth of their learning: It was apparent from student responses that the various confidence measures and feedback provided within CLASS pushed students to think about what they knew and didn't know. Additionally, several students provided evidence that this information allowed them to make better decisions regarding what and when to study, generally sharpening their SRL skills as pertaining to learning in the target course. This sharpening demonstrated itself in two distinct forms, each differing by relative level of the student's performance. For higher performers, CLASS seemed to provide a means to hone their studying on the topics they truly needed to work on and to place less emphasis on topics they had already demonstrated mastery. For example, Student 11 reported "… **it allowed me to generate quizzes, with topics that I was less confident in**. So that I was able to focus my studying, **rather than studying things that I already knew, and wasn't going to help me during the exam**." The corollary effect for lower performers is how CLASS can provide them with a pathway for how to begin studying. Student 6 noted this effect directly by stating "…it [CLASS] was really, really useful. I mean prior to having CLASS [and] I could say prior to this course I did not study a whole lot at all. … I think that was due to just, I never really had the organization to find, you know, what I need to study and get an easy way to do that, and CLASS really provides that. I think that was the biggest, biggest impact for me."

While these comments suggested that CLASS provided benefits for students in the target course, some student comments suggested it made them think of the confidence in

their answers in other settings or courses. In addition to the example provided above, Student 10 stated, “Since using CLASS and other… in other classes now like when I take an assessment, I do. I give there's a question that I'm uncertain about **I’ll think like “how confident am I in this choice versus this choice?” and I'll pick the one I'm more confident in**.” These responses detail a common thread within the interview responses that suggest that many students are considering the depth of their learning in response to feedback provided within CLASS.

Matching learning with course goals: Another prevalent thread throughout the interview responses was reference to the linkage between the assessment questions in CLASS and the course learning objectives. This connection within CLASS made the quizzes valuable in allowing students to recognize if what they were learning matched what the instructor was teaching and whether their perception of this learning was supported by their performance on questions related to the learning objectives. Student 1 noted, “Yeah, I think it did [change views on learning process], because I like keep going back to the thing about like seeing what I don't know, but it is, it is good because some other courses [where] you can like look at practice tests and you go in there and do the practice test and see what you got wrong, but **this [CLASS] actually tells you which learning objectives you're getting wrong and not just [that] you got this specific question wrong.**”

Opportunity to drive their own learning: Finally, the addition of dynamic quizzing for the Fall 2019 semester was a common thread for the students who mentioned the feature. Throughout the interview responses, 50% of participants mentioned utilizing the dynamic quizzing feature, with the students who did often mentioning it several times during the course of the interview. These comments about the utility of the self-generated quizzes

indicated that CLASS gave (some) students ownership of their learning process and that this ownership led to a perception of CLASS as a valuable learning tool in the target course. In addition to the examples above, Student 13 spoke about dynamic quizzing during the main summary of their experiences by saying "I would always make quizzes for tests. I would make 20 question long quizzes combine each learning objective I needed, and it always helped me before tests, and [to] maybe not overthink at all before answering the question [on the exams]. It was certainly a great help it helped me more than any other resource available."

3.7 Conclusion

We set out to utilize CLASS to support student learning in introductory physical geology courses at multiple institutions over multiple iterations. Student interviews revealed that all students found utility in CLASS to help them learn the geology content and that many of the design elements performed as they were intended. Informed by suggestions from literature on metacognition and SRL, CLASS can help students find gaps in their understanding and to determine what to study to fill these gaps. Student 10's comments reflected this conclusion, stating:

"I thought it was very easy to use, and it helped me **improve my learning, learn how to learn and learn how to study**. It made it easier to learn topics because **it made obvious what I already knew and what I needed to work on more**."

Aside from these more-generalized considerations of design elements and student perspectives of use, two of the primary goals of this project as a whole serve as the focus for subsequent sections of this report: the first being investigating best practices in utilizing CLASS to improve performance in a large-enrollment introductory physical geology course

at a large research university (Chapter 4) and; the second being how students from different types of institutions used CLASS and how it affected their learning outcomes (Chapter 5).

Broadly, CLASS is a new tool that can be applied in any discipline and has potential to address a widely recognized problem in student learning. The use of CLASS makes it possible for instructors to integrate the results of (meta)cognition research directly into a class by applying a common assessment method that could be readily adapted by any instructor. This project provided an opportunity to test the tool in a two-year college and a research institution. This chapter, along with subsequent chapters that follow, begin to provide information on the optimal way to use the tool to ensure improvements in both student performance and self-regulated learning skill. In addition, we have produced a suite of related materials (e.g., customizable quizzes linked to specific learning objectives) that have the potential to be incorporated into a variety of introductory geology courses. Finally, the generalizability of both the tool and essential facets of student learning may allow for not only STEM-wide, but education-wide application of findings and recommendations for practices that increase student success in undergraduate courses.

CHAPTER 4: CLASS, COURSE SETTING AND COURSE STRUCTURE: INVESTIGATING BEST PRACTICES IN UTILIZING A WEB-BASED ASSESSMENT TOOL IN AN INTRODUCTORY GEOLOGY COURSE

4.1 Background

Metacognition represents a student's ability to recognize the workings and characteristics of their knowledge and thought processes (Flavell, 1979) and can be an important influence on student learning. Metacognition is generally separated into the two distinct components, *knowledge of cognition* and *regulation of cognition* (Schraw, 1998). Three processes (planning, monitoring, and evaluation) characterize the regulation of cognition (Jacobs & Paris, 1987). Learning that is facilitated by the effective cycling among these three processes is known as self-regulated learning (SRL; Winne & Hadwin, 1998). The regulation of cognition relies upon accurate metacognitive knowledge for the SRL cycle to function properly (Serra & Metcalf, 2009). This information is provided via on-line metacognitive monitoring of a learning task as it is happening, which then influences important decisions elsewhere in the cycle (Zimmerman, 1990). This ability to effectively apply the SRL cycle allows students to accurately identify the level and quality of their learning, the content that requires further study, which study strategies should be used, and when study has been completed (Serra & Metcalf, 2009). Students must make accurate monitoring judgments related to their learning to achieve a successful result. For example, a student who is overconfident during their judgments of learning may choose to end their studying prematurely resulting in a poor performance on subsequent assessments (Dunlosky & Rawson, 2012). Similarly, confidence levels during judgments were found to predict information-seeking in decision-making contexts (Desender, Boldt, & Yeung, 2018). Thus,

as long as the relevant judgment is reliable, learners' have a rational basis for making effective decisions regarding other SRL behaviors. Unreliable judgments lead to poor study decisions (see Bjork et al., 2013, for a review). To investigate the cognitive and metacognitive processes students utilize during learning, researchers have focused upon collecting and analyzing learners' judgments in a variety of settings.

Learners' feelings regarding their level of learning can be captured via self-reported judgments of metacognitive variables. These variables can be contextualized in a number of ways depending on the theoretical construct being investigated and the research goals of the investigator. For example, an investigation of monitoring accuracy has been approached via clinical study of subjects' recall of word pairs and associated judgments of learning (JOL; Nelson & Narens; 1990; Dunlosky & Tauber, 2014). In this procedure, subjects are presented a word pair (generally for a set period of time; e.g., 4 seconds) before the pair is removed and the subject makes a JOL in terms of the percentage likelihood that they believe they will be able to recall the pair in the future (Dunlosky & Metcalfe, 2009). These judgments were made *on-line* (during the learning process) prior to each attempted recall of a word and thus theoretically rely upon the subjects' metacognitive monitoring skills for their generation (Dunlosky & Tauber, 2014). Another type of judgment is that which occurs prior to a learning task is an ease-of-learning judgment (EOL; Nelson & Narens 1990). In this procedure, subjects make a predictive judgment related to how easy they believe it will be to learn about a topic to be studied in the future (Dunlosky & Tauber, 2014). In this judgment, the metacognitive monitoring is being applied to determine the ease of a task, rather than the relative success of the learning (Nelsen & Narens, 1990). Finally, the final judgment to be discussed (as there are many others) is the one that most directly applies to this work, the

confidence in retrieved answers judgment (confidence judgments; Schraw, 2009). These are judgments made in relation to assessment performance and can be measured in a number of "grain sizes" (Hacker, Bol, & Keener, 2008; Hartwig & Dunlosky, 2017). Individual judgments made after each question of an assessment indicating the subjects' level of confidence that their answer is correct are considered local judgments (Schraw, 2009). Judgments can be made regarding overall assessment performance (i.e., what score do you think you earned on this exam?), which are considered global judgments (Schraw, 2009).

All of these judgments can be manipulated temporally (e.g., predictions vs post-dictions of global exam performance) to investigate nuances in learners' metacognitive monitoring abilities in different situations or in response to different stimuli (e.g., Hacker et al., 2000; Händel & Fritzsche, 2016; Nietfeld, Cao, & Osborne, 2006). All of these measures have their own theoretical and practical connotations and can be utilized to answer specific research questions, but all must be evaluated against some sort of criterion to provide information regarding students' monitoring abilities. Most commonly, this is achieved through the evaluation of the level of accuracy between a judgment and an actual outcome (Schraw, 2009).

Comparing the judgment to the outcome can generate several metrics that can be evaluated within and across learners in different environments (see Schraw, 2009 for a review). Two metrics of particular interest to this work are calibration accuracy and bias. Calibration accuracy represents the absolute value of the gap between a student's confidence judgment (measured via a continuous linear scale) and their measured performance (i.e., dichotomously correct or incorrect) on an individual assessment item Alexander, 2013; Nietfeld et al., 2005; Schraw, 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).

In addition to this item-level, local accuracy measure that can be collected for each question of a task (e.g., exam, quiz), many studies on calibration accuracy also record 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 (after completing an assessment) the score they anticipate receiving for the assessment (Dinsmore & Parkinson, 2013; Pieschl, 2009). Global calibration accuracy measures allow students to capture a singular judgement of confidence that can then be compared to their eventual score on the assessment. 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).

Across studies that have utilized a range of metacognitive judgments and accuracy metrics, there are several pervasive results to note: a) learners are pervasively overconfident in almost all settings (e.g., predicting their upcoming exam performance; e.g., Foster et al. 2017; Hacker et al. 2000; Hartwig & Dunlosky 2017; Miller & Geraci 2011); b) 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); c) In contrast, low-

performing students generally exhibit the widest calibration gaps as they are more likely to be over-confident in their knowledge (i.e., are "unskilled and unaware"; Hacker et al., 2000; Hadwin & Webster, 2013; Kruger & Dunning, 1999). Often these low performing students do not recognize that they are poorly calibrated, and therefore do not take steps to self-regulate their study and close their calibration gap (Dinsmore & Parkinson, 2013). These negative effects of poor calibration are not isolated to low-performers. High performers who are uncalibrated may use study time unproductively to review material that they have already learned (Maki et al. 2005, 2009).

This disagreement between confidence in knowing and actual knowing is often due to an overweighting of particular factors that are not always diagnostic of memory and/or performance (e.g., ease of remembering) and a failure to account for a number of features of learning that do influence memory and/or performance (e.g., increased practice repetitions; Kornell & Bjork 2009; Koriat 1997). Several reviews (e.g., Hartwig & Dunlosky 2012; Kornell & Bjork 2007; Yan et al. 2014) outline learning scenarios that show high relative levels confidence to be an unreliable gauge of how much has been learned. We sought to examine how students could be supported towards developing metacognitive skills to inform more accurate SRL decisions. Opportunities to foster the development of these behaviors come through thoughtful course design and the incorporation of game elements into coursework.

Many have noted similarities between effective game play and effective learning as both require the planning and use of effective strategies to achieve success (Gee 2003; Kapp 2012; McNamara et al. 2004). A gamer looking to complete a level or unlock a new feature must consider the context, make a plan, monitor their progress as they enact the plan, and

evaluate results of their attempt before recycling that information into the next task in the game. These processes (as described) model the same procedures a learner must go through to effectively self-regulate their learning. Self-regulated learning (SRL; Winne & Hadwin, 1998; Zimmerman 2000) posits that effective learning occurs in this same cyclical process of planning, monitoring, and evaluation. As a result of these observed similarities, some have included aspects of gaming into learning environments in attempts to enhance student success (Landers et al., 2015).

Course design and game elements to promote learning: In both gaming and learning, success depends on the quality of the strategies being utilized towards meeting a set goal. In gaming, this may be getting to a new level or defeating an enemy, but in educational settings this goal is effective learning of content. The strategy a student chooses to utilize in order to learn required content directly influences the success of the learning and some strategies have been shown to be more effective than others (Dunlosky et al., 2013). Dunlosky et al. (2013) outlined the relative effectiveness of several common student learning strategies and suggested that two were the most successful for criterion learning: practice testing and distributed practice.

Unfortunately, students largely rely on the ineffective strategies (i.e., rereading, highlighting) during their study and tend to underutilize the higher quality strategies (i.e., practice testing, distributed practice; Dunlosky et al., 2013). Additionally, even if students do choose to utilize a high-quality strategy, they may enact it incorrectly. For example, practice testing has been shown to be an effective study strategy (Dunlosky et al., 2013) but while some students self-test during study, it's often executed ineffectively (Karpicke, Butler, & Roediger, 2009). True practice testing involves retrieval of content from the students'

memory, which strengthens connections between the content and the student's memory and increases recall (Dunlosky et al., 2013).

If we are aware of which strategies produce successful learning, yet students are relying on low-quality strategies during their study, the question is: how do we get students to use effective study strategies? One avenue towards correcting this discordance (but as yet understudied; see Dicheva et al., 2015) is through utilizing the popularity of games/gaming to induce effective strategy use. This has been attempted via two different approaches. The use of actual games (i.e., "serious games"; Deterding et al., 2011; Sailer et al., 2017) and the incorporation of game elements into educational settings (i.e., "gamification"; Deterding et al., 2011; Landers, 2014). While actual gameplay is not the focus of this work, applying game elements and/or incentives to the traditional course experience is an approach that has been attempted in the realm of metacognition and judgment accuracy (e.g., Callender et al., 2016).

Callender et al. (2016) collected global exam judgments from a situated upper-level decision making course in attempts to characterize and foster judgment accuracy. Instructors had students make global postdictions of exam immediately after taking a course exam. Researchers gamified the judgment by providing students with an opportunity to earn bonus exam points if they predicted their exam score within a certain range (e.g., "if you are 0-2 points away from your exam score you will receive 5 bonus points"; Callender et al., 2016). This resulted in more accurate judgments and students who were provided with related feedback increased their accuracy (Callender et al., 2016).

Given the promise of incentivizing accurate metacognitive judgments and eliciting effective study/SRL behaviors via course design, we iteratively made changes to the structure

and requirements related to an introductory physical geology course. Additionally, we incorporated new features into a researcher-developed mastery quizzing website designed to collect metacognitive judgments and provide feedback. This report describes these changes and their effects on students and makes recommendations for other instructors seeking to utilize similar techniques.

Baseline data collection, *CLASS* and initial investigations: 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 4.1) of three paper-based midterm exams administered as part of an introductory physical geology course at a large southeastern university. Simple bivariate linear 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 4.1), supporting results acquired in non-discipline specific literature in educational psychology (Kruger & Dunning, 1999; Bol & Hacker, 2000; Nietfeld et al., 2005). Additionally, no feedback or training was provided regarding student judgments throughout the semester and students did not become more accurate on later exams by virtue of repetition (Jones & McConnell, 2018), 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 in science education settings (e.g., Nuhfer & Knipp, 2003; Wirth & Perkins, 2005). This lack of improvement over time was hypothesized to be due a lack of feedback on the quality of

judgments and knowledge of the level of accuracy maintained, a factor that has been shown to improve judgments in other settings (e.g., Nietfeld, Cao & Osborne, 2006).

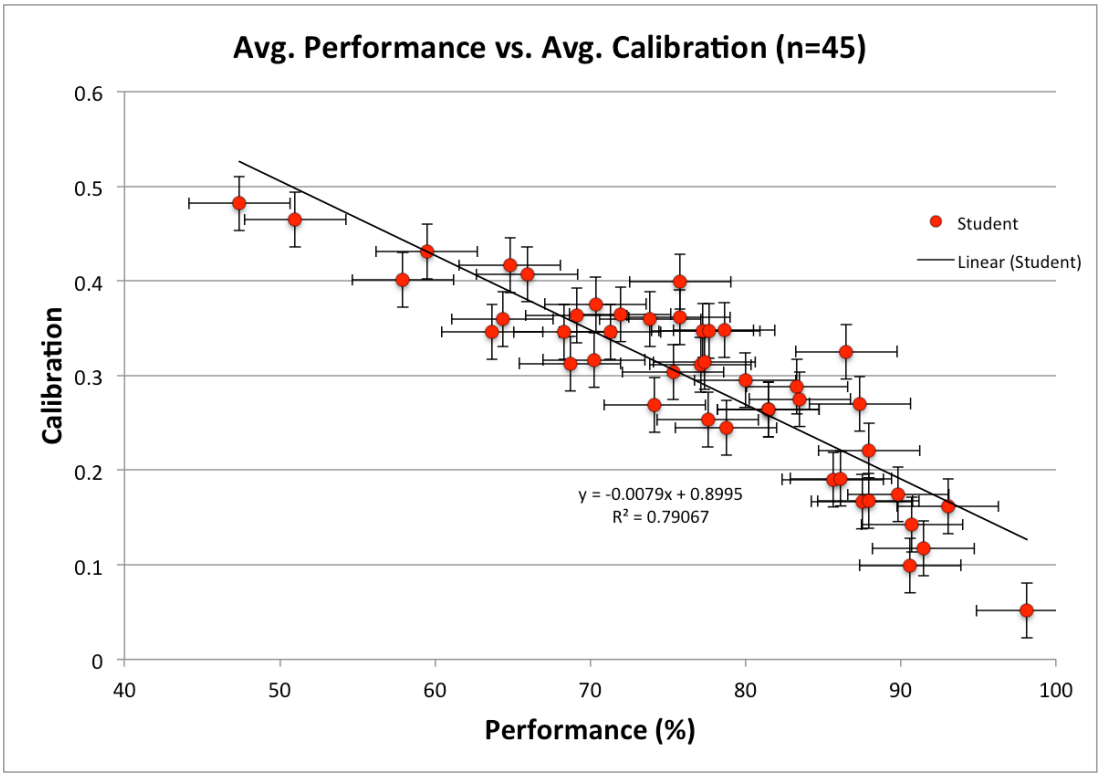

Figure 4.1 Mean performance vs. mean calibration for the baseline Fall 2014 semester.

CLASS (2016-2019): We generated a quizzing tool (CLASS, see Chapter 3) to provide students feedback to assess their knowledge and to allow information regarding students' metacognition to guide SRL behaviors. Traditionally, students in the target course were provided with an opportunity to take mastery quizzes aligned to the learning objectives for each unit of geology content material via the course management system (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

prior to exams in a no-stakes environment as the quiz attempts were not considered in the final course grade.

For the Fall 2016 semester, the Moodle quizzes were replaced with equivalent quizzes (i.e., the same bank of content questions) in a web-based assessment tool, the Confidence-based Learning Accuracy Support System (CLASS). In addition to providing students multiple choice questions to practice, CLASS automates confidence data collection and calibration calculation processes as students complete the online quiz. This allows for immediate and robust feedback related to monitoring accuracy to be provided to students following individual quizzes and globally across all quizzes that the student has completed for the course. Additionally, all questions within CLASS are tied to specific course learning objectives. Thus, CLASS provides explicit information regarding students' level of mastery of course content in an effort to combat unreliable judgments that may lead to poor study decisions (see Bjork et al., 2013, for a review). Consequently, CLASS provides data-driven feedback on what students need to study and what they know well (see Chapter 3 for full details regarding the design and development of CLASS).

The Fall 2016 class represented the first step toward characterizing student performance, confidence, and metacognitive monitoring accuracy related to their learning by introducing CLASS as an optional tool within an introductory geoscience course. Results indicated a significant increase in performance (mean difference = 4.49, $t(117) = 2.20$, $p = .03$, $d = .41$) and certain measures of accuracy (global accuracy non-significant from zero; $t(54)=1.19$; $p = 0.24$) and a general trend towards increased local accuracy overall yet were not universally conclusive (Figure 4.2 vs Figure 4.1; Jones & McConnell, 2017). The higher performance and accuracy demonstrated by the Fall 2016 cohort cannot be definitively

accounted for by CLASS as there had been no opportunity to compare pre/post measures with previous semesters to rule out that possibility that higher scores were the result of higher prior geoscientific knowledge and/or academic experience.

These results suggested both promise for future study and for the design and implementation of interventions seeking to further investigate the effects observed in the study to provide further support for student learning in introductory geoscience courses. Additionally, there were signals from further analysis of CLASS trace data and student interviews collected during the Fall 2016 pilot study indicating that many students were only using the tool in the hours leading up to an exam and that some were ignoring the feedback provided by CLASS (Jones & McConnell, 2017). For example, during an interview, one student stated, “I don’t really look at the calibration, no.” (Jones & McConnell, 2017). If students were not distributing their practice or considering feedback, what would happen if the requirements surrounding CLASS changed within the course? Could we increase student exposure to feedback and distribute their practice more broadly prior to exams?

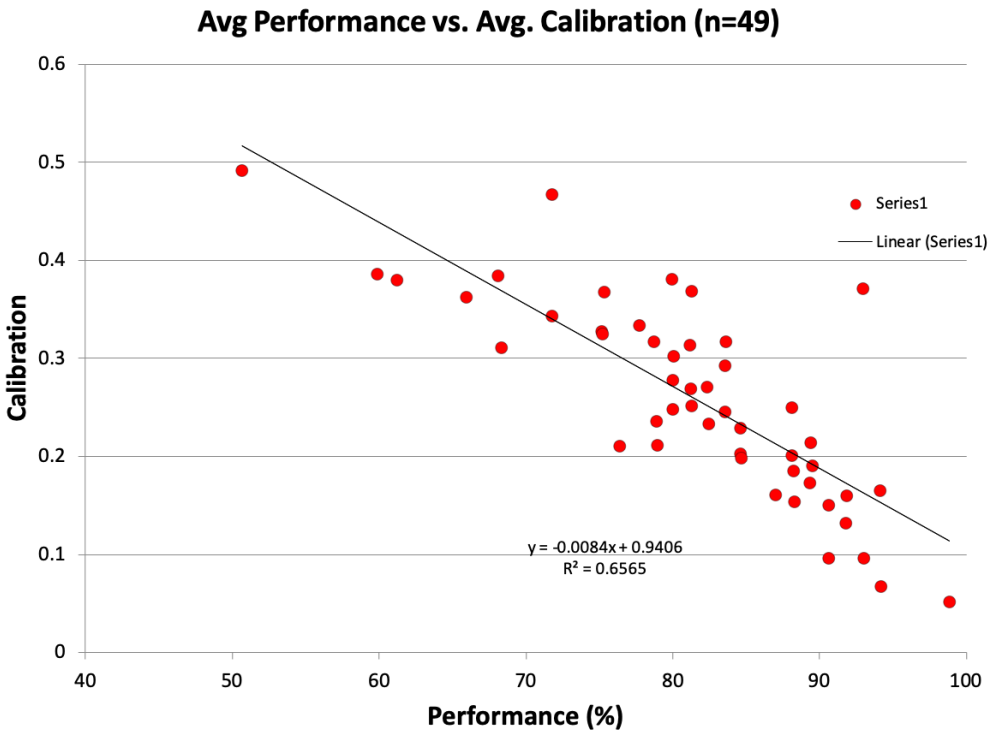

Figure 4.2 Mean performance vs. mean calibration for all exams for Fall 2016.

To answer these questions, this study sought to fulfill two broad goals: 1. To expand the consideration of how best to utilize the CLASS tool to elicit effective study practices (e.g., increased and distributed practice testing) in the target course by making iterative changes over multiple semesters and analyzing their effects on student outcomes; 2. To attempt to increase students' consideration of CLASS feedback and impact SRL behaviors by providing new features and functionality within the CLASS tool itself. To meet these goals, we derived the following research questions:

1. *What are the effects of course modifications related to CLASS quizzing over the study period on students' CLASS usage patterns in the target course? Specifically, …*
    a. *…the effects of altering course requirements related to CLASS quizzing (optional ungraded vs. required attempts ungraded vs. graded)*

b. *… the effects of altering course/quiz structure (more frequent quizzes; more questions per quiz)*

c. *…the effects of new CLASS features (e.g., dynamic quizzing)*

2. *What are the effects of these same course modifications on student exam outcomes (performance, confidence, calibration accuracy)?*
3. *What are the effects of employing gamified elements in Fall 2019 on global judgment accuracy when compared to prior semesters (Fall 2017 and 2018)?*

4.2 Methods:

A situated quasi-experimental mixed-methods study was designed and implemented across three treatment semesters (Fall 2017, 2018, 2019) to investigate these research questions and to further characterize the discipline-specific monitoring behaviors of learners in the setting. Each semester we sequentially altered an aspect of the course design and course materials to potentially elicit and isolate effects on student performance and monitoring accuracy variables. Efforts were made to control for many situational factors affecting each cohort in an attempt to mitigate confounding variables. 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. This report focuses on the quantitative data collected from students in course exams and during their interactions with the CLASS quizzing website. We will seek to triangulate to quantitative results via qualitative student perspectives collected during semi-structured interviews conducted at the end of the final (Fall 2019) semester.

Participants: Participant cohorts consisted of three convenience samples of undergraduate students enrolled in an introductory physical geology course and spread across

three experimental (2017-2019) 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 freshmen or sophomores (>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. This research was conducted in accordance with an IRB protocol declared exempt by the research institution. Participants were presented informed consent project on the first day of the course and could withdraw at any time. Signed consent forms were sealed until the final week of the semester and the instructor was not made privy of participants until after final grades were submitted. Non-consent attrition was low for each semester (76/87 in Fall 2017, 84/85 in Fall 2018, 73/78 for Fall 2019).

Course format: Throughout this study, the target course was taught utilizing an active learning format. Students had access to course resources such as copies of the lecture slides, lists of learning objectives, online practice quizzes and practice exams. Each class meeting (except for days with exams) was preceded by a homework assignment known as a learning journal featuring short videos on basic content. Students would then answer several content-related questions on the online course management system (Moodle) to assess their learning. At the beginning of the corresponding physical meeting of the class, students would typically review their learning by responding to a few (~2-4) related conceptual multiple-choice questions which they would answer via a classroom response system ("clickers"). Scores on the learning journal activities accounted for approximately 28.6% of the course grade during Fall 2017 and 25% in Fall 2018 and 2019. The remainder of the course grade was accounted for by exams (71.4% of the grade in Fall 2017; 62.5 in Fall 2018, 2019) and

quiz scores in Fall 2018 and 2019 (12.5% of grade). Course lectures included multiple formative assessment activities intended to provide an opportunity to both measure and reflect on student learning.

Student measures: *Content Pre/post Test* – During the first course meeting of each semester, students were given a fifteen-question pretest selected from the Geoscience Concept Inventory (GCI; Libarkin and Anderson, 2005). These fifteen multiple-choice questions were selected from a larger selection of questions that were validated by many geoscience education researchers over the course of multiple sequences of development (Libarkin and Anderson, 2005). The questions selected for use in this assessment were chosen due to their relevance to the concepts covered in the course. Considering the Fall 2019 pre-test dataset as an example, reliability analysis of the question set itself produced a Cronbach's Alpha value of .707; just exceeding the ≥ .70 indicating acceptable reliability as recommended by Ding and Beichner (2009). GCI pre-test scores were used to determine concept knowledge baselines for students and as a continuous control variable for the regression-based components of the study (Chapter 5).

Exam 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 true/false questions plus 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. Cronbach's Alpha values for each of the exams across the target semesters were 0.7 or greater, indicating reliability and relative consistency of performance across students on each item of each exam (Ding & Beichner, 2009).

Exam confidence and 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 4.3). 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 4.3) 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 (see Chapter 3; 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 4.3). 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).

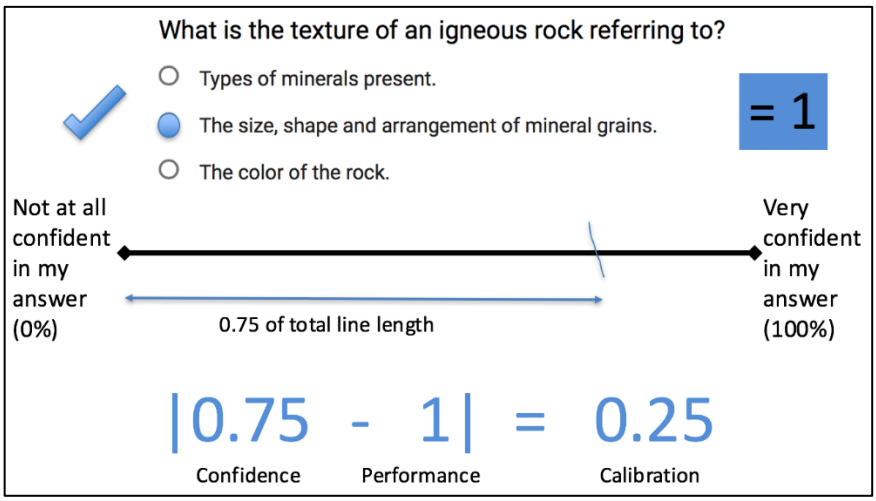

Figure 4.3 Example question and calibration calculation for local accuracy judgments.

Exam global monitoring accuracy – Throughout the study period, student confidence related to global performance was collected via a request to make a prediction of the score they predicted they would 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 (including short answer questions) 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 each experimental semester and averaged within each student to generate a mean value representing global accuracy across each semester. The verbiage and formatting of the judgment request was altered slightly in each semester to focus student attention on providing a judgment and minimizing attrition in this measure due to non-response. Finally, during the final semester (Fall 2019), this measure was the subject of gamification efforts described further below.

CLASS-derived data sources – In addition to serving as a study tool for students, student use of CLASS provides a number of "back-end" data sources that can be used to triangulate other data sources. Additionally, trace data provides opportunities to glean insight into students' formative assessment behaviors not otherwise collectable via traditional methods. For example, each submitted CLASS quiz attempt records students' responses, global and local judgments and related accuracy variables (calibration and bias) along with a timestamp (m/d/y 00:00) of when each question within the attempt was submitted. When considered in tandem, plots of these variables can be used to determine student usage patterns to see how and when students are utilizing the tool in relation to scheduled course activities (e.g., exams).

Course revisions: Several alterations were made to the structure of the course and new features were added to CLASS over the study period in attempts to foster improvements in student outcomes in the course. These changes were iteratively integrated into the course, with each being added in subsequent years the course was offered.

Changes in course structure and CLASS requirements: Results from the pilot semester provided several signals that students were a) predominantly taking quizzes in the 48-hours leading up to the exam and; b) not deeply considering the feedback provided by CLASS. As a result, we began to make iterative changes to how CLASS was utilized in the course to both distribute students' practice (an effective learning strategy; Dunlosky et al., 2013) and to motivate students to consider the learning feedback CLASS provided. As with the pilot semester using CLASS (Fall 2016), students use of CLASS was completely optional for the Fall 2017 semester. The course was structured in a "Module" format that grouped course content from 2-4 course meetings into thematic groups (e.g., Module 4 - Earthquakes and Volcanoes). Students were provided module quizzes that allowed them to formatively assess their knowledge via unlimited ten-question randomized attempts. Each attempt was generated from a quiz bank that contained approximately 50 questions per module quiz (358 questions in total across 8 module quizzes). Students were explicitly encouraged during the lectures to adopt a distributed practice approach to studying and to utilize CLASS to practice for upcoming exams and to use the feedback they received to guide their study. Each course exam then summatively assessed students on two modules of content (Figure 4.4).

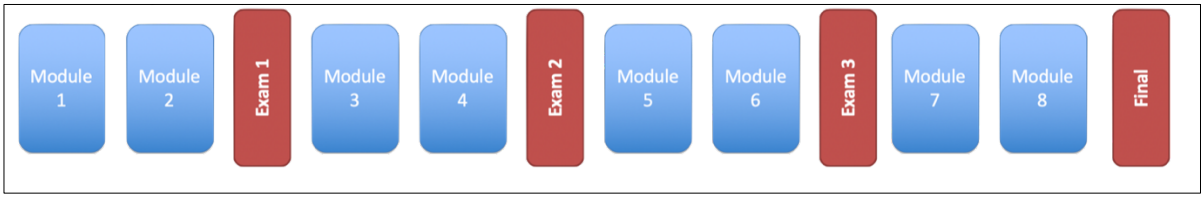

Figure 4.4 Course content grouping and exam structure for Fall 2017 semester.

Three alterations were made to this structure for the Fall 2018 semester. First, some of the modules were subdivided to focus consideration of content into smaller groups of lessons (called "topics") and to further distribute students' practice by having more frequent quizzes. For example, Module 4 (Earthquakes and Volcanoes) that comprised of

approximately five course meetings (~2 weeks) and 16 learning objectives was subdivided into Topic 6 – Earthquakes (~1 week and 8 learning objectives) and Topic 7 – Volcanoes (~1 week and 10 learning objectives). Functionally, this altered the student quizzing experience by focusing the content being assessed. For example, a student taking the former Module 4 quiz would have to expect randomized questions on both Earthquakes and Volcanoes. In contrast, a student taking the Topic 6 and Topic 7 quizzes would receive separate questions sets on earthquake- and volcano-related learning objectives, respectively. This alteration was designed to both focus effort as described, but to increase the frequency of which students assessed their learning by having more quizzes overall (Figure 4.5).

Figure 4.5 Course content grouping and exam structure for Fall 2018 and Fall 2019 semesters.

The second alteration in the Fall 2018 semester was that students were required to make five ungraded attempts for each topic quiz in CLASS. While the actual scores on the quizzes did not count toward the course grade (to remove any related grade anxiety), students received one point for completing each attempt (up to a maximum of five). These attempts accounted for small proportion of course points (up to ~1/16th total points in course). Additionally, students were provided opportunities to reflect on their CLASS use via weekly assignments that asked them to consider their results in-depth (e.g., what is the topic quiz where you demonstrate the lowest calibration?). Finally, the third alteration to how CLASS was utilized in the Fall 2018 semester was through an increase in the number of questions available to students via the subdivision of module to topic quizzes. As discussed in Chapter 3, the number of questions approximately doubled across the quizzes.

In Fall 2019, a third modality for CLASS use was implemented with two notable changes from the Fall 2018 usage. First, students were required to take topic quizzes but were allowed unlimited attempts to earn the highest grade. This grade was then counted as part of their overall course grade (up to ~1/8th total points in course). Second, we significantly augmented the question bank tied to each topic quiz to reach a criterion of a minimum of ten questions per learning objective. This was accompanied by a reduction in the number of learning objectives (by ~10%) to combine some that were considered overly specific. We significantly increased the number of questions (1301 across 11 topic quizzes), thus providing opportunities for students to both take more unique attempts, and to increase the utility of learning objective level feedback to help guide students SRL decision-making regarding studying for exams. Finally, some gamified elements were added to spur CLASS use and consideration of accuracy feedback in the Fall 2019 semester that are described in detail below. A full summary of course structure changes through the target semesters is available in Table 4.1.

New features in CLASS and dynamic quizzing: In attempts to better support students' learning and SRL behaviors in the course, several updates and new features were added to the CLASS tool over the study period to increase its functionality. These features were broadly designed so that students could more-easily isolate where gaps in their understanding or accuracy were present. The following features were first available for student use during the Fall 2018 semester. First, all tables displaying student results were made sortable by column values. For example, students could sort the table and view their quiz attempts by lowest confidence, lowest performance, highest positive bias (i.e., overconfidence) etc. and more easily identify problematic topic areas.

Table 4.1: Course structure, CLASS, and gamification features by semester.

	Feature:	Fall 2016	Fall 2017	Fall 2018	Fall 2019
Course structure changes	**CLASS Use?**	Yes	Yes	Yes	Yes
	Module or Topic?	Module	Module	Topic	Topic
	Course Requirement?	Optional	Optional	5 non-graded attempts	Highest grade
CLASS features	**Question bank size?**	358	358	536	1301
	Sortable tables?	No	No	Yes	Yes
	Results by LO?	No	No	Yes	Yes
	Results by attempt number?	No	No	Yes	Yes
	Clickable results on graphs?	No	No	Yes	Yes
	Dynamic quizzing suite?	No	No	No	Yes
Gamified elements	**Postdiction accuracy bonus?**	No	No	No	Yes
	CLASS Quiz Calibration Top 10?	No	No	No	Yes

Next, students could further delineate their results by viewing them by both specific attempt (e.g., their first attempt vs their most recent attempt) and by learning objective. The ability to view results by learning objective was designed to allow students to be able to identify which aspects of each topic they were struggling with (e.g., poor scores on learning objectives 5 and 7 in Topic 4). In addition to quantitative information related to the objective feedback, each objective was color-coded to indicate relative severity of the students' bias score (> ± 50% bias = red; ± 25 – 49.9% = yellow; < ± 24.9% bias = green; Figure 3.10), providing a qualitative indication of where to focus their attention. This was designed to help students direct their study and make better-informed SRL decisions.

These new features culminated in allowing for one final significant augmentation in CLASS's ability to support student learning. Beginning in the Fall 2019 semester, students could use the information provided via their learning objective results to generate their own dynamic quizzes to assess their understanding of specific skills. Dynamic quizzing allows students to sort quiz and learning objective results by parameters of their choice (e.g., mean performance, confidence, positive or negative bias) and to either select individual earning objectives or entire quizzes-worth of objectives to generate a customized quiz to assess the topics they want to review. Once a quiz is generated by the student, an attempt will randomly pull questions related to selected objectives for students to assess their learning. Overall, the dynamic quizzing process gives students the tools to isolate specific areas of weakness on their own and to specifically work on those areas during study, while avoiding questions on topics they have demonstrated mastery and accuracy of their perceptions.

Gamified elements (Fall 2019): In attempts to provide an additional influence on students' learning and SRL behaviors in the course and to focus students on utilizing CLASS to improve their judgment skills, we implemented gamified elements into the Fall 2019 semester. As gamification is theorized to mediate student learning through the increase of a mediating factor(s) (e.g., learning strategy use, cue utilization during learning judgments; Landers, 2014), we employed new features within the course in attempts to draw student focus upon increasing these behaviors. This gamified approach consisted of two elements: a) the potential for students to earn a bonus on the exam via accurate global postdictions of performance; b) an opportunity for students to earn bonus quiz points by being one of the top 10 most accurate students on each topic quiz.

Adopting a similar approach to Callender et al. (2016), during each exam of the semester students had the opportunity to gain up to 10 points of extra credit for an accurate postdiction of their total score. As this constituted an entire letter grade (each exam was worth 100 points of course credit), this was designed to represent a significant reward for the student and was put in place to incentivize students towards making the postdiction and to increase their focus upon making judgments in the course as a whole. The bonus structure followed the following criteria:

- If a judgment was ±½ point from their actual score they received 10 bonus points[1]
- If a judgment was ±2 points from their actual score they received 5 bonus points
- If a judgment was ±5 points from their actual score they received 2 bonus points

As Callender et al. identified postdiction feedback as a potentially important component in improving metacognitive judgments in this manner, students were given feedback related to the accuracy of their judgments and what (if any) bonus they received as a result of their judgments. Additionally, and beyond the Callender study, students had the opportunity to triangulate their judgments with local item level judgments both before they made them and in light of the result (e.g., a student could consider individual question judgments to inform global judgments).

To incentivize student consideration of CLASS feedback to increase calibration accuracy, the students who submitted the top 10 most accurate quiz attempts for each Topic quiz were awarded 2 bonus points that were added to their quiz grade for that Topic. As it is possible on a 10-question quiz attempt to earn a perfect score and a zero calibration score

[1] This was to account for grades for short answer questions which could include ½ points. Otherwise students had to estimate their exact score to earn maximum bonus points.

(i.e., full confidence on each question), if more than 10 students received zero calibration scores “tie breakers” (e.g., 100% pre- or post-diction, quiz attempt number) would be used to isolate which students received the bonus up to a maximum of no more than 15 students.

Finally, beginning with the first course meeting, students were instructed on multiple occasions throughout the semester that CLASS was a tool that was provided to help them improve their judgment accuracy and increase their chances to earn the benefits of the bonuses on offer. They were told to make careful judgments, consider accuracy feedback in the form of CLASS variables and exam results, and to use these data sources to be more mindful of their abilities and focus their study. In sum, the gamified elements were designed to increase the student focus upon CLASS feedback and to provide an incentive for students to value the information it provides

4.3 Results:

Quantitative data relating to student performance and local and global accuracy predictions across the study semesters were analyzed using either IBM Statistical Package for the Social Sciences (SPSS; for reliability and mean-based analyses) and Statistical Analysis Software (SAS; for multilevel modeling analyses). Data were screened for outliers and normality prior to performing analyses described below and during analyses all assumptions of the procedures were met unless otherwise specified. All inferential statistics were run at an alpha level of .05. Effect size considerations follow recommendations from Cohen (1988), with sizes for *d* being defined as “small” (*d* = .2 - .49) "medium,” (*d* = .5 - .79) and "large” (*d* > .8) and “small” (η_p^2 = .01 - .059), “medium (η_p^2 = .06 - .139) and “large” (η_p^2 > .14) for η_p^2.

Building upon the promising initial investigations into the use of CLASS, this study sought to further elucidate potential relationships between these variables and to iteratively

begin attempts to positively affect students' monitoring accuracy and potentially improve student learning outcomes. Results are reported in relation to each of the study's primary research questions:

1. *What are the effects of course modifications related to CLASS quizzing over the study period on students' CLASS usage patterns in the target course? Specifically, …*
    a. *…the effects of altering course requirements related to CLASS quizzing (optional vs. 5-attempts required vs. graded)*
    b. *… the effects of altering course/quiz structure (module vs. topic; augmented question banks)*
    c. *…the effects of new CLASS features (e.g., dynamic quizzing)*

To address this question, we considered a combination of trace data from students CLASS use across the study period to isolate how students were using the tool to both fulfill course requirements and prepare for exams.

Usage frequency: Usage frequency in the Fall 2017 semester (Table 4.2) was similar in magnitude to the pilot Fall 2016 semester. This was to be expected as there were no significant changes to the structure of the course or CLASS features. Students in Fall 2017 attempted between 58-102 quiz questions on average per exam period. There were six to eight class meetings between exams and students averaged 9.7 to 12.75 quiz questions per class meeting. In subsequent semesters (Fall 2018 and Fall 2019), students' quiz use increased significantly (Table 4.2). For instance, the cumulative total of questions attempted on CLASS by participants during the study period with the largest increase occurring for the Fall 2018 semester (4559 quiz attempts; 45590 questions answered). Students in Fall 2018 attempted between 128-252 quiz questions on average per exam period. There were six to

eight class meetings between exams and students averaged 21.3 to 31.5 quiz questions per class meeting. In Fall 2019 the frequency decreased, yet aggregate attempts were still double that of the Fall 2017 semester (Table 4.2).

Aggregate usage frequency also varied by exam, with the largest frequencies of practice testing occurring in relation to the first two exams of the semester across all three semesters of the interest (Table 4.2). The introduction of dynamic quizzing in the Fall 2019 semester split student use into two separate modalities. This dynamic usage, which was utilized by half of the overall participants who used CLASS in the semesters (n = 41), mirrored overall usage trends in frequency with the lowest quiz attempts being made between Exam 2 and Exam 3.

Table 4.2 Aggregate CLASS quiz question attempts across semesters by exam.

	Fall 2017 (n=65)		Fall 2018 (n=82)		Fall 2019 (n=73)		
	Assigned	Average /Student	Assigned	Average /Student	Assigned	Average /Student	Dynamic (n=41)
Exam 1	4630	71.2	14560	178	10390	142	1879
Exam 2	6630	102	20640	252	13280	182	2811
Exam 3	3740	58	10390	128	7190	98	1631
Total	15000	231	45590	556	30860	422	6321

Temporal usage patterns: To visualize students' temporal usage of CLASS quizzes throughout the target semesters, we generated scatter plots charting the date stamp collected by CLASS from each individual question attempt submitted to the website against the associated calibration accuracy value for the question (0-1). When visualized in this manner, distinct patterns of student use can be identified (Figures 4.6, 4.7, 4.8). As the Fall 2017 and Fall 2018 semester only had one use case for quiz taking (i.e., instructor-assigned quizzes) there is only one series of data representing all of CLASS use. For the 2019 semester, with

the introduction of dynamic quizzing students were able to take their own quizzes. Those data are represented by a separate series in the corresponding plot.

Fall 2017 - Quiz attempts were largely clustered around exam dates (Figure 4.6) and largely matched the Fall 2016 distribution. The blank areas of the plot represent several days to weeks of non-use between exams (Figure 4.6). The largest blank of the semester-long distribution is present between Exam 2 (which took place on October 17th) and November 2nd (16 days), with no student attempting a quiz during this period. Four class meetings took place during this interval with each covering (then) Module content that was assessed with Exam 3 (scheduled for November 9).

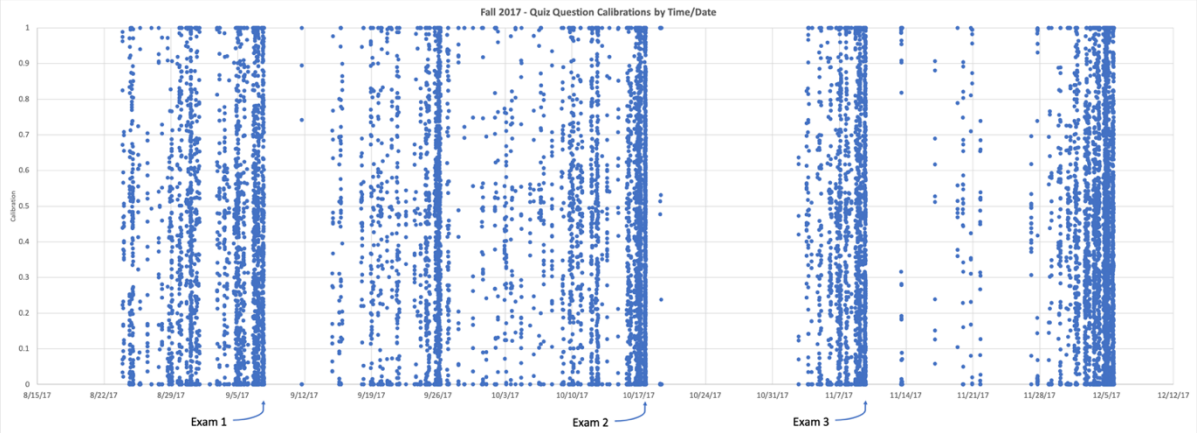

Figure 4.6 Distribution of CLASS quiz attempts during the Fall 2017 semester.

Fall 2018 – The introduction of the 5-attempt requirement for Fall 2018 and the conversion from modules to shorter topics both influenced the temporal distribution of quiz attempts in addition to frequency. Attempts within this period increased in temporal breadth with no large gaps in student use (Figure 4.7). For example, the gap between Exam 2 and the next CLASS quiz attempt for the Fall 2018 semester was reduced to three days. In practice, this means students were assessing their knowledge of new material after the very next

course meeting following Exam 2. This 3-4-day gap between the exam and next CLASS attempt was consistent across all three course exams considered in this analysis.

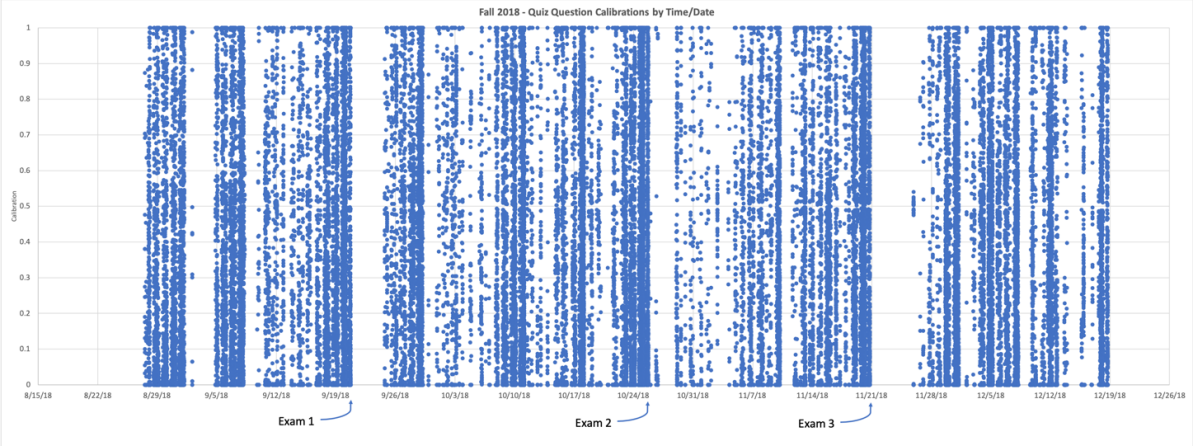

Figure 4.7 Distribution of CLASS quiz attempts during the Fall 2018 semester.

Fall 2019 – The introduction of dynamic quizzing provided two separate series of data to plot over the semester and doing so highlighted student usage patterns. Students could take an unlimited number of quiz attempts with their highest score counting toward their grade. This resulted in decrease in overall number of attempts from Fall 2018 to Fall 2019 but attempts were distributed in a similar manner to the Fall 2018 dataset (Figure 4.8). Plotting the dynamic quiz attempts (orange points) along with the attempts elicited by the requirement (blue points) details that students were largely using the dynamic quizzing feature in relation to course exams. The densest distributions of data points occurred within the 48-hours or so prior to the course exams (Figure 4.8).

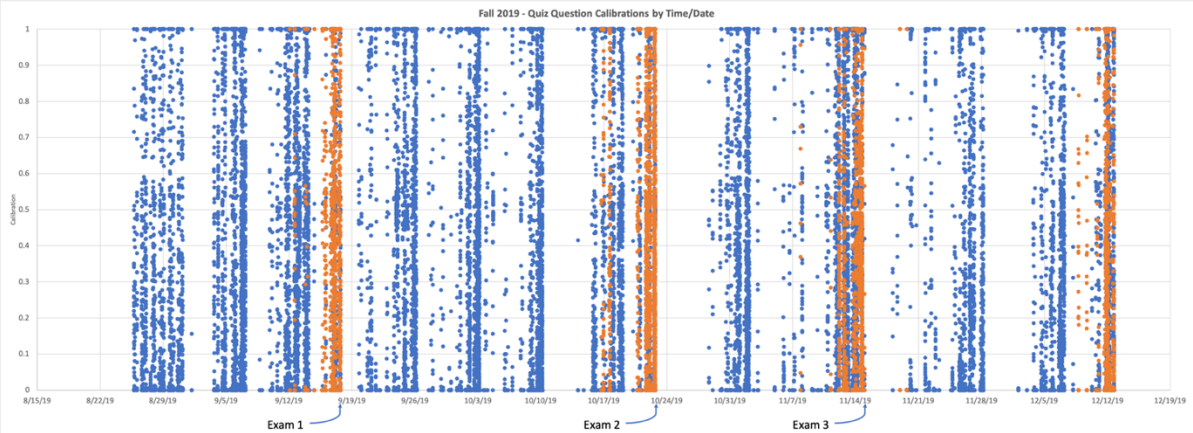

Figure 4.8 Distribution of CLASS quiz attempts during the Fall 2019 semester.

2. *What are the effects of these same course modifications over on student exam outcomes (performance, confidence, calibration accuracy) in the target course across the study period?*

To address this question, we utilized a combination of mixed between-within ANOVAs on exam performance and confidence and a multilevel model predicting exam calibration accuracy with a variety of independent variables.

Pre/Post measures: To determine if there were any significant differences between the three semesters pre-GCI scores (i.e., # of questions correct), one-way ANOVAs were conducted and independent-samples t-test for the two post-GCI scores (Fall 2018 GCI post data was not collected). Paired samples t-tests were run for each of the two semesters with post-test data to determine if statistically significant gains were made from pre- to post-test for each iteration of the course. Finally, we calculated a change score by subtracting students matched post-pre score to represent GCI change as a result of taking the course.

There were no significant differences in students GCI pre-test scores ($F(2, 224) = 0.70$, $p = .49$, $\eta_p^2 = 0.01$) or post-test scores between semesters ($F(1, 101) = -1.195$, $p = .24$, $d = .24$). Significant gains with a large effect size were seen in both courses where post

measures were present to establish gains; Fall 2017 ($t(48) = 8.42$, $p < .0001$, $d = 2.43$) and Fall 2019 ($t(52) = 7.81$, $p < .0001$, $d = 2.17$). Finally, Fall 2019 had statistically significantly higher gains pre- to post-test (M=2.81 SD=2.64) than the Fall 2017 semester (M=1.94 SD=1.61), although this result was only slightly significant with a small effect size ($t(86.96) = 2.03$; $p = .05$, $d = .44$).

Exam performance: *General trends by exam* – Results for three One-way ANOVA analyses (one per exam) indicated that there were no significant differences in Exam 1 scores across the three semesters ($F(2, 226) = 1.30$, $p = .276$, $\eta^2 = .011$) though students in the Fall 2018 and Fall 2019 semesters generally scored higher than their Fall 2017 counterparts (Figure 4.9; Table 4.3). Exam 2 performance analysis, however, resulted in a significant difference between semesters ($F(2, 219) = 4.15$, $p = .017$, $\eta^2 = .037$). Bonferroni corrected post-hoc tests revealed that the Fall 2019 cohort scored significantly higher than the Fall 2017 group (mean difference = 6.0, $p = .016$). Finally, each semester demonstrated almost equivalent performance on Exam 3 with no significant differences present ($F(2, 226) = .161$, $p = .851$, $\eta^2 = .002$). Raw means and standard deviations relating to student performance from each semester of the study are presented below in Table 4.3. Each representative *n* is derived from matched data.

Table 4.3 Performance variable mean and standard deviation across semesters.

	Fall 2017 (*n*= 67)		Fall 2018 (*n*= 69)		Fall 2019 (*n* = 69)	
	M	SD	M	SD	M	SD
Exam 1	78.84	13.84	80.18	13.40	82.19	12.30
Exam 2	75.91	13.57	80.33	12.00	81.57	12.39
Exam 3	75.50	15.70	75.66	15.21	76.31	11.87
Student mean	76.17	14.37	78.72	13.53	80.02	12.52

Exam by Semester – To investigate changes across each exam and across target semesters, a mixed between-within analysis of variance was conducted. The between-subjects factor was semester (0=Fall 2017, 1=Fall 2018, 2=Fall 2019) and within-subjects variable performance on each of the three Exams. Mauchly's test was not significant ($p >$ .05), suggesting that the sphericity assumption was not violated. Thus, the univariate estimation to repeated measures is reported. As expected from one-way analyses, there was a significant main effect of exam on student performance from semester to semester ($F(2, 404) = 16.157$, $p <.001$, $\eta_p^2 = .084$) with post-hoc tests determining Exam 3 performance was significantly lower than Exams 1 and 2 ($p <.001$). The two-way exam*semester interaction was non-significant ($F(4, 404) = 1.757$, $p = .137$, $\eta_p^2 = .084$). Additionally, the between-subjects effect of semester was non-significant $F(2, 202) = 1.401$, $p = .25$, $\eta_p^2 = .014$, with no comparisons between semester-long mean performance values being statistically different. (Figure 4.9). Generally, however, performance trended higher for Exams 1 and 2 from semester to semester aside from the significant decrease in performance demonstrated by the Fall 2019 semester during Exam 3.

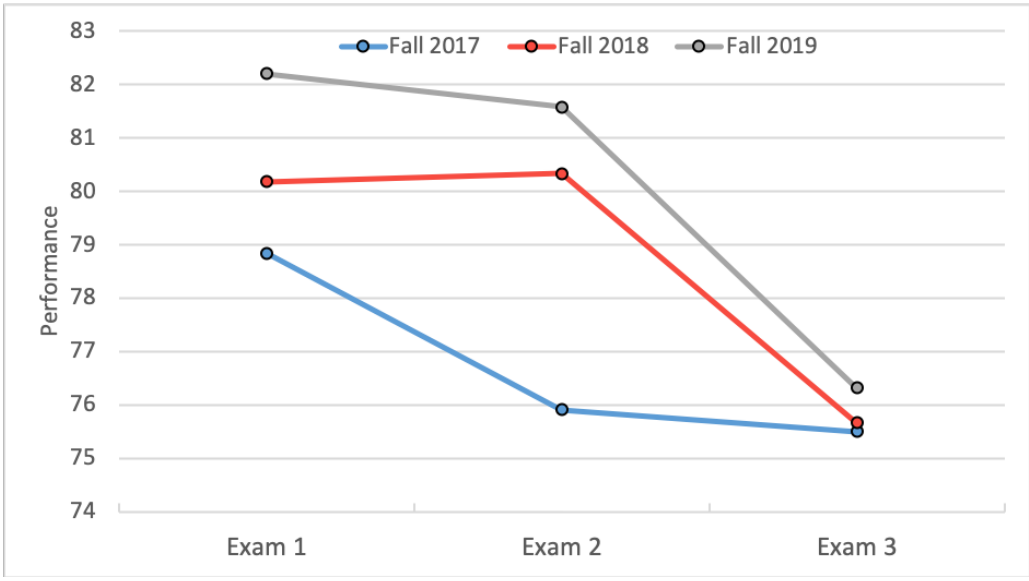

Figure 4.9 Mean performance between each exam across each target semester.

Confidence –A mixed between-within ANOVA was conducted to analyze student confidence derived from local accuracy judgments both across semesters and by each exam. Similar to performance, the between-subjects factor was semester (0=Fall 2017, 1=Fall 2018, 2=Fall 2019) and the within-subjects variable mean confidence collected from item level judgments from the three Exams. Results revealed there was a significant main effect of exam on student confidence throughout the semester, $F(2, 362) = 4.846$, p=.01, η_p^2= .026, with post-hoc tests determining that students regardless of semester were significantly least confident in their judgments during Exam Three (Figure 4.10).

The two-way within-subjects Exam*Semester interaction was non-significant $F(4, 362)$ =.144, p = .97, η_p^2= .002, meaning there were no significant changes in exam-specific confidence from semester to semester for each subsequent exam. The between-subjects effect of semester was significant $F(2, 181)$ =8.345, p <.001, η_p^2= .084, with post hoc analysis revealed all comparisons between semesters (averaged across exams) to be significant other

than the difference between the Fall 2018 and Fall 2019 semesters, which both averaged higher mean confidence than Fall 2017 (Figure 4.10).

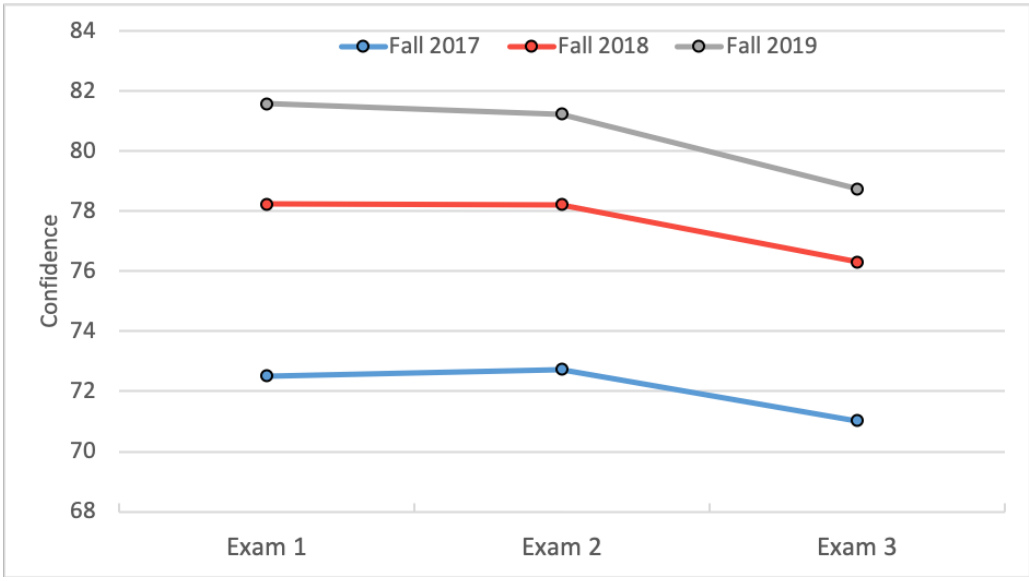

Figure 4.10 Mean item-level confidence between each exam across each target semester collected from course exams.

Local Accuracy – To analyze student calibration accuracy results both across semesters and by each exam per semester, a mixed between-within ANOVA was conducted. As prior, the between-subjects factor was semester (0=Fall 2017, 1=Fall 2018, 2=Fall 2019) and the within-subjects variable mean calibration accuracy derived from each of the three Exams. Results revealed there was a significant main effect of exam on student calibration accuracy throughout the semester, $F(2, 362) = 11.231$, $p < .001$, $\eta_p^2 = .058$, with post-hoc tests determining that students regardless of semester were significantly most calibrated in their judgments (i.e., had lower calibration scores) during Exam One and that this accuracy declined (i.e., calibration scores increased) as the semester progressed. The between-subjects effect of semester was non-significant $F(2, 181) = 1.246$, $p = .29$, $\eta_p^2 = .014$, meaning mean calibration accuracy across all three exams was not significantly different between semesters.

Additionally, there was a significant two-way interaction between exam and semester, $F(4, 362) = 2.603$, $p = .04$, $\eta_p^2 = .028$. Decomposing the interaction, no pairwise comparisons of calibration accuracy were significant between exams across semesters (i.e., no significant differences between exams across semesters), but there were significant increases across exams between semesters with all significant changes being due to a decrease in accuracy during Exam 3 of the Fall 2019 semester (Figure 4.11). Generally, calibration accuracy trended lower (increased in accuracy) for each subsequent semester aside from the significant increase in calibration (decrease in accuracy) demonstrated by the Fall 2019 semester during Exam 3. Local accuracy descriptive statistics are provided in below in Table 4.4.

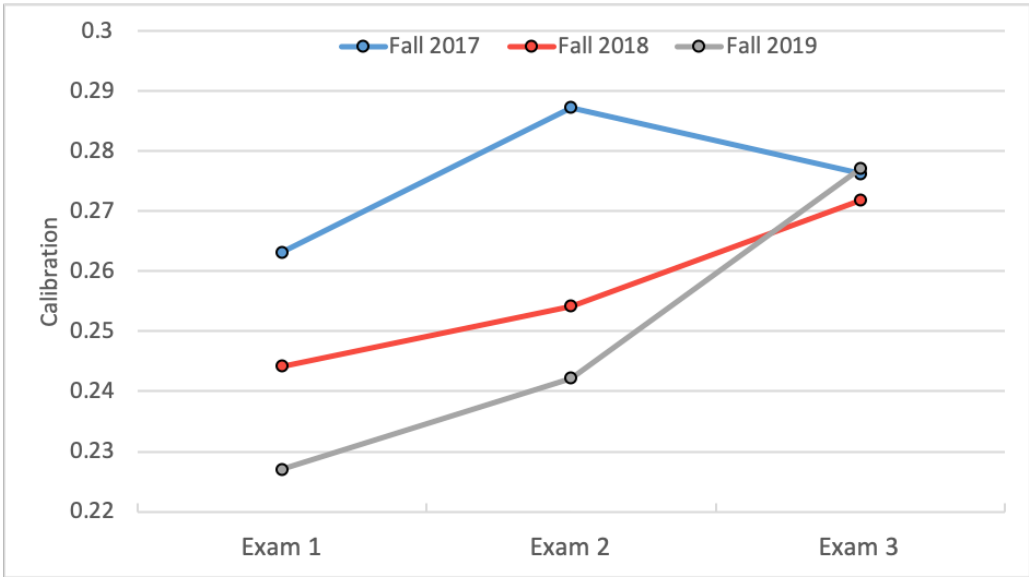

Figure 4.11 Mean calibration between each exam across each target semester.

Table 4.4 Local accuracy variables mean and standard deviation from 2017-2019.

	Fall 2017 (*n*=57)		Fall 2018 (*n*=58)		Fall 2019 (*n*=69)	
	M	SD	M	SD	M	SD
Exam 1 Calibration	0.26	0.11	0.24	0.11	0.23	0.10
Exam 1 Confidence	72.51	14.77	78.24	11.59	81.58	11.06
Exam 2 Calibration	0.29	0.11	0.25	0.11	0.24	0.11
Exam 2 Confidence	72.73	14.29	78.21	12.00	81.52	9.83

Table 4.4 Continued

Exam 3 Calibration	0.27	0.12	0.28	0.11	0.27	0.10
Exam 3 Confidence	71.03	16.79	76.71	13.52	78.74	13.16
Mean Calibration	0.27	0.11	0.25	0.11	0.24	0.10
Mean Confidence	72.02	15.28	77.72	12.37	80.61	11.35

Multilevel modeling of calibration accuracy: For traditional 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. Often in this course, students would neglect to provide judgments. This led to attrition for local accuracy analysis in prior investigations (up to 30+ individuals during the pilot investigation). To help remedy this phenomenon and to determine sources of variability in calibration accuracy, multilevel modeling was used to analyze the influences on exam calibration (e.g., GCI pre-test score). This analysis and model results from this setting will described in detail in Chapter 5 in reference to its comparison to a regional community college.

3. *What are the effects of employing gamified elements in Fall 2019 on global judgment accuracy when compared to prior semesters (Fall 2017 and 2018)?*

To address this question, we analyzed students' global exam postdictions across the target semesters comparing the accuracy of students in the two semesters (Fall 2017, 2018) with no gamified elements to the accuracy demonstrated in the gamified semester (Fall 2019).

Judgment characteristics (Fall 2017/2018 vs Fall 2019): Considering the general features of students' global postdiction judgments across the semesters, students in the Fall 2019 semester provided a larger frequency of unique values than their peers in previous semesters. The Fall 2017 and Fall 2018 judgments were multi-modal and were predominantly whole number approximations that corresponded to grade cut-offs (e.g., 70, 80, 90; Figure 4.12). While judgments from the Fall 2019 semester still featured these judgments, judgments were more fine-grained with a larger frequency of unique values than in prior semesters (Figure 4.12).

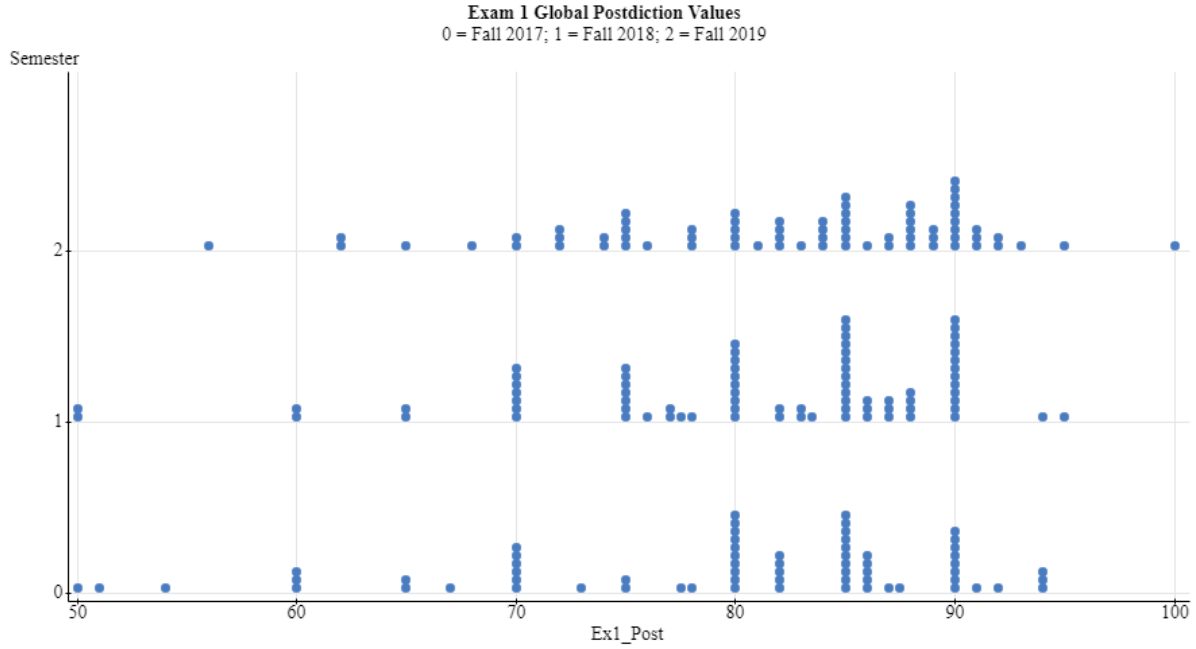

Figure 4.12 Dot plot of Exam 1 postdiction values by semester.

Global accuracy across semesters: We examined global postdictions across both the Fall 2018 (similar course and CLASS features, non-gamified) and Fall 2019 (gamified) semesters to analyze student accuracy. The global signed measures of accuracy data (the signed difference between a student's postdiction judgment and their performance on each exam) were not normally distributed, rejecting the null hypothesis for both the Shapiro-Wilk

test for normality and Levene's test for the assumption of equal variances. As a result, to compare the effects of the implementation of the gamified elements on student postdictions non-parametric tests were used. To determine if there were differences in average signed accuracy between the two semesters, Mann-Whitley U tests indicated that global signed accuracy between the Fall 2018 and Fall 2019 were not significantly different ($U = 5966.5$, $p = 0.2612$) with the Fall 2019 semester being indicated to be approximately 2 percent overconfident (M=2.12 SD= 5.70).

Changes across Fall 2019 by group – While there was no significant difference between global signed accuracy between the Fall 2018 and Fall 2019 semesters, we also sought to determine if there were any fluctuations in accuracy between the Fall 2018 and Fall 2019 semester by performance quartile. To do so, students were separated into performance quartiles based on their final course grade (1 = lowest performers; 4 = highest performers). Considering mean signed accuracy by group, Wilcoxon Signed Ranks tests on each group were performed to see if each group's accuracy was significantly different than zero. Results revealed hypothesized trends in that higher performers were underconfident (negative signed values) and the lowest performers were the most overconfident (Table 4.5). Additionally, looking between Fall 2018 and Fall 2019, the lowest quartile demonstrated less overconfidence in 2019 than 2018 (Table 4.5). Similarly, the highest performers demonstrated less underconfidence (Table 4.5).

Table 4.5 Wilcoxon signed ranks of global bias by performance quartile.

	Quartile	*n*	Median Est.	*W*	*p*
Fall 2018	1	20	7.85	173.5	.01
	2	20	3.33	158	.04
	3	18	-0.25	80	.83
	4	21	-4.96	2	<.0001
Fall 2019	1	18	6.17	159	.0015
	2	18	4.67	150.5	.005

Table 4.5 Continued

	Quartile	*n*	Median Est.	*W*	*p*
Fall 2019	3	18	0.58	100.5	.5277
	4	19	-2.75	7	.0004

Looking at the Fall 2019 semester by exam, Figure 4.13 reports all students' signed global postdictions by performance for each exam. In the figure, each individual student's bias for each exam is represented by bar. Positive values indicate overconfidence and negative values represent underconfidence. The shaded areas represent the earned grade for that exam (red < 60%, orange 60-70%, yellow 70-80%; pale green 80-90%; darker green 90-100%). Students who earned below a 70% on the exam were most often overconfident, estimating exam scores that were often 10 points or more above their actual scores. In contrast, students who earned above 90% on exams were typically underconfident, predicting scores that were several points lower than they earned (Figure 4.13). Students earning between 70% and 90% were generally the most accurate with their postdictions and were often within ±5 points off their actual score (Figure 4.13).

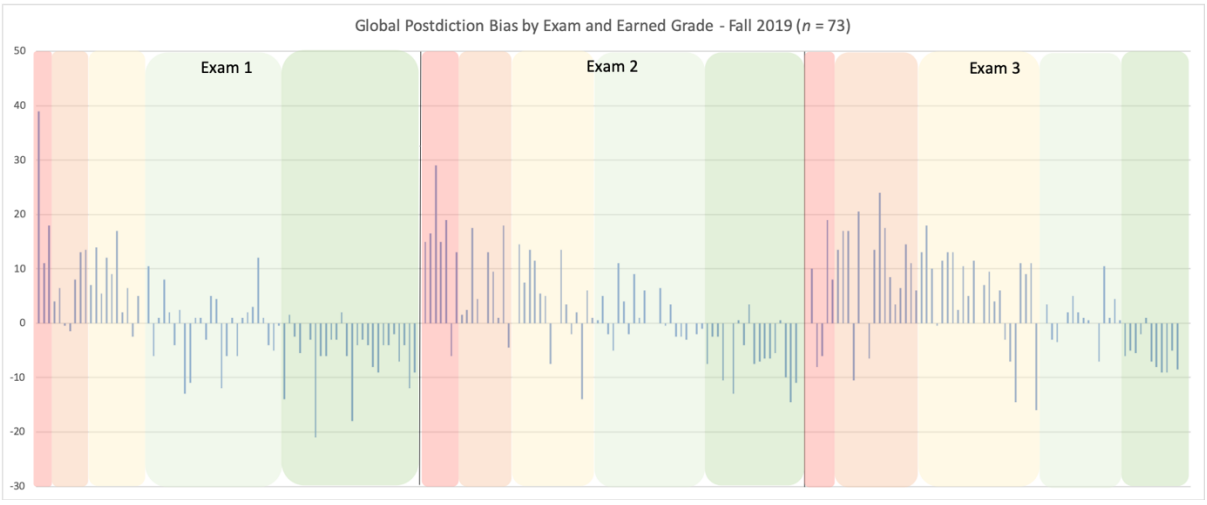

Figure 4.13 Global postdiction values (%) by exam and earned grade for Fall 2019.

Bonus allocation – To determine the net effect of the bonus for Fall 2019 and to compare results to semesters prior to the bonus, we binned global exam postdictions from prior semesters into the ranges that corresponded with the bonus ranges for the Fall 2019. Overall, roughly half students in the Fall 2019 received a bonus of any magnitude for first two exams (50% for Exam 1 and 46.2% for Exam 2; Table 4.6). This frequency diminished with Exam 3 as less earned bonuses than during prior exams (35.7%; Table 4.6). Expanding this consideration to prior semesters, students in the Fall 2018 would have received the most total bonuses (45% of possible) than the Fall 2017 semester (41%) though the Fall 2019 semester earned the second most overall, though only by one percent (44%). The total number and percentages of bonuses that were earned (or potentially earned) by students is presented in Table 4.6 below. Note that each percentage is cumulative in that, for example, 4.2% of students in the Fall 2019 semester earned the 10-point bonus during Exam 1 while an additional 18% of students earned a 5-point to generate the 22.2% value reported and an additional 29.2% of students earned a 10-point bonus to generate the 51.4% value reported as the total percentage of students who earned a bonus of any amount (Table 4.6).

Table 4.6 Global postdiction incentive bonus percentage by semester.

	Semester:	+10	+5	+2	*n*	Total Bonus	No Bonus
Exam 1	Fall 2017*	4.8%	20.6%	46.0%	63	29	34
	Fall 2018*	4.0%	14.7%	37.3%	75	28	47
	Fall 2019	4.2%	22.2%	51.4%	74	37	37
Exam 2	Fall 2017*	1.9%	17.0%	43.4%	53	23	30
	Fall 2018*	1.5%	18.5%	47.7%	65	31	34
	Fall 2019	9.0%	23.9%	47.8%	67	30	35
Exam 3	Fall 2017*	5.6%	22.2%	51.9%	54	28	26
	Fall 2018*	8.1%	29.0%	51.6%	62	32	30
	Fall 2019	10.0%	20.0%	35.7%	70	25	45

*Bonus was only in place for the Fall 2019 semester. Remaining values are for comparison

4.4 Discussion

Major findings: This analysis sought to characterize the relationship between iterative course changes and alterations to a researcher designed quizzing website to help determine the best practices surrounding its use in an introductory physical geology course. In addition, we sought to improve the performance, confidence, and SRL behaviors of undergraduate students. Towards this aim, we implemented a suite of alterations that were hypothesized to make students more aware of their metacognitive monitoring and SRL behaviors via increased interaction with CLASS and the utilization of the feedback it provides. Below we will consider implications of quantitative signals in relation to each of the study's primary research questions:

1. *What are the effects of course modifications related to CLASS quizzing over the study period on students' CLASS usage patterns in the target course? Specifically, …*
    a. *…the effects of altering course requirements related to CLASS quizzing (optional vs. 5-attempts required vs. graded)*
    b. *… the effects of altering course/quiz structure (module vs. topic; augmented question banks)*
    c. *…the effects of new CLASS features (e.g., dynamic quizzing)*

As expected, there was a significant increase in CLASS quiz usage as a result of the 5-quiz requirement that was implemented between Fall 2017 and Fall 2018. The more than doubling of quiz attempts and overall increase in student usage provided students with significantly more opportunities for practice leading up to course exams. This alteration along with the subdivision of course structure units/quizzes into smaller "chunks" of material elicited an increase in the temporal distribution of students' practice across the semester with

almost no gaps in use (as opposed to multiple weeks without practice in prior semesters with optional CLASS use). Thus, this result provides evidence that instructors' decisions regarding how they choose to structure their course (if mindful to effective study strategies) can directly lead students towards adopting effective learning strategies (e.g., distributed practice and practice testing; Dunlosky et al., 2013).

The introduction of a new use modality (dynamic quizzing) and the second alteration of course requirements, led to the splitting of student use into two distinct modes that aligned with the alterations in the course requirements: taking quizzes for a grade and taking quizzes to prepare for exams. While altering the five-attempt quiz requirement into a highest-grade requirement lowered the overall number of questions attempted in an aggregate sense (45550 attempts in Fall 2018 to 37181 total attempts in Fall 2019 including dynamic quizzes), the combination of this alteration with the new feature provided insight into how students utilized each use. Though it may be self-evident, students largely respond to tasks they are required to complete within a course. Dividing the course content into smaller "chunks", increasing focus of that content's breadth by adding more questions, and embedding effective study strategies into course design is one manner to guide students towards effective learning behaviors.

Finally, these results provide insight into instructor's decision-making during course design. These decisions regarding how to structure a course are directly tied to the experiences a student has while in the course. For example, the Module format of the target course and having optional mastery quizzes led to students not assessing themselves (not via instructor provided resources at least) for multiple weeks of new content instruction. Students learn more effectively the earlier they are assessed on their knowledge of new information

(e.g., the testing effect; see Eisenkraemer, Jaeger, & Stein, 2013 for a review). The initial decision to make CLASS assessment optional resulted in students generally declining to engage with the quizzes until a short time before the exam. Once the course structure was altered, students accessed the quizzes earlier and more frequently. As a result, these results provide an example of how an instructor can directly embed effective learning strategies into a course through conscientious course design.

2. *What are the effects of these same course modifications over on student exam outcomes (performance, confidence, calibration accuracy) in the target course across the study period?*

The course changes led to generally positive effects, with some mixed results. Each sequential cohort of students generally performed better and were better calibrated than their peers in the preceding semester for Exams One and Two (Table 4.4; Figure 4.9). However, this trend disappeared for Exam 3 which had little difference among the three semesters. Additionally, while the performance of Fall 2019 students decreased relative to their peers for Exam 3, their mean confidence did not show a parallel decrease (Figures 4.10, 4.11). As a result, they were less calibrated than for their first two exams. This affected the statistical significance of semester-wide, mean-based analyses of both performance and calibration accuracy results.

Important to the pursuit of potential discipline-based variations in student performance and metacognitive monitoring, and similar to prior work (Jones & McConnell, 2017), students in each of the semesters of this study were best-calibrated in their local calibration accuracy during the first exam of the semester (though not always significant; Figure 4.11). This runs contrary to work in chemistry which suggested 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). In our course, we saw the opposite effect, perhaps indicating a stronger influence of topic content on students' accuracy than signaled in other STEM-related calibration studies.

While performance and local accuracy results were mixed, one result that was not mixed was the mean level of confidence displayed during exams by each cohort. The semesters with required quizzing (Fall 2018/9) saw significantly increased confidence than the prior semester (Figure 4.10). As CLASS quiz attempts in these semesters increased significantly as a result of the change in requirements, students had many more opportunities to practice and become familiar with the questions and question styles used for course exams. This likely impacted their experience and contributed to increased confidence during exams.

This disparity between the increase in confidence but a lack of significant gains in local accuracy (calibration) may of course be influenced by the specific geology content assessed during those modules/topics, but related to this study's design it is important to note that the subdivisions of Modules into Topics were not applied to the units providing content for Exam 3 (i.e., the Modules from Fall 2017 are the same as the current Topics in terms of content and breadth). Almost serving as a control between semesters, the lower quiz attempts during this period and the decrease in performance for each of the (before then) higher performing/more accurate semesters provides interesting suggestions of the effect of the course structure changes. Future iterations of the course will seek to subdivide these topics and investigate results.

3. *What are the effects of employing gamified elements in Fall 2019 on global judgment accuracy when compared to prior semesters (Fall 2017 and 2018)?*

From a practical research perspective, as attrition due to non-response (e.g., not making global or local judgments) is a common issue in prior iterations of the course where confidence data was collected, the byproduct of the gamification process of increasing student response rates was a positive result. The Fall 2017 global judgment analysis across all exams, for example, was missing judgments from more than 50 students. This non-response across all exams was reduced to 10 global judgments for the Fall 2019 semester. A similar effect was found for the local judgments (with no local accuracy attrition in Fall 2019), and as the students most likely to not respond are lower-performing, this provides some insight into the results of the comparisons between the Fall 2017/18 and Fall 2019 semester. While we attempted to correct this via our research design, the influence of this potential (lack of) selection bias must be considered and likely contributed to higher mean global accuracy values for the Fall 2019 semester.

Considering global accuracy by course grade performance quartile, students in the course largely followed the trend from the literature in which lower performers overestimated their performance and the higher performers slightly underestimate their performance (e.g., "unskilled and unaware" (Kruger & Dunning, 1999). There were suggestions that effect was lessened in the Fall 2019 semester with lower performers decreasing their overconfidence and higher performers decreasing their underconfidence (see Figure 4.12, Figure 4.13 and Table 4.6), but no equivocal statements can be made. Further investigation to within-group variation and fluctuation in accuracy across the semester is warranted.

Finally, the net effect of the gamification elements added to the Fall 2019 were generally positive (e.g., increase in intent, increase in judgment participation). While there were some signals of these changes having an impact on accuracy, students in the Fall 2019 semester largely mirrored the Fall 2018 cohort in these results. As they shared the same increased level of interaction with CLASS (compared to Fall 2017), it is likely that these changes cannot be attributed to gamification elements.

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. Another limitation was the significant attrition experienced for the comparisons of local accuracy during the Fall 2017 and Fall 2018 semesters as a result of the requirement for paired data in repeated measures ANOVAs. This problem, however, was much less than historical semesters of the course perhaps due to the focus on judgments as a result of CLASS use. 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. As noted above, this problem was completely mitigated in Fall 2019 due to the gamification-related requirement which was a positive effect of its implementation. This is important to consider in light of results as many of the participants who did not complete the dataset in prior semesters (e.g., Fall 2017) were often the lower-performing students. As these students are often overconfident relative to their performance, this likely introduced a negatively-skewing influence on mean considerations of local accuracy variables (e.g., performance, confidence,

and calibration) in the mixed-design ANOVAs and positively inflated global judgment analyses, particularly for the Fall 2019 semester.

In addition to this influence, it is important to note potential ceiling effects associated with these analyses. The mean performance for the first two exams of the Fall 2019 semester were above 80% and average calibration accuracy values were less than .23. When looking for evidence of improvement via mean-based analyses, natural variability limits the potential for improvement. Future work will expand the numbers of participants and seek to look beyond means into more-individualized and regression-based analyses to investigate group variability.

Additionally, CLASS provides directed information related to student's individual strengths and weaknesses as they evolve throughout the course. This information can be utilized by the student to improve their understanding. This information can make the study process (and employing effective SRL behaviors) much more efficient. Measures of this potential efficiency, however, were not measured as part of this study. For example, one may imagine student (particularly a successful one) studying for several hours leading up to a geology exam in prior iterations of the course, that may now only need to study for one hour to feel adequately prepared for the exam as a result of the feedback CLASS provides. While this feedback may influence judgment accuracy (and likely does based on these results), information from CLASS may be used to alter study behaviors that lead to increased performance and/or confidence that may have been present when the student took the exam regardless. CLASS could have just made the end result easier to achieve. Future work will seek to investigate this potential for CLASS to influence students' SRL and study behaviors.

Conclusion and recommendations for practice: Overall, these iterative course changes made over the three target semesters succeeded in increasing overall CLASS quiz usage, distributed this use over time, and positively influenced student exam outcomes in the course. While the effects of gamification were not potent, the process served to focus student intent in making global judgments and led to all students participating in the process. Lessons to be learned for instructors are that decisions made regarding how to structure your course have direct effects on not only student behavior, but their learning process. Making course design decisions with the explicit intent of inducing effective learning strategies (e.g., distributed practice, practice testing) can impact student learning. As CLASS itself is built with these principles in mind (e.g., backward design - questions tied to objectives, SRL – feedback to inform study practices, etc.), this work further suggests the potential for CLASS as a powerful tool for instructors to support student learning for courses of any discipline.

CHAPTER 5: CLASS AND INSTITUTION TYPE: USING A WEB-BASED ASSESSMENT TOOL TO ANALYZE THE RELATIONSHIP BETWEEN STUDENT PERCEPTIONS OF ABILITY AND EXAM PERFORMANCE ACROSS DIFFERENT TYPES OF INSTITUTIONS

Prepared for submission to *Internet and Higher Education*

5.1 Background

In recent years several reports have highlighted the relative dearth of STEM degrees being conferred each year as compared to the number of students who start out pursuing degrees in STEM fields (e.g., NCES, 2019). This problem has particular relevance for the geosciences as current projections suggest that by 2025 industry will require more than 150,000 geoscientists to fill professional need (AGI, 2011) and that there is currently deficit of students rising to meet this growing demand (Wilson, 2017a).

One higher education environment that has exerted a growing influence on the total number of geoscience students is the two-year college (or 2YC). Approximately a quarter (27%) of those who earned a geoscience degree in 2016 spent at least a semester in a two-year college (Wilson, 2017b). Additionally, student experiences while in a 2YC setting (e.g., academic opportunities, mentorship) have been shown to significantly affect students' intent to major in geosciences (Wolfe, 2016). We sought to explore how factors such as metacognition and self-regulated learning (SRL) in 2YC students compared with those of students enrolled in a similar geoscience introductory course in a research-intensive university.

Factors leading to student success – Metacognition and SRL: There are many factors that influence the level of student success on a specific learning task. One important

contributor is metacognition, one's ability to recognize the workings and characteristics of their own knowledge and thought processes (Flavell, 1979). Metacognition 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 linked processes: planning, monitoring, and evaluation (Jacobs & Paris, 1987). Generally speaking, the concept of ones "self-regulation" represents the cycling of these processes during a learning task. Functionally, this 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 they interact with and internalize course content. Effective 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 correct deficiencies (Zimmerman, 1990). Within an educational context, the process is referred to as self-regulated learning (SRL; Zimmerman, 1990). Interventions that seek to foster student use of SRL behaviors, especially monitoring and controlling one's learning, interventions that focus on metacognitive support have the potential to foster students' successful learning (e.g., Bannert & Reimann, 2012; Künsting, Kempf, & Wirth, 2013). A robust body of research has confirmed the beneficial effects of metacognitive support on learning in computer-based learning environments (Devolder, van Braak, & Tondeur, 2012; Zheng, 2016). Further, several studies have suggested that learners who are better able to identify and employ effective SRL strategies demonstrate higher academic achievement (e.g., Mega, Ronconi, &

De Beni, 2014; Zimmerman & Martinez-Pons, 1988), more intrinsic motivation (Pintrich & Zusho, 2002), and greater persistence (Pintrich & De Groot, 1990) than learners that do not.

SRL has been the subject of many theoretical models that have been generated by different researchers over the last thirty years. A review of the six most prominent SRL models (Panadero, 2017) reveals that all contain some common features; a) consideration of specific learning tasks; b) the student's pre-task behaviors (e.g., planning/selection of strategies); c) monitoring of strategies employed; and, d) post-task evaluation regarding the relative success of the learning strategies that can then be applied for future study events. On a practical level, the manner in which students navigate their lives both during and after their experiences in higher education will be dictated by how effectively they can interact with and learn new information. Whether they are on a path to a future geoscience career or are simply fulfilling a course requirement, students need to be able to effectively self-regulate their learning instead of memorizing specific facts that will change as scientific understanding alters with new discoveries. Additionally, during the process of geoscience learning specifically, a learner must process large quantities of specific and unfamiliar terminology to be successful (Kortz, Grenga, & Smay, 2018), thus making the ability to employ effective SRL practices an important factor for success.

SRL and 2YCs: As two-year colleges often seek to support students with limited academic experiences, students may arrive in these environments with less academic preparation than their peers at research universities. This creates a unique opportunity to investigate how to implement practices and provide tools that foster students' metacognitive skills to support effective SRL behaviors. Metacognitive awareness is not guaranteed and has been suggested to be independent of general ability (e.g., Pressley & Ghatala, 1990; Schraw,

1998), investigating best practices for supporting discipline-specific SRL and metacognition in this environment represents a significant gap in the literature.

Pape-Lindstrom Eddy & Freeman's (2018) report on research findings in biology course represents a rare investigation of SRL in a 2YC institution. They reported the results of a multi-semester study on reading assignments and associated online multiple-choice quizzes. These assignments were presented as tools to increase course structure and support students' SRL by providing feedback during self-study that students would theoretically use to regulate their learning, therefore better preparing them for future active learning lecture sessions (Pape-Lindstrom, Eddy & Freeman, 2018). Results revealed that the semesters with the reading quizzes (controlling for other student variables such as non-biology GPA) saw a significant increase (~5%) in exam performance (Pape-Lindstrom, Eddy & Freeman, 2018).

In summary, the two-year college environment has been shown to be a growing resource for the geoscience community in providing transfer students that increase the number of geology majors to fill the impending professional need. Additionally, as a result of situational factors such as academic preparedness and a prevalence (geoscience-wide) of didactic lecturing (Egger, Viscupic, & Iverson, 2019), these environments may benefit from the addition of metacognitive support and tools to aid in learning. We sought to investigate baseline judgment accuracy data in a two-year college setting prior to supporting the incorporation of a web-based assessment tool that provides metacognitive and SRL-related feedback.

Baseline data collection and *CLASS*: We previously (Fall 2014) collected item-level confidence judgments (Figure 3.1) on paper-based midterm exams administered as part of an introductory physical geology course at a large southeastern research university. Students

were asked to report their level of confidence that their answer was correct for each exam question. When compared to performance on the question, these measures generate a metric for metacognitive monitoring accuracy. For more details of this analysis and subsequent investigations of best practices regarding supporting its use please see Chapter 4. There was a significant correlation between student performance and the absolute accuracy of these judgments (or calibration; Figure 5.1; Jones & McConnell, 2018).

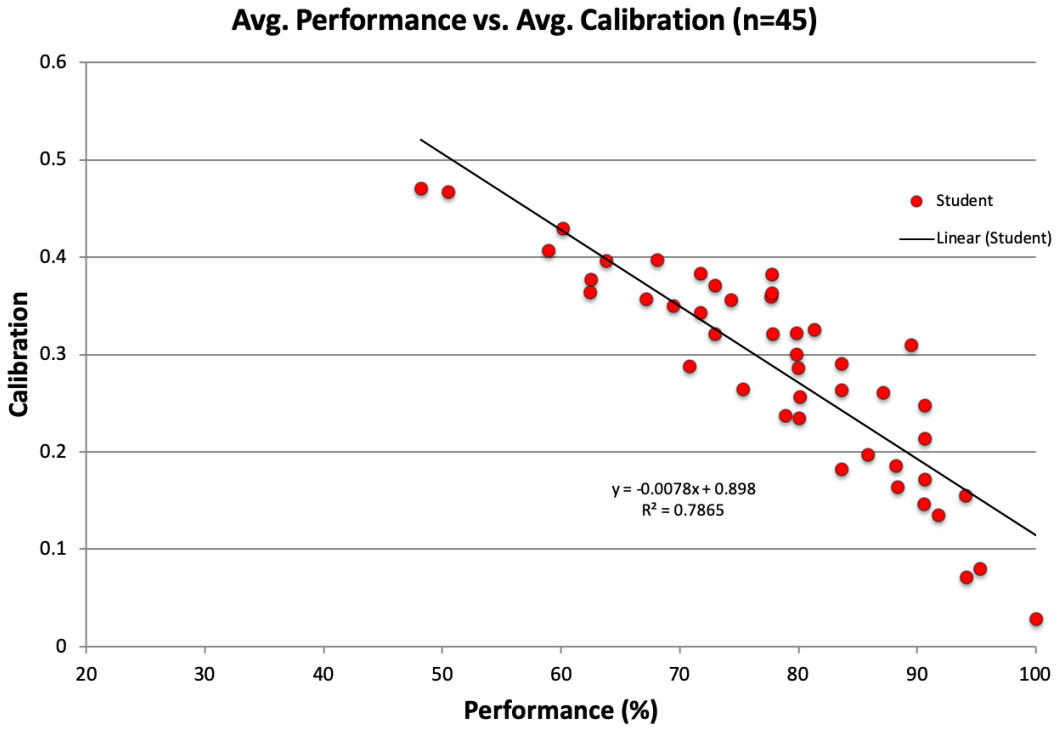

Figure 5.1 Mean performance vs. mean calibration for the baseline Fall 2014 semester.

We subsequently generated a quizzing tool called the Confidence-based Learning Accuracy Support System (CLASS) to turn this process of data collection into a novel mechanism to provide students metacognitive feedback to guide SRL behaviors. (For more on the features and development of CLASS, refer to Chapter 3.) Students in the target course were provided with an opportunity to take optional CLASS mastery quizzes aligned to the

learning objectives for each unit of geology content material. CLASS quizzes were intended to provide give students an opportunity to assess their learning of the content in a no-stakes environment. The CLASS tool allowed students to assess their content knowledge while providing them with real-time feedback regarding the accuracy of their perceptions. The target course contained extensive instructional structure to support students' progress through the course (e.g., active learning exercises during course meetings, flipped-learning environments requiring students to complete homework-style assignments prior to class) prior to the addition of CLASS. Regardless, students in the pilot semester demonstrated higher performance and improvements in some metrics of accuracy (e.g., global postdictions of exam performance; Jones & McConnell, 2018).

Research Questions: Building upon this baseline correlation and promising introduction of CLASS to the research institution, this study sought to further elucidate potential relationships between these variables in the 2YC setting. After developing a baseline characterization of the relationship between performance, confidence and calibration accuracy, we sought to investigate the effects of adding CLASS to an equivalent introductory geoscience course at a regional 2YC. Specifically, we sought to investigate the following questions:

1. *How do students from different types of institutions (Research University vs Community College) compare in their baseline performance and judgment accuracy?*
2. *What are the differential effects of providing CLASS quizzes and accuracy feedback to students at these institutions? Specifically, …*
    a. *…student use variations across semesters between institutions*

b. *...the effects of providing increased course structure and CLASS quizzes on students' exam outcomes in community college setting*

5.2 Methods

To investigate these research questions and to further characterize the discipline-specific monitoring behaviors of college-level geoscience learners across each setting, a situated quasi-experimental mixed-methods study was designed and implemented across multiple semesters. In attempts to mitigate confounding variables, efforts were made to control for many situational factors affecting each cohort. For details regarding research university data collection, please review Chapter 4. Each target semester at the Regional Community College (RCC) in the southeastern US was co-taught by the same pair of instructors, with each alternating as lead instructor between academic years. The course was offered during the same time of day (late morning), and measured student performance via one of two equivalent exams (one set per lead instructor). This report focuses upon the quantitative data collected from participants' exams and during their interactions with a researcher-designed online quizzing website (CLASS).

Participants and setting: RCC participant cohorts consisted of five convenience samples of undergraduate students (~5-20 students per semester; see Table 5.1). Participants were students enrolled in an introductory physical geology course that consisted of two lecture meetings (80 min each) and one lab meeting (150 min) per week. Data collection was spread across one baseline academic year (2017-2018) and three sequential experimental semesters where CLASS quizzes were offered to students with differing course requirements (Fall 2018 – Fall 2019). Students self-selected into the course via enrollment. The sample population contained a male majority (62.4% across the sample). Other demographic data

was not collected/available. This research was conducted in accordance with an IRB protocol declared exempt by the research institution and adopted by the RCC IRB office. Participants were presented informed consent project on the first day of the course and could withdraw at any time. Signed consent forms were sealed until the final week of the semester and the instructors were not made privy of participants. From each course population, study-related subsets were included for analysis based on minimum age requirements (18+) and signed consent as per IRB approval.

Table 5.1 Student gender, sample size and lead instructor for each target semester.

	Fall 2017		Sp. 2018		Fall 2018		Sp. 2019		Fall 2019	
	count	%	count	%	count	%	count	%	count	%
Male	10	52.6	12	60	12	52.2	13	76.5	12	66.6
Female	9	47.4	8	40	11	47.8	4	23.5	6	33.3
Total	19	100	20	100	23	100	17	100	18	100
Lead inst.	Instructor A		Instructor A		Instructor B		Instructor B		Instructor A	
CLASS use?	No		No		Optional		Optional		Required	

Course features: While courses at both institutions focused upon introductory college-level geology concepts, there were several differences in their course features, student support structures and approaches to instruction. We will describe these features in relation to each institution below:

Research university – The research university introductory physical geology course sample was taken from a large-enrollment course with high structure (e.g., several smaller assignments). In-class meetings featured frequent active learning exercises (e.g., peer instruction, small group discussion) and several support structures demonstrated to support student learning outside of the classroom (e.g., lecture slides shared, short videos to support learning of basic concepts, online mastery quizzing). Additionally, all of these activities were

united throughout the course via explicit learning objectives (~100/semester). Worded as directives (e.g., "I can define the water table."), these objectives served as the basis for adopting CLASS quizzing as each question in the system is tied to a specific learning objective to facilitate objective level feedback (see Chapter 3 for more information). Finally, students in the research university setting were assessed via three summative midterms and one final (not included in analysis) that in total represented between approximately two-thirds of a student's final grade.

Regional community college – The RCC course sample was taken from a small-enrollment introductory geology course with moderate structure consisting of homework assignments administered via an online platform provided by the textbook publisher. Course meetings consisted of didactic lecturing interspersed with call and response questioning to assess individual students' understanding of lecture topics. Formal support for students consisted of the textbook and reference information present in the online homework platform. The course had one explicit learning outcome prior to the experimental semesters (*"Upon completion, students should be able to recognize the most common minerals and rocks that can be found in nature and recognize and describe basic geological processes that shape the earth."*). The course content was subdivided in relation to the 16 textbook chapters covered in the course and students were evaluated via 5-6 lecture exams (each assessing 3-4 textbook chapters of geology content and approximately 40% of a student's final grade). The first four exams are analyzed for this report (Table 5.2).

Course revisions: *Pre-work* – Exam judgments at the RCC were collected to determine baseline performance, confidence and accuracy for the first academic year (2017-18) prior to the incorporation of the CLASS tool. The instructor-pair met with the author to

learn about the details of project and to isolate a suite of learning objectives for the target course in preparation for the adoption of CLASS quizzing during the second academic year (2018-2019). These learning objectives mirrored the grain size of the research university course's objectives and allowed for quizzes to be created within CLASS to support student learning in the course. Beginning at the end of this meeting and continuing over the Fall semester, question banks for fifteen quizzes were generated. Research university course objectives were used wherever learning objectives overlapped. Consequently, many of the questions are shared between the two institutions in their respective quizzes.

Introduction of CLASS – These fifteen CLASS quizzes were made available to RCC students beginning in the Fall 2018 semester. During the first academic year (2018-19), similar to the research university setting, the quizzes were optional and not considered as a part of the course grade. The author visited the course to describe the project, collect informed consent and pre-measures and provide a demonstration (~10-15 min) on how to sign up for CLASS, take quizzes, and view/consider results. These requirements were changed for Fall 2019, with students being required to take quizzes in CLASS as part of their course grade. Requirements matched those of the research university setting for this same semester (Chapter 4) in that students had unlimited attempts to earn the highest grade (out of 10). Each quiz was worth 10 course points (150 in total or 16.7% of the total course grade).

Student measures: *Content Pre/post-test* – During the first course meeting of each semester, students were given a fifteen-question selection from the Geoscience Concept Inventory (GCI; Libarkin & Anderson, 2005) that was identical to the one used at the research institution. These fifteen multiple-choice questions were selected from a larger collection of questions that were validated by many geoscience education researchers over

the course of multiple sequences of development (Libarkin & Anderson, 2005). The questions selected for use in this assessment were chosen due to their relevance to the concepts covered in the course. Considering the Fall 2019 pre-test dataset as an example, reliability analysis of the question set itself produced a Cronbach's Alpha value of .797; exceeding the ≥ .70 indicating acceptable reliability as recommended by Ding and Beichner (2009; see Table 5.2). GCI pre-test scores were used to determine concept knowledge baselines for students and as a continuous control variable for the regression-based components of the study.

Exam Performance – For this analysis, student performance was measured via four summative midterm exams that were distributed in time throughout the course, each assessing five to eight 75-minute lessons of geology content. Each exam varied in length (34-50 questions) and by instructor teaching the course but were consistent within instructor across semesters (see Table 5.2 for a summary of exam features). Each question was dichotomously scored (correct or incorrect) and was either multiple choice or fill in the blank. Additionally, on each exam there were 2-4 short answer questions (3-4 sentences) that are not considered in this analysis. The multiple-choice questions were predominantly from lower-order recognition levels of Bloom's taxonomy (i.e., knowledge and comprehension; Bloom et al., 1956). 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. Cronbach's alpha (α) values for each set of exams were .652 or above, indicating reliability and relative consistency of performance across students on each item of each exam. Total exam information including total number of questions, question type and alpha values is presented in Table 5.2 below.

Table 5.2: Exam features and example reliability for each exam/instructor.

	Measure	Questions	MC	FIB	α
Instructor 1	GCI Pre	15	15	-	.797*
	Exam 1	40	36	4	.805*
	Exam 2	50	50	0	.898*
	Exam 3	40	32	8	.714*
	Exam 4	35	30	5	.684*
Instructor 2	GCI Pre	15	15	-	.775†
	Exam 1	41	41	-	.761†
	Exam 2	46	46	-	.820†
	Exam 3	47	47	-	.768†
	Exam 4	47	47	-	.775†

Note: "MC" = Multiple-choice; "FIB" = Fill in the blank; * = Fall 2019; † = Fall 2018

Exam confidence and local monitoring accuracy – To measure students' monitoring accuracy as they were completing each exam, a five-inch line was placed below each question 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 5.2). 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 5.2) 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). These individual values were averaged across the exam to generate a measure of exam-level accuracy or entered directly into a multilevel model as a dependent variable at the individual judgment level.

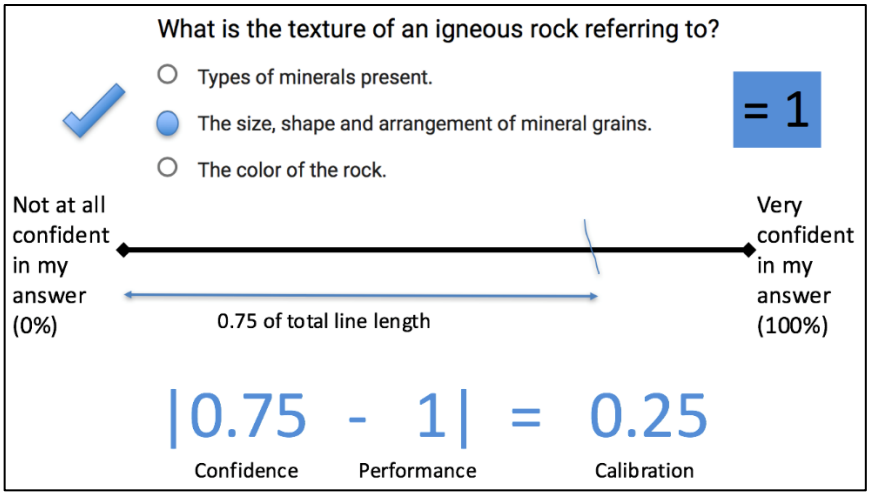

Figure 5.2 Example question and calibration calculation for local accuracy judgments.

CLASS-derived data sources – In addition to serving as a study tool for students, student use of CLASS provides a number of trace data sources that can be used to triangulate exam data and provide insight into student study strategies (e.g., timing, frequency). Additionally, these trace data provide opportunities to glean insight into students' formative assessment behaviors not otherwise collectable via traditional methods. For example, each submitted CLASS quiz attempt records students' responses, global and local judgments and related accuracy variables (calibration and bias) along with a timestamp (m/d/y 00:00) of when each question within the attempt was submitted. When considered in tandem, plots of these trace data variables can be used to determine student usage patterns related to the tool to see how and when students are utilizing the tool in relation to course activities (e.g., exams).

5.3 Results

Quantitative data relating to student performance and local and global accuracy predictions across the study semesters were analyzed using either IBM Statistical Package for the Social Sciences (SPSS; for mean-based analyses) and Statistical Analysis Software (SAS; for multilevel modeling analyses). Data were screened for outliers and normality prior to performing analyses described below and during analyses all assumptions of the procedures were met unless otherwise specified. 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) or as "small" (η_p^2 = .01 - .059), "medium (η_p^2 = .06 - .139) and "large" (η_p^2 > .14) for η_p^2.

1. *How do students from different types of institutions (Research university vs Community College) compare in their baseline performance and judgment accuracy?*

To answer these questions, we analyzed the exam data from five semesters of the RCC setting divided into three groups by academic year (2017-2018, 2018-2019 and Fall 2019). This aggregation was completed to both increase the sample size for comparisons as enrollments for each individual semester is approximately 20 students (Table 5.1). Additionally, each semester within an academic year was taught by the same lead instructor and utilized the same course schedule, instructional elements (e.g., lecture PowerPoints) and situational factors (e.g., time of day, room, etc.). The only differing factor between the two academic years were the lead lecture instructor and the exams used to assess student mastery of course content. Instructor A was the lead instructor for academic year 2017-2018 and Fall 2019, while Instructor B was the lead for 2018-2019 (Table 5.1). The lead instructor taught all lecture meetings and administered equivalent exams of their own authorship at each

iteration. Finally, the Fall 2019 semester featured CLASS more heavily than prior semesters (see below). We will broadly consider trends of each data source (e.g., performance, confidence) before describing a multilevel modeling analysis of calibration accuracy within each setting:

Research university: See Chapter 4 for a full investigation into semester-based fluctuations in exam outcomes at the research university setting. In summary, however, the students in the physical geology course at the research university setting demonstrate accuracy that is highly correlated with their performance (Figure 5.1). Pre/post measures across multiple semesters demonstrated significant learning gains with a large effect size ($d >$ 2) on a standard content test. Across all exams through three semesters, students reported between 71-82% confidence values as a group (Figure 4.10). High performers in the research university course generally slightly underestimate their exam performance while lower performers often overestimate their abilities. Student performance on summative exams in this environment average around 80% and individual exam local calibration averages are traditionally in the .25-.30 range (Figure 5.3; Chapter 4; Jones & McConnell, 2018). Summary statistics from the most recent investigation of exam outcomes and monitoring accuracy in the research university physical geology setting are available in Table 4.4 and 4.5 of Chapter 4 for reference.

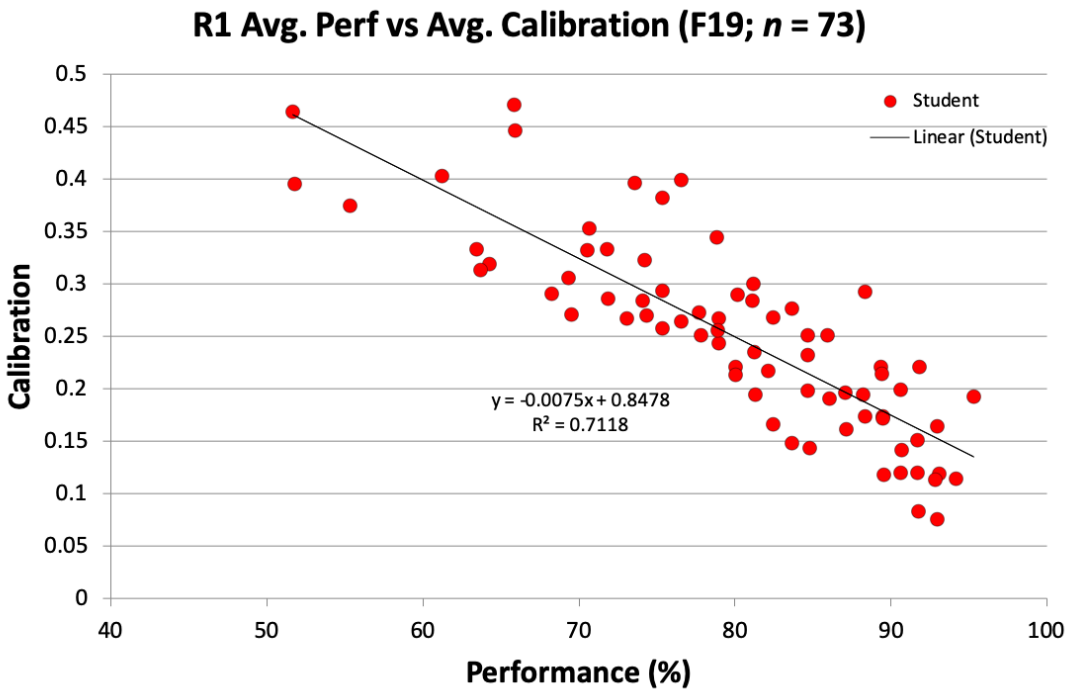

Figure 5.3 Mean performance vs. mean calibration for the research institution Fall 2019 semester

RCC: *Pre/post measures* – To determine if there were any significant differences between the five semesters of 2YC pre-GCI scores (i.e., # of questions correct), one-way ANOVAs were conducted on the measures across semesters and independent-samples t-test for the two post-GCI scores (Fall 2018 and Fall 19 GCI post data was not available). Paired samples t-tests were run for each of the two semesters with post-test data to determine if statistically significant gains were made from pre- to post-test for each iteration of the course. Finally, we calculated a change score by subtracting students matched post-pre score to represent GCI change as a result of taking the course.

There were no significant differences in students GCI pre-test scores ($F(4, 90) = 0.82$, $p = .52$, $\eta_p^2 = 0.035$) or post-test scores between semesters ($F(2, 26) = 2.95$, $p = .07$, $\eta_p^2 = 0.185$). Significant gains with a large effect size were seen in two of three semesters where post measures were present to establish gains; Fall 2017 (mean difference = 1.91, $t(10) =$

2.70, $p = .02$, $d = 1.71$), Spring 2018 (mean difference = 1.89, $t(8) = 3.69$, $p = .01$, $d = 2.61$). Gains were not significant in Spring 2019 ($t(6) = 1.77$, $p = .13$, $d = 1.45$), though sample size was low due to students dropping the course and the course having low attendance the day post measures were administered. There were no significant differences in pre-post change scores (i.e., no semester gained significantly more than any other; Table 5.3).

Table 5.3 GCI pre- and post-test mean (and standard deviations) and gains for each institution.

	RCC			Research Institution		
	Pre-	Post-	Gain†	Pre-	Post-	Gain†
Fall 2017	9.11 (2.75)	11.27 (2.53)	1.91	9.32 (2.57)	11.57 (2.48)	1.94
Spring 2018	7.67 (3.58)	10.0 (2.32)	1.89	-	-	-
Fall 2018	8.60 (3.38)	*	-	9.13 (2.92)	*	-
Spring 2019	7.47 (2.50)	8.71 (1.11)	1.71	-	-	-
Fall 2019	8.16 (3.55)	*	-	9.66 (2.96)	12.13 (2.26)	2.79
Institution mean	8.24 (3.18)	10.17 (2.34)	1.85	9.36 (2.82)	11.86 (2.37)	2.37

Note: † = Gain values from paired-t-test results and matched data. * = Measures unable to be collected/analyzed.

Performance – We analyzed exam performance via either one-way ANOVA or Kruskal-Wallis tests depending on the distributions of each variable. The distribution for Exam 1 was not normal despite square root transformation, so a Kruskal-Wallis test revealed no significant differences in performance between semesters ($X^2 = 0.139$, $p = .94$) with all semesters performing around 70% on average (Table 5.4). The remaining distributions for exam performance were normal so three separate one-way ANOVAS (one per exam) were conducted. Exam 2 analysis indicated that there were no significant differences in scores across the groupings ($F(2, 63) = 2.89$, $p = .06$, $\eta^2 = .084$) with students performing near 60% for each semester grouping (Table 5.4). Exam 3 performance analysis, however, resulted in a significant difference between semesters ($F(2, 54) = 41.67$, $p < .0001$, $\eta^2 = .607$). Bonferroni

corrected post-hoc tests revealed that the student from 2018-2019 scored significantly higher (at 72%) than the both the 2017-2018 group (mean difference = 35.7, $p < .001$) and the Fall 2019 group (mean difference = 23.0, $p < .001$). Additionally, the Fall 2019 group outscored the 2017-2018 group (mean difference = 12.8, $p = .02$)

Finally, there were significant differences between groups on Exam 4 ($F(2, 63) = .161$, $p = .85$, $\eta^2 = .229$) with the 2018-2019 cohort outscoring both the 2017-2018 group (mean difference = 12.7, $p = .011$) and the Fall 2019 group (mean difference = 21.6, $p < .001$). The difference between 2017-2018 and Fall 2019 was not significant ($p = .28$). Raw means and standard deviations relating to student performance from each semester are presented in Table 5.4.

Table 5.4 Performance variable mean and standard deviation across semesters.

	2017-2018 (*n*= 37)		2018-2019 (*n*= 32)		Fall 2019 (*n* = 18)	
	M	SD	M	SD	M	SD
Exam 1	69.26	17.55	70.20	11.44	69.31	13.98
Exam 2	59.67*	21.03	63.11	14.47	50.00	19.24
Exam 3	36.33*	9.23	72.07	14.41	49.12	12.44
Exam 4	56.16	17.92	68.92	12.18	47.76	12.80
Student mean	55.36	16.43	68.57	13.12	54.04	14.61

Note: * *n* = 18 due to missing data

Confidence – We applied the same approaches to analysis for students' line-item confidence measures. Similar to performance analysis, Exam 1 confidence was not normally distributed despite a square-root transformation. A Kruskal-Wallis test revealed no significant differences in performance between semesters ($X^2 = 4.09$, $p = .13$) with all semesters reporting between 57-70% confidence as a group (Table 5.5; Figure 5.4). Exam 2 analysis indicated that there were no significant differences in exam confidence across the groupings ($F(2, 63) = 1.58$, $p = .22$, $\eta^2 = .049$) with groups of students reporting an

aggregate of between 50-60% confidence across exam questions (Table 5.5). Exam 3 was similar with no significant differences between semesters ($F(2, 50) = 2.27$, $p = .12$, $\eta^2 = .115$). Students reported between 53-68% confidence on the third exam (Table 5.5). Finally, there were no significant differences between groups on Exam 4 ($F(2, 54) = 0.41$, $p = .67$, $\eta^2 = .015$) with students reporting between 56-64% confidence. Raw means and standard deviations relating to student performance from each semester are presented in Table 5.5.

Table 5.5 Confidence variable mean and standard deviation across semesters.

	2017-2018 (*n*= 37)		2018-2019 (*n*= 32)		Fall 2019 (*n* = 18)	
	M	SD	M	SD	M	SD
Exam 1	66.10	14.39	70.23	13.82	57.34	25.21
Exam 2	61.02*	20.15	61.93	21.47	50.17	23.28
Exam 3	64.06*	18.46	67.60	17.40	52.49	27.29
Exam 4	58.99	18.36	63.57	24.90	56.10	29.59
Student mean	62.54	17.84	65.83	19.40	54.02	26.34

Note: * *n* = 18 due to missing data

Performance vs confidence – Generally, while average performance and confidence demonstrated during exams was low, they were similar in magnitude. Mean confidence did not fluctuate between exams as measured by local item judgments. Additionally, these values similar to mean performance on the same exam aside from one significant deviation (overconfidence during Exam 3 of 2017-2018; Figure 5.4). Even during the Fall 2019 semester, where performance was significantly lower than prior semesters, student confidence was similar. For example, the highest bias between confidence and performance was demonstrated during the first exam where students were 11% underconfident in their performance. This signal of agreement between confidence and performance suggests some level of accuracy in perception.

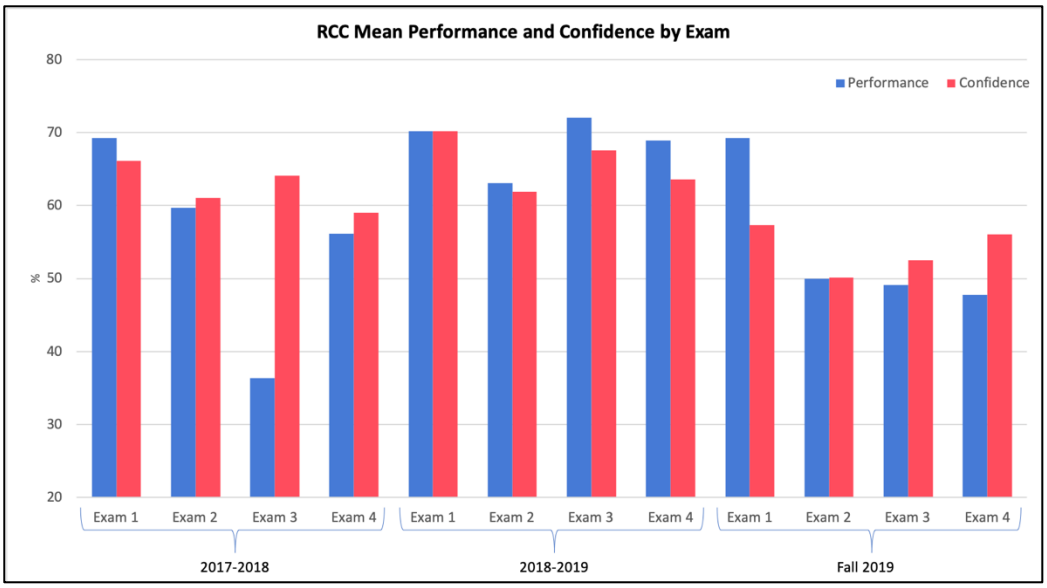

Figure 5.4 Mean exam performance and item-level confidence by RCC course exam.

Performance vs. calibration – As with Figure 5.1 above, one way to visualize the relationship between a student's metacognitive monitoring accuracy and their performance is to generate a simple bivariate regression of the two variables. As several students were missing data from the cohort, we generated the plots below from raw values of students' mean performance and calibration accuracy for students with qualifying data from at least three of the four course exams. As the research university semester consists of three exams, this allows for a similar comparison between the two institutions. Consider the three groupings of the RCC sample and the baseline research university dataset in Figure 5.5. While some subsets have similar slopes to the research university baseline (e.g., RCC 2018-2019), others are much less steep, with higher performing students displaying higher calibration values than would be predicted from the research university sample (e.g., RCC 2017-2018, Fall 2019). These distributions of performance and local monitoring accuracy highlight the wide variability expressed within the dataset.

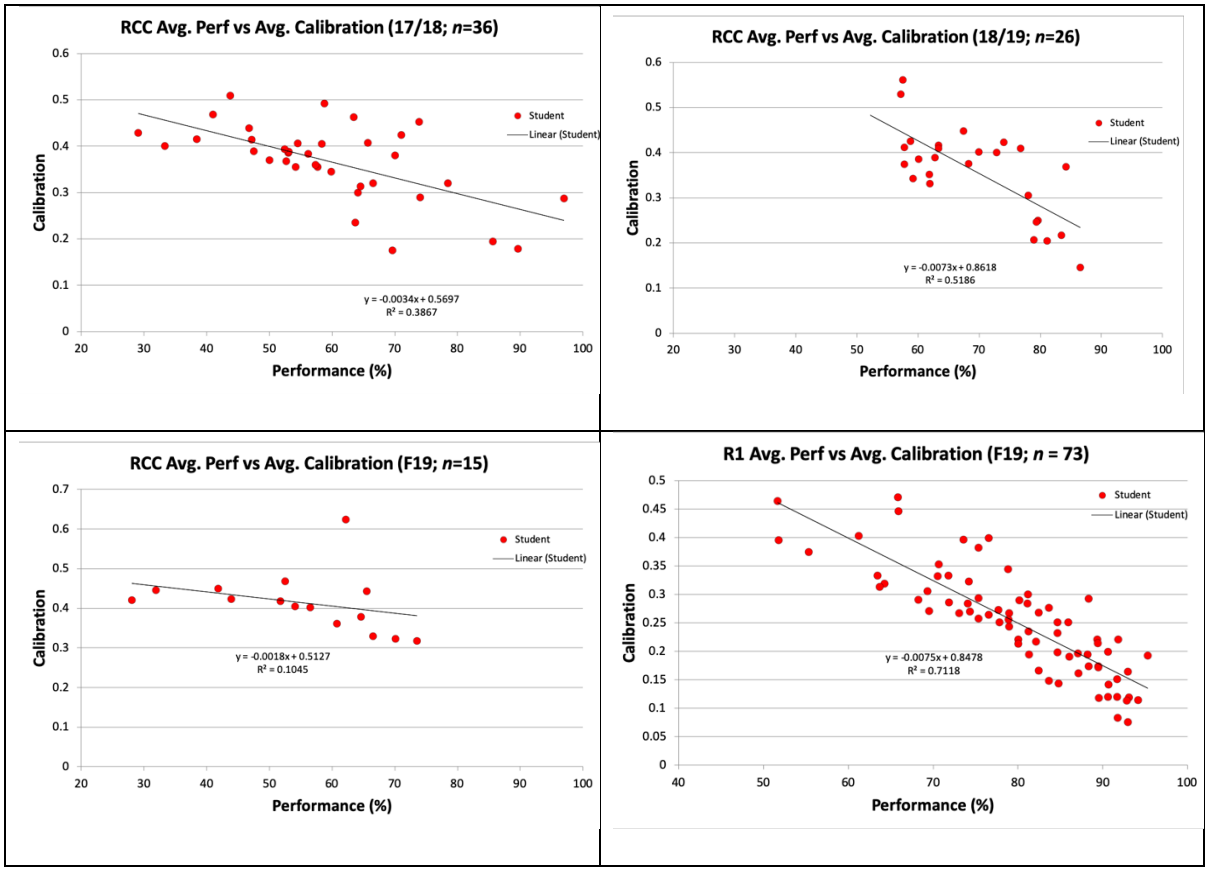

Figure 5.5 Mean performance vs mean calibration for all RCC subsets and for research university Fall 2019 semester.

Multilevel modeling of calibration accuracy – A procedure for the compiling and collation of student calibration data was followed rejecting participants with more than a fifteen-percent non-response rate to prepare inferential measures of local accuracy (calibration) using mean-based analysis (e.g., ANOVA). Due to the requirement for paired data in repeated measures designs, students who failed to complete a sufficient number of confidence judgments for questions on each exam were to be rejected from the larger sample population due to their incomplete datasets. This led to the possibility of significant attrition for local accuracy analysis in the already limited 2YC dataset. Consequently, unlike the research university data, we did not have sufficient complete sets of exam confidence line items from a large enough population of students to obtain an accurate analysis. To help remedy this phenomenon, multilevel modeling was used to analyze exam calibration data as

multilevel models can be estimated when data are partially missing (Curran et al. 2010). We used multilevel modeling to analyze the individual judgments made by students at large within the setting (i.e., aggregated between semesters) to begin to characterize potential differences in demonstrated accuracy between each institutional setting.

In the multilevel modeling framework, individual change/variability is represented through a 2-level hierarchical model (Curran et al. 2010). At Level 1, each person's variability (e.g., fluctuation in calibration values measured from multiple judgments) is represented by an intercept and slope that become the outcome variables in a Level 2 model in which they may depend on person-level characteristics (e.g., pre-test score, gender; Nezlek, 2012). As estimates of both between-person effects and within-person variability are possible with multilevel models (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999), inferences can be made towards the partitioning of variability between people (Student A's judgment accuracy vs Student B's) and within people (how Student A's judgments fluctuate) across observations. This provides an opportunity to mitigate attrition due to missing data related to students dropping the course, missing exams, or neglecting some questions on an exam. As a result, we utilize multilevel modeling to investigate the research questions for this work as multiple judgments of learning that are used to generate calibration values (Level 1) are nested within students (Level 2) who participated in the study. Additionally, these observations can be considered by exam (a Level 1 variable serving as a time proxy) and control for individual variability in prior knowledge (a Level 2 covariate). All of these features of the dataset suggest multilevel modeling as the most prudent means of analysis (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999).

Similar to a regression analysis, the Greek letters in multilevel modeling results represent the values of coefficients related to the target variables (e.g., the intercept and slope). In Level 1, the intercept, β_{0it}, is defined as the expected level of calibration for person *i*. The calibration slope, β_1, is the expected change in calibration associated with subsequent judgments. The error term, r_{it}, represents a unique effect associated with person *i* (i.e., how much that individual fluctuates or varies over time). The individual intercepts (β_{0i}) and slopes (β_1) become the outcome variables in the Level 2 equations, where the average calibration level for the sample at baseline (i.e., when exam/time = 0) is represented by γ_{00} and the average change over time for the sample is represented by γ_{10}. The extent to which people vary from the sample average of calibration is represented by u_{0i}. In building the model, we followed the precedent of prior calibration researchers during their multilevel modeling of measures of calibration (e.g., Hadwin & Webster, 2013; Foster, Was & Dunlosky et al., 2017) and followed a similar line of decision making (e.g., model selection, model fit decisions) as their prior work modeling calibration accuracy.

RCC calibration model – To begin a multilevel modeling investigation it is recommended to conduct a preliminary analysis to ensure that there is sufficient variability at Level 1 and Level 2 to allow a continuation of analyses (e.g., Nezlek, 2012; Raudenbush & Bryk, 2002). This preliminary analysis is generally referred to a fully unconditional model (or null model), in which no term other than the intercept is included at any level (Nezlek, 2012). Results from this analysis with RCC calibration as the dependent variable indicated that students global average calibration predicted by the model was 0.36, which is significantly greater than zero ($t(232) = 41.63$, $p < .0001$) and the ICC equaled .085. This indicated that 91.5% of the variability in calibration accuracy was within people ($\tau_{00} = 0.01$, *z*

= 5.94, $p < .0001$) and 8.5% was between people ($\sigma^2 = 0.08$, $z = 71.83$, $p < .0001$). Therefore, the fully unconditional model indicated that there was sufficient variability at each level to warrant further analyses.

After conducting the fully unconditional model to determine distribution of variability between- and within-students, the second step was to add a variable at Level 1 to determine the fluctuation of calibration accuracy across the study period using exam number as a proxy for time across the semester. The results indicated that there were significant differences in the calibration accuracy over time ($\gamma_{10} = .0019$, $t = 3.07$, $p = .002$) with students on average making less accurate judgments (higher calibration values) with each subsequent exam. This model accounted for 0.63% of the within-person variability in calibration.

To control for individual variability in incoming geoscience knowledge, each individual's centered GCI pre-score was included as a Level 2 predictor. The slopes in Level 2 were constrained in the model as this generated a better model fit. The results indicated that students with higher GCI pre-scores were more accurate than students with lower scores as time progressed ($\gamma_{01} = -.011$, $t = 3.42$, $p = .0013$). This model accounted for 4.99% of the between-person variability and 0.76% of the within-person variability in calibration accuracy. Finally, considering the results of all three models together, after controlling for GCI pre-scores at time/exam zero, the model-predicted calibration value for the setting is 0.32. Theoretically, this signals on average 32.1% inaccuracy between students' judgments and their demonstrated performance on test items in the RCC setting. A summary of each model of community college student exam calibration values and associated statistics (e.g., fixed effects, random effects, and maximum-likelihood fit statistics, etc.) is provided in Table 5.6.

Table 5.6 Unstandardized coefficients (and standard errors) of multilevel models of calibration accuracy from RCC setting.

Fixed Effects		Model 1	Model 2	Model 3
Calibration, β_0				
	Intercept, γ_{00}	0.36 (.01)***	0.32 (.02)***	0.32 (.01)***
Calibration Slope, β_1				
	Exam, γ_{10}	-	0.02 (.01)***	0.02 (.00)**
	GCI Pre, γ_{01}	-	-	-0.01 (.00)**
Random Effects				
L1 Variance (τ_{00})		0.01(.00)***	0.01(.00)***	0.01(.00)***
Within-person fluctuation (σ^2)		0.08 (.00)***	0.08(.00)***	0.08(.00)***
ML Fit statistics:	-2LL	4058.2	4006.8	3878.2
	AIC	4062.2	4010.8	3888.2
	BIC	4076.4	4016.0	3887.3
Note: * $p < .05$, ** $p < .01$, *** $p < .001$				

Research university calibration model – To compare exam judgment accuracy between the two institutions, a multilevel model was generated using the same variables and methods as the RCC model described above. Results from this analysis with research university calibration as the dependent variable indicated that students global average calibration predicted by the model was 0.28, which is significantly greater than zero ($t(232) = 41.63$, $p < .0001$) and the ICC equaled .15. This indicated that 85% of the variability in calibration accuracy within the sample was within people ($\tau_{00} = 0.01$, $z = 9.58$, $p < .0001$) and 15% was between people ($\sigma^2 = 0.001$, $z = 93.18$, $p < .0001$). Therefore, the fully unconditional model indicated that there was sufficient variability at each level to warrant further analyses.

After conducting the fully unconditional model to determine distribution of variability between- and within-students, the second step was to add a variable at Level 1 to determine

the fluctuation of calibration accuracy across the study period using exam number as a proxy for time across the semester. The results indicated that there were significant differences in the calibration values over time (γ_{01} = .0015, t = 6.05, p < .0001) with (similar to the RCC setting) students on average making less accurate judgments with each subsequent exam. This model accounted for 0.22% of the within-person variability in calibration.

To control for individual variability in research university students' incoming geoscience knowledge, each individual's centered GCI pre-score was included as a Level 2 predictor. The results indicated that students with higher GCI pre-scores were more accurate than students with lower scores (γ_{01} = -.018, t = 6.22, p < .0001). This model accounted for 17.21% of the between-person variability and 0.79% of the within-person variability in calibration accuracy at the research university institution. Finally, considering the results of all three models together, after controlling for GCI pre-score and time/exam through the semester, the model-predicted calibration value for the research university setting is 0.24. Theoretically, this represents on average a 24.3% absolute inaccuracy between students' judgments and their demonstrated performance on test items in the research university setting. A summary of each model related to the research institution setting and associated statistics are provided in Table 5.7.

Table 5.7 Unstandardized coefficients (and standard errors) of multilevel models of calibration accuracy from the research university setting.

Fixed Effects		Model 1	Model 2	Model 3
Calibration, β_0				
	Intercept, γ_{00}	0.28 (.01)***	0.25 (.01)***	0.24 (.01)***
Calibration Slope, β_1				
	Exam, γ_{10}	-	0.02 (.01)***	0.01 (.01)**
	GCI Pre, γ_{01}	-	-	-0.02 (.01)**
Random Effects				

Table 5.7 Continued

L1 Variance (τ00)		0.01(.00)***	0.01(.00)***	0.01(.00)***
Within-person fluctuation (σ^2)		0.08 (.00)***	0.08(.00)***	0.08(.00)***
ML Fit statistics:	-2LL	5563.6	5537.1	5291.7
	AIC	5567.6	5541.1	5295.7
	BIC	5574.5	5548.0	5302.6

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

2. *What are the differential effects of providing CLASS quizzes and feedback to students at these institutions? Specifically, …*
    a. *…student use variations across semesters between institutions*
    b. *…the effects of providing increased course structure and CLASS on students' exam outcomes in the community college setting*

To investigate these questions, we considered a combination of trace data collected from CLASS and a subset of the RCC exam dataset from the cohort that was required to interact with CLASS (Fall 2019).

Usage frequency: After screening the CLASS trace data for consented individuals, we considered usage across the study period. CLASS was provided to students in the RCC setting as optional in Fall 2018/Spring 2019 and as required in Fall 2019 (three semesters). Usage frequency in the Fall 2018 semester was negligible, with 2 consenting students making only 21 attempts prior to each course exam. Despite being introduced to the course in the same manner (e.g. via a short presentation and the administering of pre-measures/consent), students the subsequent semester (Spring 2019) attempted even fewer attempts; with one student completing 11 attempts throughout the entire semester. Table 5.8 reports these frequencies both in aggregate and delineated by the individual students generating the attempts ("Student 1" and "Student 2" for Fall 2018 and "Student 1" for Spring 2018).

Beginning in the Fall 2019 semester, students' CLASS use was mandatory in the same manner as was implemented in the research university setting, with students being allowed unlimited randomized attempts to earn the highest grade possible. This requirement increased the frequency of quizzing significantly (by a factor of 10, on average). This increase in frequency, when averaged across users, was similar in magnitude to the average student usage leading to some exams during semesters at the research university when quizzing was optional (e.g., 58.46 questions per student for RCC Exam 4 Fall 2019 vs 58 questions per student for research university Exam 3 Fall 2017). While dynamic quizzing was available to students in the Fall 2019 semester (see Chapter 3), they were not utilized.

Table 5.8 Aggregate CLASS quiz question attempts across semesters by exam.

	Fall 2018 (*n*= 2)		Spring 2018 (*n*= 1)		Fall 2019 (*n*=13)		
	Assigned	S1/S2	Assigned	S1	Assigned	Average/ Student	Dynamic (*n*=0)
Exam 1	120	10/110	0	0	1170	90	0
Exam 2	30	30/0	0	0	610	46.9	0
Exam 3	60	60/0	70	70	0	0	0
Exam 4	0	0/0	40	40	760	58.46	0
Total	210	-	110	-	2540	65.12	0

Note: S1 = "Student 1"; S2 = "Student 2"

Temporal usage patterns: To visualize students' temporal usage of CLASS quizzes throughout the target semesters, we generated scatter plots plotting the date stamp collected by CLASS from each individual question attempt submitted for the Fall 2019 assigned quizzes against linear time across the semester (as usage was too infrequent during the optional academic year). Each dot represents one question and its associated calibration accuracy value as recorded via submission of the question (Figure 5.6; compare with Figures 4.6, 4.7 and 4.8). The change in requirements pertaining to CLASS quizzes caused more students to use the tool (Table 5.8). Results demonstrate that course requirements directly

have an effect on the timing (and completion) of student learning tasks. Students' mastery quizzing on CLASS during the semester was not clustered around impending exams.

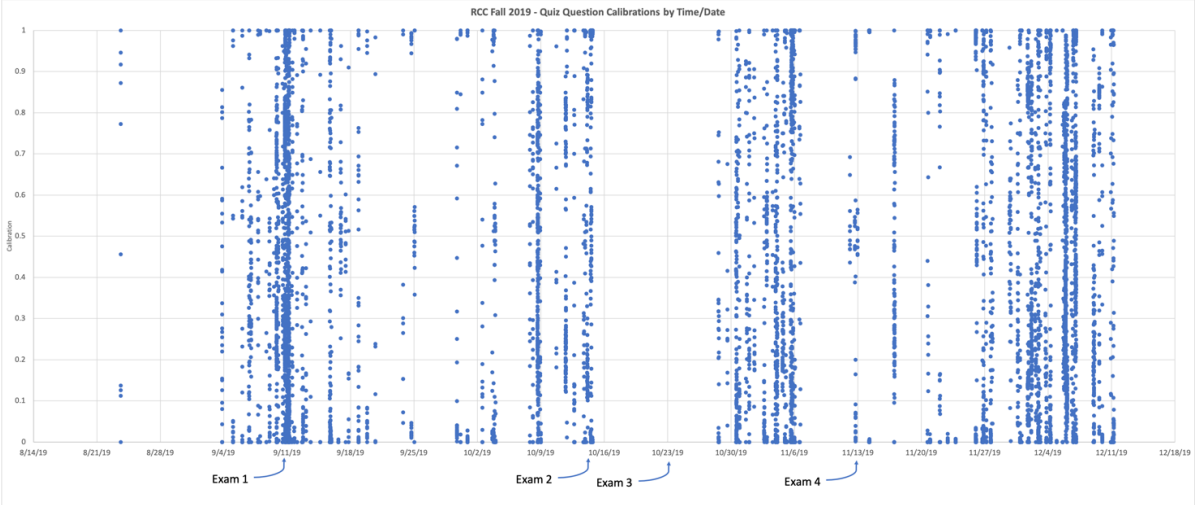

Figure 5.6 Distribution of RCC CLASS quiz attempts during the Fall 2019 semester.

Changes in exam outcomes: As outlined above, effects related to the introduction of CLASS to the target course is at first a null result due to the low frequency of student use during the first two semesters it was implemented. As a result, we focused our analyses upon Fall 2019 (where usage was required as part of the course structure) to investigate potential effects. Throughout the five semesters of the study, 17 students consented to participate and used CLASS across the study period (17.5% of the students sampled), with 13 of the users enrolled during the Fall 2019 semester.

There are no discernable patterns between student use of CLASS and their exam outcomes during Fall 2019 (Figure 5.7). The Fall 2019 cohort was consistently among the lowest performing and least confident groups in the study (Figure 5.4). Figure 5.7 reports the frequency of CLASS quiz attempts (not total questions as in Table 5.7), mean performance, and mean line-item confidence of each student in the Fall 2019 cohort. Considering students in the figure, some with high attempts returned low exam performance and confidence (e.g.,

Student 3; Figure 5.7) and others who have modest attempts were some of the better performers on average. As sample sizes were too low to generate meaningful inferential statistics, these inconsistent patterns present a need for further characterization of the user experience and their interaction (or non-interaction) with feedback provided by CLASS.

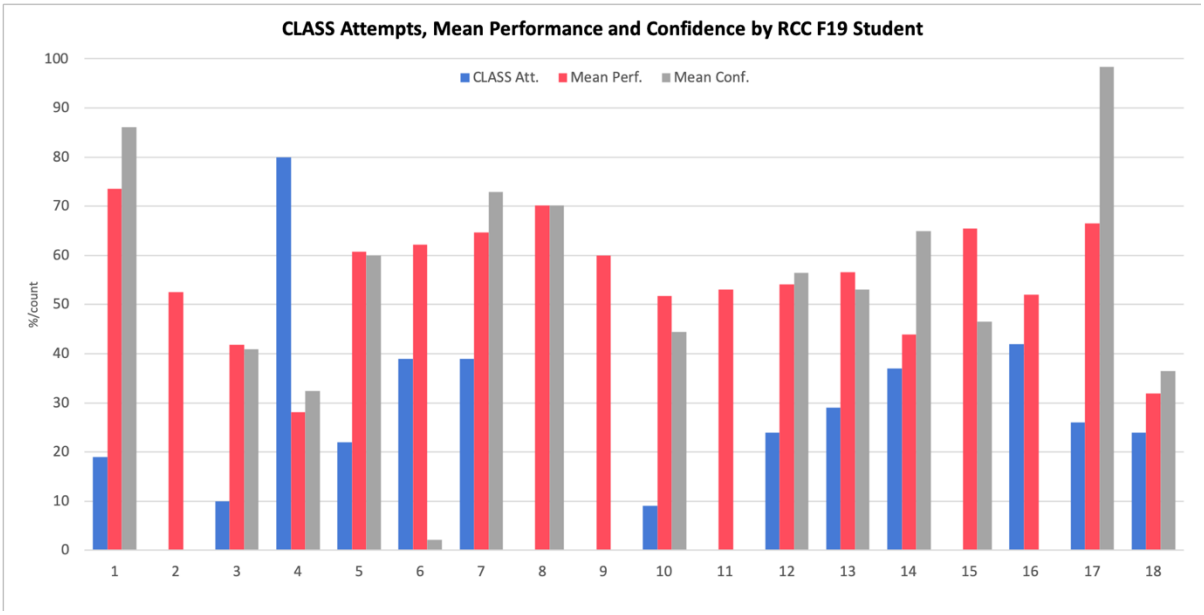

Figure 5.7 CLASS attempts, mean performance, and confidence by RCC students during the Fall 2019 semester.

5.4 Discussion

Major findings: This analysis sought to begin to characterize potential differences between introductory geology student metacognitive awareness and exam outcomes in community college and research institutions.

1. *How do students from different types of institutions (Research university vs Community College) compare in their baseline performance and judgment accuracy?*

Comparing the two samples collected from each course at each institution, there are several different phenomena to note. The RCC students were lower-performing on average, less confident and exhibited higher (less accurate) calibrations when compared to their

research university counterparts. Within this trend however, there were some students who performed well who were less calibrated and some students who were more calibrated than their performance would predict. This baseline characterization suggests that the consideration of local accuracy within this setting requires more data sources than simply performance and confidence judgments. Though prior geoscience knowledge as measured by the GCI was shown to influence accuracy through the semester, data sources collecting information geared towards more SRL-related variables such as reported use of (low-level) study strategies, time spent studying, school/life balance and self-efficacy can potentially further this understanding of the metacognition and SRL of these learners.

2. *What are the differential effects of providing CLASS quizzes and feedback to students at these institutions? Specifically,…*
    a. *…student use variations across semesters between institutions*
    b. *…the effects of providing increased course structure and CLASS on students' exam outcomes in community college setting*

The results of the frequency and distribution analysis support the findings of Chapter 4 in that by requiring students to interact with course support via a modest amount of course credit throughout the semester (e.g. 10 points/quiz or ~15% of the total course grade for the RCC), instructors can guide students towards effective learning behaviors (e.g., practice testing). What was unique, however, was the significant impact of this requirement to students in the setting. When CLASS quizzing was optional and proposed as a learning tool to help them succeed in the course, adoption was near zero. Once required, mean values of quizzing neared a frequency and distribution similar to that of the research university setting. In other words, within this sample in particular, "optional" appeared to translate as "not

necessary" to the students. Collecting data pertaining to student motivations for learning in the setting and the types of learning strategies they feel help them to find success is necessary to isolate the factors contributing to this non-use.

The addition of another element of structure to the course (CLASS quizzes), did not to have a significant effect upon student performance outcomes as the final Fall 2019 semester of the study demonstrated the lowest mean performance within the sample. This poses salient questions for the support of metacognition and effective SRL behaviors in this setting in the future. While similar endeavors in biology 2YC environments have shown increases in exam performance, Pape-Lindstrom et al. (2018) suggested the potential of a conditional "dosage effect" for online quizzing to impact student outcomes. They state that while increasing the support for students in the 2YC setting may improve performance outcomes in some settings, they purport that these increases are the effects of moderation by active learning and evidence-based teaching (Pape-Lindstrom et al., 2018) This hypothesis has been supported by other work in biology and elsewhere (Freeman et al., 2011; Connell et al., 2016; Elliot et al., 2016). In the words of Pape-Lindstrom et al. (2018 p. 6), "dosage may matter, *if* instructors are using evidence-based prescriptions." As the learning environment sampled for this work was very much instructor-centered, the lack of an effect of CLASS-provided practice testing and metacognitive feedback may not have had the chance to take root. More work in the 2YC setting across multiple courses with varying instructional approaches is needed, however, to further investigate this hypothesis.

Limitations/future work: In light of these mixed results, there are several limitations to consider. While efforts were made to control for as many confounding variables as was possible, collecting real-world classroom data is always subject to external influences outside

of the researcher's control. For this population in particular (i.e., 2YC), the student population in this environment is often more diverse in background, academic experience, and outside of school complexities than the traditional research university student (Ma & Baum, 2016). As a result, it is important to measure more of these variables to be able to triangulate results and recognize significant trends. There is a need to know more about the participants via increased collection of student data (experiences, demographics, motivational instruments, etc.) and better documentation of the instructional setting. For example, while cursory observations were made of each classroom and the style of instruction therein, there is a need to empirically characterize instruction via observational instruments (e.g., RTOP; Sawada et al., 2002) to better investigate effects of in-classroom influences on performance and metacognitive accuracy outcomes.

Additionally, while attempts were made to measure and analyze equivalence between them, each of the lead instructors did not use the same exams and these two sets of exams were different than the exams used to measure learning at the research university institution. While the exams assessed the same units of geology content, were similar in format, and had similar measures of reliability, students during Instructor B's academic year scored higher than Instructor A's. The potential for differential difficulty of the geology content and relative complexity of the material is present and can affect metacognitive monitoring outcomes (Schraw & Roedel, 1994; Pressley & Ghatala, 1990). Future work will seek to narrow this uncertainty by further characterizing the exams in target courses through processes such as the rating of the complexity of exam questions on Bloom's Taxonomy scale to control for potential exam-level variations in difficulty (e.g., Freeman et al., 2011). With that, however, given anecdotal familiarity with the exams gained during data processing

(e.g., line measuring) the complexity of the RCC exams was generally lower than that of the research university exams (e.g., more knowledge-level questions). Also, analyzing the results not only by exam/temporal occurrence, but by the specific geologic topics (e.g., # of questions about plate tectonics, etc.) covered between the two settings could also help isolate disparities in topic-based performance, confidence or monitoring accuracy.

In order to answer important questions regarding differential effects of interventions on diverse populations present in the 2YC environment, an inherent structural limitation in the setting is small sample sizes. As enrollment in these courses are often capped around 20-25 student per section, there is an inherent challenge in sampling enough students and having enough treatment sections to allow for the required statistical power to answer questions related to student demographics, etc. Additionally, as attrition (either via dropping the course or students not providing data) was high, this study was doubly affected. In future work, we will seek to widen data collection to settings with multiple sections and look to build in incentives for students to provide necessary accuracy judgments for analysis (e.g., Chapter 4).

Conclusion: We sought to investigate the relationship between metacognitive monitoring and performance within students learning geoscience in a 2YC environment and to support students via a web-based tool. While there were challenges with this pilot program (e.g., low buy-in, attrition, working with instructors unfamiliar with research techniques), this pilot investigation provided some insight into the setting. Students in the sampled class varied widely in their performance and judgment behavior, but were generally lower performing, less confident, and less accurate than students sampled in the research university environment. They were reluctant to utilize the support provided by the tool yet responded to

alterations in requirements surrounding the tool. Their exam outcomes, however, were not predictably altered by this increase in use. Finally, multilevel modeling analysis additionally revealed that they were generally less accurate than students in the research university setting.

As CLASS has been shown to have utility to the students who use it frequently (Chapter 3), we seek to "widen the net" of data collection to more 2YC institutions, with more instructors and more robust data sources regarding both the student experience *before* (i.e. academic history), *during* (i.e., classroom observations) and *after* instruction (i.e., CLASS). This is to better characterize best practices for supporting students' metacognition and SRL behaviors. Ultimately, while valuable, this experience serves as a lesson learned and one to build upon for future investigations.

The 2YC setting provides more diverse range of individuals, with more diverse backgrounds than traditional research settings and is an important frontier for both future discipline-based education research in general (Pape-Lindstrom et al., 2018) and for the geosciences as a community specifically (Wilson, 2017b). As a result, it is our suggestion that any intervention related to fostering metacognitive abilities and/or performance in the 2YC should be provided with enough structural support (i.e., requirements, directed interaction with feedback, instructor support) to ensure adoption and "buy-in" of behaviors and enough measurement of within-person variables to allow for the investigation of success.

CHAPTER 6: CONCLUSION

This work detailed the development and implementation of the Confidence-based Learning Accuracy Support System (CLASS) over a three-year project in introductory college geology courses at two separate institutions. We created CLASS to automate the collection of student confidence data and the generation of accuracy feedback during the online practice quizzing process. This feature allows for immediate feedback to be provided to students on both individual assessments and globally across all quizzes that the student completes for a course. We designed feedback provided by CLASS to be used by students to hone metacognitive monitoring accuracy and/or to inform self-regulated learning behaviors. For instructors, student data from CLASS can be used to know more about students' progress and to provide evidence-based interventions that are directly related to gaps in students' mastery of course concepts as needed.

After supplying a review of metacognition and SRL literature for the geoscience education community (Chapter 2), we detailed the features and empirical influence on design elements of CLASS before gathering student voices regarding their perceptions of the tool and the feedback it provides (Chapter 3). Next, we investigated the effects of altering CLASS and the requirements surrounding its use in the research institution (Chapter 4), including a gamified element to students' post-exam judgments. Results were mixed, yet it was found that students at the research institution significantly increased the frequency and distribution of their practice testing behaviors as a result of the course alterations. Performance, confidence, and accuracy during the study trended upward during the study period, but not to statistical significance.

Finally, the final study (Chapter 5) sought to characterize and improve student performance, confidence, and metacognitive monitoring accuracy related to learning in a geoscience course at a regional community college. This was attempted by providing students CLASS quizzes in both optional and required formats. While there were difficulties with attrition, results indicated a wide variability in student outcomes over the study period and with (on average) lower performance than the research institution setting and students having little interaction with CLASS. Additionally, multi-level models comparing the two settings revealed increased inaccuracy within the 2YC setting over the study period. These results suggest that just as the two-year college is a more-nuanced environment and one that requires future investigations to collect as much information regarding the student experience outside of exam and quiz results as possible to isolate potential factors that lead to success. Regardless, this work further suggests the potential for CLASS as a powerful tool for instructors to support student learning in only STEM-wide, but education-wide applications.

REFERENCES

Adams, W. K., & Wieman, C. E. (2015). Analyzing the many skills involved in solving complex physics problems. *American Journal of Physics, 83*(5), 459–467. https://doi.org/10.1119/1.4913923

Alexander., P.A., (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction, v.24*, p.1-3.

Azevedo, R., (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist, 40*, 199-209.

Azevedo, R. (2009). Theoretical, methodological, and analytical challenges in the research on metacognition and self-regulation: A commentary. *Metacognition & Learning, 4,*87-95.

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, NJ: Prentice-Hall.

Bannert, M., & Reimann, P. (2012). Supporting self-regulated hypermedia learning through prompts. *Instructional Science*, 40(1), 193–211.

Bellhäuser, H., Lösch, T., Winter, C., & Schmitz, B. (2016). Applying a web-based training to foster self-regulated learning — Effects of an intervention for large numbers of participants. *Internet and Higher Education, 31*, 87–100. https://doi.org/10.1016/j.iheduc.2016.07.002

Bielaczyc, K., Pirolli, P., & Brown, A.L., (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. Cognition and instruction, v.13, #2, p.221-252.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *Annual Review of Psychology, 64*(1), 417–444.

Bloom, B. S., Krathwohl, D. R., and Masia, B. B. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals*. New York, NY: D. McKay.

Boekaerts, M. (2011). "Emotions, emotion regulation, and self-regulation of learning," in *Handbook of Self-Regulation of Learning and Performance*, eds B. J. Zimmerman and D. H. Schunk (New York, NY: Routledge), 408–425. Charmaz, K. and Bryant, A. (2007). The SAGE handbook of grounded theory. London: SAGE.

Bol, L., & Hacker, D. J. (2012). Calibration research: Where do we go from here? Frontiers in Psychology, *v.3(JUL)*, p.1–6.

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, v.73, #4,* p.269–290.

Bråten, I., & Samuelstuen, M. S. (2007). Measuring strategic processing: comparing task-specific self-reports to traces. *Metacognition and Learning, 2*(1), 1–20.

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning, 11*(2), 215–235. https://doi.org/10.1007/s11409-015-9142-6 Costa Ferreira et al., 2016

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Connell, G. L., Donovan, D. A., & Chambers, T. G. (2016). Increasing the use of student-centered pedagogies from moderate to high improves student learning and attitudes

about biology. CBE—Life Sciences Education, 15(1), ar3. doi: 10.1187/cbe.15-03-0062

Crouch, C.H. & Mazur, E. (2001). Peer instruction: ten years of experience and results. *Am J Phys, v.69,* p.970–977.

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognitive Development, 11*,121–136.

Dancy, M., and Henderson, C., (2010). Pedagogical practices and instructional change of physics faculty. American Journal of Physics, v.78(10), p.1056–1063. doi:10.1119/1.3446763

De Backer, L., Van Keer, H., & Valcke, M. (2012). Exploring the potential impact of reciprocal peer tutoring on higher education students' metacognitive knowledge and regulation. *Instructional Science, 40*(3), 559–588. https://doi.org/10.1007/s11251-011-9190-5.

Derting, T.L., & Ebert-May, D., (2010). Learner-centered inquiry in undergraduate biology: Positive relationships with long-term student achievement. CBE-Life Sciences Education, v.9, p.462-472.

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. Psychological Science, 29(5), 761-778.

Deterding, S., Khaled, R., Nacke, L. E., & Dixon, D. (2011). Gamification: Toward a definition. Conference Paper. Vancouver, BC. Retrieved: http://gamification-research.org/wp-content/ uploads/2011/04/02-Deterding-Khaled-Nacke-Dixon.pdf

Devolder, A., van Braak, J., & Tondeur, J. (2012). Supporting self-regulated learning in computer-based learning environments: systematic review of effects of scaffolding in

the domain of science education. *Journal of Computer Assisted Learning*, *28*(6), 557–573. https://doi.org/10.1111/j.1365-2729.2011.00476.

Dicheva, D., Dichev, C., Agre, G., & Angelova, G. (2015). Gameification in Education: A Systematic Mapping Study. *Educational Technology & Society*, *18*(3), 75–88.

Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning, 3*(3), 231–264. https://doi.org/10.1007/s11409-008-9029-

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. Physical Review Special Topics - Physics Education Research, 5(2), 1–17. doi:10.1103/ PhysRevSTPER.5.020103

Dinsmore, D. L., Alexander, P. a., & Loughlin, S. M. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review, 20*(4), 391–409. https://doi.org/10.1007/s10648-008-9083-6

Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, v.24*, p.4-14.

Dörrenbächer, L., & Perels, F. (2016). More is more? Evaluation of interventions to foster self-regulated learning in college. *International Journal of Educational Research, 78*, 50–65. https://doi.org/10.1016/j.ijer.2016.05.010

Dixon, W. J., & Massey Jr., F. J. (1969). *Introduction to statistical analysis (3rd ed.)*. New York: McGraw-Hill Book Co.

Duffy, M. C., & Azevedo, R. (2015). Motivation matters: Interactions between achievement goals and agent scaffolding for self-regulated learning within an intelligent tutoring system. *Computers in Human Behavior, 52*, 338–348. https://doi.org/10.1016/j.chb.2015.05.041

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition.* Beverly Hills, CA: SAGE.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self-evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271–280. http://dx.doi.org/10.1016/j.learninstruc.2011.08.003.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques. *Psychological Science in the Public Interest*, *v.14*, *#1*, p.4–58.

Dunlosky, J., & Tauber, S. K. (2014). Understanding People's Metacognitive Judgments: An Isomechanism Framework and Its Implications for Applied and Theoretical Research. In *The SAGE Handbook of Applied Memory* (pp. 444–464). 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd. https://doi.org/10.4135/9781446294703.n25

Eagan, M. K., Stolzenberg, E. B., Berdan Lozano, J., Aragon, M. C., Suchard, M. R. & Hurtado, S. (2014). Undergraduate teaching faculty: The 2013–2014 HERI Faculty Survey. Los Angeles: Higher Education Research Institute, UCLA.

Eddy, S.L., & Hogan, K.A., (2014). Getting under the hood: How and for whom does increasing course structure work? CBE-Life Science Education, v.13, p.453-468.

Egger, A.E., Viskupic, K., and Iverson, E.R. (2019). Results of the National Geoscience Faculty Survey (2004-2016). National Association of Geoscience Teachers, Northfield, MN. 82 p.

Eisenkraemer, R. E., Jaeger, A., & Stein, L. M. (2013). A systematic review of the testing effect in learning. *Paidéia, 23*(56), 397–406.

Elliott, E. R., Reason, R. D., Coffman, C. R., Gangloff, E. J., Raker, J. R., Powell-Coffman, J. A., & Olgilvie, C. A. (2016). Improved student learn- ing through a faculty learning community: How faculty collaboration transformed a large-enrollment course from lecture to student- centered. CBE—Life Sciences Education, 15(2), ar22. doi: 10.1187/ cbe.14-07-0112

Ericsson, K. A. 2006. "Protocol Analysis and Expert Thought: Concurrent Verbalizations of Thinking During Experts' Performance on Representative Tasks." In *The Cambridge Handbook of Expertise and Expert Performance*, edited by K. A. Ericsson, N. Charness, R. Hoffman and P. J. Feltovich, 223–42. Cambridge, MA: Cambridge University Press.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist, v.34*, #10, p.906.

Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2017). Even after thirteen class exams, students are still overconfident: the role of memory for past exam performance in student predictions. *Metacognition and Learning*, *12*(1), 1–19. https://doi.org/10.1007/s11409-016-9158-6

Freeman, S., Haak, D., & Wenderoth, M. P. (2011). Increased course structure improves

performance in introductory biology. *CBE-Life Sciences Education, v.10, #2,* p.175-186.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, v.111, #23,* p.8410-8415.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, v.141, #1*, p.2–18.

Gajjar NB (2013). The role of technology in 21st century education. *Int J Res Educ v.2,* p. 23–25.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S., (2001). What makes professional development effective? Results from a national sample of teachers. American Educational Research Journal, v.38, #4, p.915–945. doi:10.3102/00028312038004915

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave Macmillan.

Gilbert, L., Stempien, J., McConnell, D., Budd, D., van der Hoeven Kraft, K., Bykerk-Kauffman, Jones, M., Knight, C., Matheney, R., Perkins, D., and Wirth, K. (2012). Not just "Rocks for Jocks": Who are introductory geology students and why are they here? *Journal of Geoscience Education, 60*, 360-371

Greene, J. and Azevedo, R. (2009). A macro-level analysis of SRL processes and their relations to the acquisition of a sophisticated mental model of a complex system. *Contemporary Educational Psychology*, 34, 18-29.

Gutierrez, A. P., & Schraw, G. (2015). Effects of Strategy Training and Incentives on Students' Performance, Confidence, and Calibration. *Journal of Experimental Education, v.83, #3,* p.386–404.

Haak, D.C., Hille RisLambers, J., Pitre, El, & Freeman, S., (2011). Increased structure and active learning reduce the achievement gap in introductory biology. Science, v.332, June 3, p.1213-1216.

Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 429–455). New York: Taylor & Francis.

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, v.92, #1,* p.160–170.

Hadwin, A.F., & Webster, E.A., (2013). Calibration in goal setting: examining the nature of judgments of confidence. *Learning and Instruction, v.24,* p.37-47.

Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., & Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition and Learning*, *2*(2–3), 107–124. https://doi.org/10.1007/s11409-007-9016-7

Hartwig, M., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19*,126–134.

Hartwig, M. K., & Dunlosky, J. (2017). Category learning judgments in the classroom: Can students judge how well they know course topics? *Contemporary Educational Psychology*, *49*, 80–90. https://doi.org/10.1016/j.cedpsych.2016.12.002

Hattie, J. (2013). Calibration and confidence: Where to next? Learning and Instruction, v.24, #1, p.62–66

Hawker, M. J., Dysleski, L., & Rickey, D. (2016). Investigating general chemistry students' metacognitive monitoring of their exam performance by measuring postdiction accuracies over time. *Journal of Chemical Education, acs.jchemed.5b00705.*

Hilpert, J.C., Stempien, KJ., Katrien J. van der Hoeven Kraft. (2013). Evidence for the Latent Factor Structure of the MSLQ: A New Conceptualization of an Established Questionnaire. *Sage Open*, 1-10. https://doi.org/10.1177/2158244013510305

Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning, v.4, #2,* p.161–176.

Hurme, T. R., Merenluoto, K., & Järvelä, S. (2009). Socially shared metacognition of pre-service primary teachers in a computer-supported mathematics course and their feelings of task difficulty: A case study. *Educational Research and Evaluation, 15*(5), 503–524. doi:10.1080/13803610903444659

Hurtado, S., Eagan, M. K., Pryor, J. H., Whang, H., & Tran, S. (2012). Undergraduate teaching faculty: The 2010–2011 HERI Faculty Survey. Los Angeles: Higher Education Research Institute, UCLA.

Jacobs, J.E. & Paris, S.G. (1987). Children's metacognition about reading: Issues in definition, measurement, and instruction. *Educational Psychologist, v.22*, p.255–278.

Jamieson-Noel, D., & Winne, P. H. (2002). Exploring students' calibration of self reports about study tactics and achievement. *Contemporary Educational Psychology*, *27*(4), 551–572. https://doi.org/10.1016/S0361-476X(02)00006-1

Janssen, J., Erkens, G., Kirschner, P. A., & Kanselaar, G. (2012). Task-related and social regulation during online collaborative learning. *Metacognition and Learning, 7*(1), 25–43. doi:10.100711409-010-9061-5

Järvelä, S., & Hadwin, A. F. (2013). New Frontiers: Regulating Learning in CSCL. *Educational Psychologist, 48*(1), 25–39. https://doi.org/10.1080/00461520.2012.748006

Jones, J. P. & McConnell, D. A. (2016). CLASS: A new tool for characterizing student awareness of their learning in geoscience courses. *Geol. Soc. Am. Abst. with Prog., v. 48, #7*.

"" (2017). "Better learning through feedback: Improving student performance and judgment accuracy in an introductory geoscience course." In *Geological Society of America Abstracts with Programs, vol. 49 no. 6.*

"" (2018). How do effort and judgments of learning during online practice quizzes predict exam outcomes in an introductory physical geology course*? Geol. Soc. Am. Abst. With Prog. V.50, #6*.

"" (2019). "Completing the Self-Regulation cycle: Utilizing online trace data to characterize student learning behaviors in introductory physical geology courses" In *Geological Society of America Abstracts with Programs, vol. 51* no. 6.

Karpicke, J., Butler, A., and Roediger, H. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17*(4), 471-479.

Kastens, K., and Manduca, C. (2012). *Earth and Mind II: A synthesis of Research on Thinking and Learning in the Geosciences.* Geological Society of America Special Paper, Volume 486, https://doi.org/10.1130/SPE486

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*(2), 219–224.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*(4), 449.

Kortz, K.M, Grenga, A. M., & Smay, J. J. (2017) Establishing and applying literature-based criteria for effective communication of science to novices via introductory geology textbooks, *Journal of Geoscience Education, 65*:1, 48-59, DOI: 10.5408/16-205.1

Krathwohl, D.R., Bloom, B.S., and Masia, B.B., 1964, Taxonomy of educational objectives: Book 2, affective domain: New York, Longman, p. 196.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, v.77, #6,* p.1121–1134.

Kuncel, N.R., Crede, M., & Thomas, L.L., (2005). The validity of self-reported grade point

averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research, v.75, #1*, p.63-82.

Landers, R. N. (2014). Developing a Theory of Gamified Learning. *Simulation & Gaming*, *45*(6), 752–768. https://doi.org/10.1177/1046878114563660

Landers, R. N., Bauer, K. N., Callan, R. C., & Armstrong, M. B. (2015). Psychological Theory and the Gamification of Learning. In T. Reiners & L. C. Wood (Eds.), *Gamification in Education and Business* (pp. 165–186). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-10208-5_9

Libarkin, J.C., and Anderson, S.W., 2005, Assessment of learning in entry-level geoscience courses: results from the Geoscience Concept Inventory: Journal of Geoscience Education, v. 53, no. 4, p. 394–401.

Lopez, E. J., Nandagopal, K., Shavelson, R. J., Szu, E., & Penn, J. (2013). Self-regulated learning study strategies and academic performance in undergraduate organic chemistry: An investigation examining ethnically diverse students. *Journal of Research in Science Teaching, 50*(6), 660–676. https://doi.org/10.1002/tea.21095

Lovett, M. C. (2013). Make exams worth more than the grade: Using exam wrappers to promote metacognition. In M. Kaplan, N. Silver, D. LaVaque-Manty, & D. Meizlish (Eds.), *Reflection and Metacognition in College Teaching* (pp. 18–52). New York, NY: Stylus.

Lukes, L. A., 2012, "Going to the source: Student perspectives on factors that influence learning in college-level introductory geology courses," *Geological Society of America Paper 75*-2.

Lukes, L. A. and McConnell, D., 2014, "Comparing the exam preparation strategies of high and low performing students in college-level introductory geoscience courses," *Geological Society of America Paper, 72*-5.

Lukes, L. A., and McConnell, D., 2013, "What motivates introductory geology students? A qualitative analysis of student interviews from multiple institutions across the U.S.," *Geological Society of America Paper 192*-9.

Lukes, L. A., and McConnell, D.A., 2014, What motivates introductory geology students to study for an exam? *Journal of Geoscience Education,* v.62, #4, p.725-735.

Lynch, D. J. A. Y., & Trujillo, H. (2011). Motivational beliefs and learning strategies in organic chemistry. *International Journal of Science and Mathematics Education, (9),* 1351–1365.

Ma, J., & Baum, S. (2016). Trends in community colleges: enrollment, prices, student debt, and completion. College Board Research Research Brief, April 2016. 1-23.

Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Education & Psychology, v.97, #4*, p.723–731.

Maki, R. H., Willmon, C., & Pietan, A. (2009). Basis of metamemory judgments for text with multiple-choice, essay, and recall tests. *Applied Cognitive Psychology, v.23, #2*, p.204–222.

Manduca, C. and Mogk, D. (2006). *Earth and Mind: How Geologists Think and Learn About the Earth*, Geological Society of America Special Paper, Volume 413, https://doi.org/10.1130/SPE413

Marton, F. (2015). Reflections on post-exam reflections: Using exam wrappers at a two-year college. *Geological Society of America Abstracts with Programs. Vol. 47*, No. 7, p.662

McConnell, D. A. (2019). Research-based Instructional Reform in the Geosciences: Building a Community of Practice. In *Levers for Change: An Assessment of Progress on Changing STEM Instruction.* American Association for the Advancement of Science.

McConnell, D. A. and van Der Hoeven Kraft, K. J. (2011). Affective Domain and Student Learning in the Geosciences. *Journal of Geoscience Education, 59*, 106-110.

McConnell, D. A., Chapman, L., Czajka, C. D., Jones, J. P., Ryker, K. D., & Wiggen, J. (2017). Instructional Utility and Learning Efficacy of Common Active Learning Strategies. *Journal of Geoscience Education*, 65, 604–625. https://doi.org/10.5408/17-249.1

McCardle, L., & Hadwin, A. F. (2015). Using multiple, contextualized data sources to measure learners' perceptions of their self-regulated learning. *Metacognition and Learning, 10*(1), 43–75. https://doi.org/10.1007/s11409-014-9132-0.

McCrindle, A.R., & Christensen, C.A., (2005). The impact of learning journals on metacognitive and cognitive processes and learning performance. *Learning and instruction v.5, #2,* p.167-185.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavior Research Methods, Instruments, and Computers, 36*, 222–233.

McNeal, K. S.; van der Hoeven Kraft, K.; Nagy-Shadman, E.; Beck, M.; and Jones, J. P. (2018). "Research on Geoscience Students' Self-Regulated Learning, Metacognition,

and Affect". In St. John, K (Ed.) (2018). *Community Framework for Geoscience Education Research. National Association of Geoscience Teachers*. Retrieved from DOI https://doi. org/10.25885/ger_framework/10

Mega, C., Ronconi, L., & De Beni, R. (2014). What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement. *Journal of Educational Psychology*, 106(1), 121–131.

Metzger, K., Smith, B., Brown, E. & Soneral, P. (2018). SMASH: A diagnostic tool to monitor student metacognition, affect and study habits in an undergraduate science course. *Journal of College Science Teaching, 47*(3), 88–99.

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: the influence of incentives and feedback on exam prediction. *Metacognition and Learning, v.6, #3,* p.303–314.

"" (2011). Unskilled but aware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *v.37*, *#2*, p.502–506.

Moos, D. C., & Azevedo, R. (2008). Exploring the fluctuation of motivation and use of self-regulatory processes during learning with hypermedia. *Instructional Science, 36*(3), 203–231.

National Research Council (1996). National Science Education Standards: Washington D.C., National Academies Press - National Committee on Science Education Standards and Assessment, 272 p.

National Research Council. (2012). *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering.* S. R. Singer, N.

R. Nielsen, and H. A. Schwingruber, (Eds.) Washington, DC: The National Academies Press.

National Science Board (NSB). (2007). *A National Action Plan for Addressing the Critical Needs of the U.S. Science, Technology, Engineering, and Mathematics Education System* Arlington, VA: National Science Foundation.

Nelson, T. O., & Narens, L. (1990). Metamemory: a theoretical framework and new findings. *The Psychology of Learning and Motivation, 26*, 125–141.

Nezlek, J. B. (2012). Multilevel modeling for psychologists. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbooks in psychology®. APA handbook of research methods in psychology, Vol. 3. Data analysis and research publication* (p. 219–241). American Psychological Association.

Nietfeld, J. L. (2002). The Effect of Knowledge and Strategy Training on Monitoring Accuracy. *The Journal of Educational Research*, *v.95*, *#3*, p.131–142.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning, v.1, #2*, p.159–179.

Nietfeld, J. L., Cao, L., Osborne, J. W., Taylor, P., & Osborne, J. W. (2005). Classroom metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Education*, *v.74*, *#1*, p.7–28.

Nilson, L., (2013). Creating self-regulated learners: Strategies to strengthen students' self-awareness and learning skills. Sytlus Publishing, 180 p.

Nuhfer, E., & Knipp, D. (2003). The knowledge survey: A tool for all reasons. *To improve*

the academy v.21, p.59-78.

Nunez, B., Lukes, L., and Rushing, M., 2015, "Type, Frequency, and Timing of Student Study Strategies within an Introductory Historical Geology Course in an Active Learning Technology (ALT) Classroom," *Geological Society of America Paper 208*-4.

Oktay, J.S. 2012. Grounded theory [electronic resource]. Oxford, UK: Oxford University Press. Available at http://www.oxfordscholarship.com.prox.lib.ncsu.edu/view/10.1093/acprof:oso/9780199753697.001.0001/acprof-9780199753697 (accessed 1 June 2012).

Panadero, E. (2017). A Review of Self-regulated Learning: Six Models and Four Directions for Research. *Frontiers in Psychology, 8*(April), 1–28. https://doi.org/10.3389/fpsyg.2017.00422

Panadero, E., & Järvelä, S. (2015). Socially Shared Regulation of Learning: A Review. *European Psychologist, 20*(3), 190–203. https://doi.org/10.1027/1016-9040/a000226

Pape-Lindstrom, P., Eddy, S., & Freeman, S. (2018). Reading quizzes improve exam scores for community college students. *CBE Life Sciences Education, 17*(2), 1–8. https://doi.org/10.1187/cbe.17-08-0160

Perkins, D. & Wirth, K. (2011). What can we do about declining student performance in our classrooms?. *Geological Society of America Abstracts with Programs. Vol. 43*, No. 5, p.635

Perry, N. E., & Winne, P. H. (2006). Learning from learning kits: gStudy traces of students' self-regulated engagements with computerized content. *Educational Psychology Review, 18*(3), 211–228

Pieschl, S. (2009). Metacognitive calibration and extended conceptualization and potential applications. *Metacognition and Learning, v. 4*, p.3-31.

Pintrich, P. (2000). The role of goal orientation in self-regulated learning. In Boekarts, Pintrich, and Zeidner (Eds.). *Handbook of self-regulation.* London: Academic Press.

Pintrich, P. R., & De Groot, E. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*, 33–40.

Pintrich, P.R., & Zusho, A. (2002). The development of academic self-regulation: The role of cognitive and motivational factors. In A. Wigfield, & J.S. Eccles (Eds.), *Development of achievement motivation* (pp. 249–284). San Diego, CA: Academic.

Pintrich, P. and Zusho, A. (2007). Student Motivation and self-regulated learning in the college classroom. In R.P. Perry and J.C. Smart (Eds.). *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, 731-810.

Pollock, S.J., & Finkelstein, N.D., (2008). Sustaining educational reforms in introdutory physics. Physical Review Special Topics – Physics Education Research, v.4, p.1-8. DOI 10.1103/PhysRevSTPER.4.010110

President's Council of Advisors on Science and Technology (PCAST) (2012). Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics. Report to the President. (accessed 15 April 2016).

Pressley, M., & Ghatala, E. S. (1990). Self-regulated learning Monitoring learning from text. *Educational Psychologist, v.25, #1*, p.19-33.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage Publications.

Riggs, E., Balliet, R., Lieder, C. (2009). Effectiveness in problem solving during geologic field examinations: insights from analysis of GPS tracks at variable time scales. In Whitmeyer, S., Mogk, D., and Pyle, E. eds., Field Geology Education: Historical Perspectives and Modern Approaches: *Geological Society of America Special Paper 461*, p. 323-340.

Rovers, S. F. E., Clarebout, G., Savelberg, H. H. C. M., de Bruin, A. B. H., & van Merriënboer, J. J. G. (2019). Granularity matters: comparing different ways of measuring self-regulated learning. *Metacognition and Learning, 14*(1), 1–19. https://doi.org/10.1007/s11409-019-09188-6

Sabel, J. L., Dauer, J. T., & Forbes, C. T. (2017). Introductory biology students' use of enhanced answer keys and reflection questions to engage in metacognition and enhance understanding. *CBE Life Sciences Education, 16*(3), 1–12. https://doi.org/10.1187/cbe.16-10-0298

Sailer, M., Hanse, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. Computers in Human Behavior, 69, 371–380.

Schellings, G. L. M. (2011). Applying learning strategy questionnaires: problems and possibilities. *Metacognition and Learning, 6*(2), 91–109.

Schellings, G. L. M., van Hout-Wolters, B. H., Veenman, M. V., & Meijer, J. (2013). Assessing metacognitive activities: the in-depth comparison of a task-specific questionnaire with think-aloud protocols. *European Journal of Psychology of Education, 28*(3), 963–990.

Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Educational Psychology*. https://doi.org/10.1006/ceps.1994.1013

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science, v.26, #1*, p.113–125.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, v.4*, #1, p.33–45.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, v.19*, p.460–475.

Schraw, G., & Roedel, T. D. (1994). Test difficulty and judgment bias. Memory & Cognition, 22, 63 –69.

Schraw, G., Crippen, K., Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education 36*, 111-139.

SERC (2007). Student Motivations and Attitudes: The Role of the Affective Domain in Geoscience Learning. Retrieved 1 August 2014 from https://serc.carleton.edu/NAGTWorkshops/affective/workshop07/index.html

SERC (2008). The Role of Metacognition in Learning. Retrieved 1 August 2014 from https://serc.carleton.edu/NAGTWorkshops/metacognition/workshop08/index.html

Serra, M. J., & Metcalfe, J. (2009). Effective implementation of metacognition. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), Handbook of metacognition in education (pp. 278e298). New York, NY: Routledge

Singer, S.R., Nielsen, N.R., Schweingruber, H.A. (2012). *Discipline-Based Education*

Research: Understanding and Improving Learning in Undergraduate Science and Engineering. Washington, D.C.: National Academies Press

Sletten, S. R. (2017). Investigating Flipped Learning: Student Self-Regulated Learning, Perceptions, and Achievement in an Introductory Biology Course. *Journal of Science Education and Technology, 26*(3), 347–358. https://doi.org/10.1007/s10956-016-9683-8

Smith, B. A., Metzger, K., & Soneral, P. (2019). Investigating introductory nonmajor biology students' self-regulated learning strategies through the implementation of a reflective-routine. *Journal of College Science Teaching, 48*(6), 66–76.

Snijders, T. & Bosker, R. (1999). Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. SAGE Publicaitons.

Son, L. and Simon, D. (2012). Distributed Learning: Data, Metacognition, and Educational Implications. *Educational Psychology Review, 24*(3), 379-399.

Spencer, D., Thomson, M. M., & Jones, J. P. (2018). Socially Shared Metacognition Among Undergraduate Students During an Online Geology Course. In *Handbook of Research on Student-Centered Strategies in Online Adult Learning Environments* (pp. 406–439). https://doi.org/10.4018/978-1-5225-5085-3.ch019

St. John, K. (Ed.) (2018). A Community Framework for Geoscience Education on Research. National Association on of Geoscience Teachers. Retrieved from http://commons.lib.jmu.edu/ger_framework/15

St. John, K., and McNeal, K.S. 2017. The strength of evidence pyramid: One approach for characterizing the strength of evidence of geoscience education research (GER) community claims. *Journal of Geoscience Education, 65*(4):363–372.

Stanton, J. D., Neider, X. N., Gallegos, I. J., & Clark, N. C. (2015). Differences in metacognitive regulation in introductory biology students: When prompts are not enough. *CBE Life Sciences Education, 14*(2), 1–12. https://doi.org/10.1187/cbe.14-08-0135

Stelzer, T., Gladding, G., Mestre, J.P., and Brookes, D.T. (2009). Comparing the efficacy of multimedia module with traditional textbooks for learning introductory physics content. *American Journal of Physics, v.77, #2*, p.184-190.

Strayer, J. F. (2012). How learning in an inverted classroom influences cooperation, innovation and task orientation. *Learn Environ Res, v.15*, p.171–193.

Thiede, K.W., Griffin, D.,Wiley, J., & Redford, J. S. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), Handbook of metacognition in education (pp. 85–106). New York, NY: Routledge.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of a survey response*. Cambridge: Cambridge University Press

van Der Hoeven Kraft, K. J., Srogi, L., Husman, J., Semken, S., & Fuhrman, M. (2011). Engaging Students to Learn Through the Affective Domain: A new Framework for Teaching in the Geosciences. *Journal of Geoscience Education, 59*(1), 71–84. https://doi.org/10.5408/1.3543934

van der Hoeven Kraft, K. J. (2017). Developing Student Interest: An Overview of the Research and Implications for Geoscience Education Research and Teaching Practice. *Journal of Geoscience Education, 65*(4), 594–603. https://doi.org/10.5408/16-215.1

Veenman, M. V. J. (2005). The assessment of metacognitive skills: What can be learned from multi-method designs? In C. Artelt & B. Moschner (Eds.), *Lernstrategien und*

Metakognition: Implikationen für Forschung und Praxis (pp. 77–99). Münster: Waxmann.

Veenman, M. V. J. (2011). Alternative assessment of strategy use with self-report instruments: a discussion. *Metacognition and Learning, 6*(2), 205–211.

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: conceptual and methodological considerations. *Metacognition and Learning, 1*(1), 3–14. https://doi. org/10.1007/s11409-006-6893-0

Volet, S. E., Summers, M., & Thurman, J. (2009). High-level co-regulation in collaborative learning. How does it emerge and how is it substained? *Learning and Instruction, 19*, 128–143. doi:10.1016/ j.learninstruc.2008.03.001

Weinstein, C. E., & Palmer, D. R. (2002). *Learning and study strategies inventory* (2nd. ed.). Clearwater, FL: H & H Publishing.

Wetzel, C., Radtke, P., & Stern, H. (1994). Instructional effectiveness of video media. Hillsdale, NJ: Lawrence Erlbaum Associated.

Winne, P. (1995). Self-regulation is ubiquitous but its forms vary with knowledge. *Educational Psychologist, 30,* 4, 223-228.

Winne, P., & Hadwin, A. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 227-304). Mahwah, NJ: Erlbaum.

Winne, P. H. (2013). Learning strategies, study skills, and self-regulated learning in post-secondary education. In M. B. Paulsen (Ed.), *Higher education: Handbook of theory and research* (pp. 377–403). Netherlands: Springer. http://dx.doi.org/10.1007/978-94-007-5836-0_8

Winne, P., & Hadwin, A. (2008). The weave of motivation and self-regulated learning. In D. Schunk & B. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and applications* (pp. 297-314). NY: Taylor & Francis.

Winne, P. & Jamieson-Noel, D. (2002). Exploring students' calibration of self-reports about study tactics and achievement. *Contemporary Educational Psychology, 27*, 551-572

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 531–566). Orlando: Academic Press.

Winne, P., Hadwin, A. F., & Perry, N. E. (2013). Metacognition and computer-supported collaborative learning. In C. E. Hmelo-Silver, A. O'Donnell, C. Chan, & C. Chinn (Eds.), *International handbook of collaborative learning.* New York, NY: Taylor & Francis.

Wiggins, G.P., and McTighe. J.,2005, Understanding by design. Ascd.

Wilson, C. (2017a). Status of the geoscience workforce 2014. Alexandria, VA : American Geosciences Institute.

Wilson, C. (2017b). Status of recent geoscience graduates 2014. Alexandria, VA : American Geosciences Institute.

Wirth, K.R.., and Perkins, D. 2005, Knowledge surveys: An indispensable course design and assessment tool. Presented at the Innovations in the Scholarship of Teaching and Learning at Liberal Arts Colleges, St. Olaf, Northfield, MN.

Wirth K. & Perkins, D. (2011). Students learning about their own learning: Why not in your course?. *Geological Society of America Abstracts with Programs. Vol. 43*, No. 5, p.635

Wolfe, B. A. (2016). Tapping the Geoscience Two-Year College Student Reservoir: Factors that Influence Student Transfer Intent and Physical Science Degree Aspirations. Doctoral dissertation. University of Kansas.

Wysession, M.E., LaDue, N., Budd, D.A., Campbell, K., Conklin, M., Kappel, E., Lewis, G., Raynolds, R., Ridky, R.W., Ross, R.M.., Taber, J., Tewksbury, B., and Tuddenham, P., (2012). Developing and applying a set of earth science literacy principles. Journal of Geoscience Education, v.60, #2, p.95-99.

Yan, V. X., Thai, K. P., & Bjork, R. A. (2014). Habits and beliefs that guide self-regulated learning: do they vary with mindset? *Journal of Applied Research in Memory and Cognition, 3*(3), 140–152.

Zhang, D., Zhou, L., Briggs, R., & Nunamaker, Jr., J. (2006). Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. *Information and Management, v.43, #1*, p.15-27.

Zimmerman, B. J. (1986). Becoming a self-regulated learner: which are the key subprocesses? *Contemp. Educ. Psychol. 11*, 307–313. doi: 10.1016/0361-476x(86)90027-5

“” (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology, 81*, 329–339.

“” (1990). Self-Regulated Learning and Academic Achievement: An Overview. *Educational Psychologist*, *v.25, #1*, p.3-17.

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal v.45, #1*, p.166-183.

Zimmerman, B. J., and Martinez-Pons, M. (1986). Development of a structured interview for assessing student use of self-regulated learning-strategies. *Am. Educ. Res. J. 23*, 614–628. doi: 10.3102/00028312023004614

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 299–315). New York: Routledge.

Zimmerman, B. J., & Schunk, D. H. (Eds.). (2011). *Handbook of self-regulation of learning and performance.* New York: Routledge.

Zohar, A., & Barzilai, S. (2013). A review of research on metacognition in science education: current and future directions. *Studies in Science Education*, *v.49, #2*, p.121–169

Zusho, A., Pintrich, P. R., & Coppola, B. (2003). Skill and will: The role of motivation and cognition in the learning of college chemistry. *International Journal of Science Education, 25*(9), 1081–1094.